Artificial Intelligence Review (2020) 53:1655–1720
https://doi.org/10.1007/s10462-019-09716-5

Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges

Ghulam Murtaza[1,2] · Liyana Shuib[1] · Ainuddin Wahid Abdul Wahab[1] · Ghulam Mujtaba[1] · Ghulam Mujtaba[3] · Henry Friday Nweke[1,4] · Mohammed Ali Al-garadi[5] · Fariha Zulfiqar[1] · Ghulam Raza[6] · Nor Aniza Azmi[7]

Published online: 25 May 2019

Abstract
Breast cancer is a common and fatal disease among women worldwide. Therefore, the early and precise diagnosis of breast cancer plays a pivotal role to improve the prognosis of patients with this disease. Several studies have developed automated techniques using different medical imaging modalities to predict breast cancer development. However, few review studies are available to recapitulate the existing literature on breast cancer classification. These studies provide an overview of the classification, segmentation, or grading of many cancer types, including breast cancer, by using traditional machine learning approaches through hand-engineered features. This review focuses on breast cancer classification by using medical imaging multimodalities through state-of-the-art artificial deep neural network approaches. It is anticipated to maximize the procedural decision analysis in five aspects, such as types of imaging modalities, datasets and their categories, pre-processing techniques, types of deep neural network, and performance metrics used for breast cancer classification. Forty-nine journal and conference publications from eight academic repositories were methodically selected and carefully reviewed from the perspective of the five aforementioned aspects. In addition, this study provided quantitative, qualitative, and critical analyses of the five aspects. This review showed that mammograms and histopathologic images were mostly used to classify breast cancer. Moreover, about 55% of the selected studies used public datasets, and the remaining used exclusive datasets. Several studies employed augmentation, scaling, and image normalization pre-processing techniques to minimize inconsistencies in breast cancer images. Several types of shallow and deep neural network architecture were employed to classify breast cancer using images. The convolutional neural network was utilized frequently to construct an effective breast cancer classification model. Some of the selected studies employed a pre-trained network or developed new deep neural networks to classify breast cancer. Most of the selected studies used accuracy and area-under-the-curve metrics followed

✉ Ghulam Murtaza
gmurtaza@iba-suk.edu.pk

✉ Liyana Shuib
liyanashuib@um.edu.my

Extended author information available on the last page of the article

by sensitivity, precision, and F-measure metrics to evaluate the performance of the developed breast cancer classification models. Finally, this review presented 10 open research challenges for future scholars who are interested to develop breast cancer classification models through various imaging modalities. This review could serve as a valuable resource for beginners on medical image classification and for advanced scientists focusing on deep learning-based breast cancer classification through different medical imaging modalities.

Keywords Breast cancer classification · Deep learning · Medical imaging modalities · Convolutional neural network

1 Introduction

The World Health Organization (2018) reported that cancer is the leading cause of non-accidental deaths worldwide; in specific, approximately 8.8 million people globally died of cancer in 2015. Breast cancer (BrC) is a common and fatal disease among women worldwide. BrC is the third highest fatal disease among different cancer types, such as lung, liver, and brain. According to WHO, 1.7 million (11.3%) casualties reported in 2015 were related to BrC (World Health Organization 2018). In addition, the number of new BrC patients is expected to increase by 70% in the next 20 years. Therefore, early and precise diagnosis plays a pivotal role to improve the prognosis and increase the survival rate of patients with BrC from 30 to 50% (World Health Organization (2018). In general, breast tumor has two types, benign and malignant. Benign is a noninvasive (non-cancerous) while malignant is an invasive (cancerous) type of tumor. Both tumors have further subtypes that need to be diagnosed individually because each may lead to different prognosis and treatment plans. Proper diagnosis requires accurate identification of each subcategory of BrC, also called BrC multi-classification. Medical imaging modalities are more commonly adopted and effective for BrC detection than any other testing method. Well-known medical imaging modalities for BrC diagnosis are mammography (breast X-ray images), ultrasound (US) imaging or sonograms, magnetic resonance imaging (MRI), computed tomography (CT) and Histopathology (HP) image (Beutel et al. 2000; Goceri 2017; Kasban et al. 2015). Medical imaging is usually performed manually by one or more expert doctors (radiologist, sinologist, or pathologist). An absolute decision is made after consensus if more than one pathologist is available for BrC HP image analysis; otherwise, findings are reported by one pathologist only. Nonetheless, manual HP image analysis faces three main issues (Gurcan et al. 2009; Sophie Softley Pierce 2017). First, more than one expert pathologist at one place is usually unavailable in developing countries. Second, the procedure of image analysis for the multi-class classification of BrC is cumbersome and time consuming for pathologists. Therefore, pathologists may experience fatigue and deteriorated attention during image analysis. Finally, a reliable BrC subtype identification depends on the professional experience and domain knowledge of an expert pathologist. These issues may cause misdiagnosis, especially in the early stages of BrC. However, computer-aided diagnosis (CAD) systems can serve as a second opinion to solve BrC multi-classification problems. A CAD system is an affordable, readily available, fast, and reliable source of early diagnosis (Doi 2007; Sadaf et al. 2011). This system assists radiologists and physicians in identifying abnormalities by using various imaging modalities, which have reduced the mortality rates from 30 to 70% (Schneider and Yaffe 2000).

The advent of digital images in medical science has provided an edge to artificial intelligence (AI) for pattern recognition using a CAD system. CAD systems are designed to assist

doctors by automatic image interpretation. Hence, such a system reduces human dependency, increases diagnosis rate, and reduces the overall treatment expenses by reducing false positive and false negative (FN) predictions (Goceri and Songul 2018). Moreover, increased FN rate (sensitivity) may lead to no treatment for a BrC carrier, and misdiagnoses usually occur in the early stages of BrC. Sadaf et al. (2011) reported that the use of a CAD system for BrC classification increases sensitivity by 10%. Apart from classification (Rouhi et al. 2015; Spanhol et al. 2017; Zheng et al. 2017), CAD systems have also been developed to perform other diagnosis-related tasks, such as BrC lesion detection (Ertosun and Rubin 2015; Wang et al. 2017; Wang and Yang 2018; Yousefi et al. 2018), segmentation (Lo et al. 2014; Pan et al. 2017; Shan et al. 2016), registration (Adoui et al. 2017), and grading (Cao et al. 2016; Wan et al. 2017).

Recently, several articles have been published to solve BrC classification, segmentation, registration, detection, or grading problems by using traditional machine learning (ML) approaches (e.g., support vector machine, Naïve Bayes, and decision tree) or by using state-of-the-art artificial neural network (ANN)-based approaches [e.g., shallow neural networks (SNNs) and deep neural networks (DNNs)]. A SNN is based on a single hidden layer between input and output layers, whereas DNNs mostly consist of two or more than two hidden layers along with input and output layers. However, only few review articles (Chen et al. 2017b; Goceri and Goceri 2017; Jalalian et al. 2017; Lee and Chen 2015; Litjens et al. 2017; Mehdy et al. 2017; Nahid and Kong 2017a; Sathish et al. 2016; Yassin et al. 2018) are available to recapitulate BrC classification using medical imaging modalities. For instance, studies (Chen et al. 2017b; Jalalian et al. 2017; Lee and Chen 2015) reviewed publications related to traditional ML approaches using hand-engineered features (HEFs) for the analysis of cancer images, including BrC images. In addition, Chen et al. (2017b) summarized recent works on the use of hematoxylin and eosin (H&E) HP images for BrC prognosis. The authors discussed and analyzed different medical imaging modalities, image pre-processing tasks, and image detection, segmentation, and feature extraction techniques. Finally, traditional ML-based studies were evaluated with future direction. Lee and Chen (2015) focused on the detection of common forms of cancer, such as breast, lung, prostate, and skin, using multimodalities, such as X-ray, US, and CT. The authors analyzed traditional ML techniques adopted for each cancer detection, segmentation, and classification. In addition, the use of various imaging modalities for cancer detection was discussed and compared. Finally, future directions were suggested for new researchers. However, previous review studies mainly focused on traditional ML approaches by using imaging modalities usually for binary classification. Conversely, recent review studies (Mehdy et al. 2017; Sathish et al. 2016) have emphasized on ANNs using multimodalities for BrC analysis. For instance, Mehdy et al. (2017) focused on state-of-the-art ANN techniques by using breast image multimodalities. They analyzed various types of ANN adopted for BrC analysis by using multimodalities, such as Mg, US, MRI, and thermal imaging. Sathish et al. (2016) studied various medical imaging modalities and ANN-based CAD approaches for BrC detection. They performed a comparative analysis of the imaging procedures, benefits, and limitations of Mg, US, MRI, and thermography. However, the aforementioned reviews provided a generic analysis of the applications of various ANN-based models by using multimodalities. Furthermore (Litjens et al. 2017; Nahid and Kong 2017a; Yassin et al. 2018), studies explored all major types of CAD systems for BrC diagnosis using many types of medical images, including HP images. For instance, Yassin et al. (2018) performed a systematic review of the various types (e.g., traditional ML, SNN, and deep learning) of CAD system for BrC diagnosis. The authors followed a systematic approach to select relevant studies from authentic sources to ensure high review quality. In addition, they investigated state-of-the-art CAD diagnostic approaches for

BrC by using well-known medical imaging modalities, analyzed all types of classifiers and feature extraction methods, and briefly studied the performance measures commonly used to compare BrC image diagnoses. Finally, future directions were suggested to highlight areas that need further investigations. Litjens et al. (2017) explored deep learning-based CAD systems for medical image classification, detection, segmentation, and registration. Moreover, the authors provided a brief review of different types of cancer, such as brain, eye, chest/lung, heart, abdominal, bone/joints, and BrC. Open issues and future directions were also discussed to motivate new researchers in medical image diagnosis through deep learning.

Most review studies published on BrC detection focused on traditional ML algorithms, generic ANNs, or SNNs where feature extraction is especially involved. Existing review studies presented grayscale images and rarely discussed other modalities, such as HP images. In addition, their discussion on deep learning techniques was incomprehensible. Thus, to overcome the aforementioned limitations, this study presents a systematic and critical review of existing state-of-the-art DNN-based CAD systems for BrC image classification from five aspects, namely, BrC imaging modalities, datasets, image pre-processing, DNNs, and performance measurements. The review adopts a systematic review methodology for searching and selecting studies from well-known sources to ensure the authenticity and quality of selected literature. Furthermore, this review provides a critical analysis of DNN performance on different publicly available datasets. Finally, this review presents 10 new research directions and research challenges for future researchers who intend to work in BrC image classification using DNNs.

This review is organized as follows. Section 2 describes the research methodology for the selection of studies. BrC state-of-the-art DNN types are given in Sect. 3. Discussion and future works are detailed in Sects. 4 and 5, respectively. Finally, a conclusion is provided in Sect. 6.

2 Research methodology

This review incorporates the systematic literature review guidelines suggested by Kitchenham and Charter (Keele 2007). The four fundamental phases of this review are planning, searching and filtering of primary studies, information extraction, and information synthesis. The planning phase defines the review goals, scope, and protocols (Sect. 2.1). The second phase, study selection procedure and criteria, comprises formulation of search keywords, segregation of keywords in groups, and writing search queries (Sect. 2.2). The third phase is composed of screening criteria and quality evaluation criteria for the collected studies (Sect. 2.3). The information extraction approach is described in Sect. 2.4. The last phase, a systematic review, involves information synthesis and critical analysis (Sect. 3).

2.1 Survey scope identification

This review aims to identify various studies related to BrC classification using numerous medical imaging modalities through DNNs. The primary scope of this study is to find the answers to the following research questions:

1. What are the medical imaging modalities used for BrC classification?
2. What are the medical image datasets used to develop DNN-based classification models?
3. What are the pre-processing techniques to improve the classification results?
4. What are the DNN types currently applied to BrC classification using medical imaging modalities?

5. What are the evaluation metrics used to assess the performance of DNN-based classification models?

Moreover, current challenges and their possible solutions along with future research direction are discussed. In the conclusion, the cruxes of overall review are summarized.

2.2 Studies searching strategy and results

Considering the review scope, the authors unanimously prepared seven groups of search criterion (Table 1) to identify relevant studies. Moreover, each group was coupled with the "AND" operator, and search keywords within the first four groups were coupled with the "OR" operator. However, the search criterion was applied on eight journal archives, including Association for Computing Machinery, Institute of Electrical and Electronics Engineers Xplore, Medical Literature Analysis and Retrieval System Online, PubMed, Science Direct, Scopus, SpringerLink, and Web of Science. These eight journal archives contain most of the research articles related to the five research questions listed above. The first four groups of search criterion consisted of search tokens (Table 1). Group 1 contained key words such as "breast" and "breast cancer" to ensure that the study is related to BrC only. Group 2 was composed of key words "mammogram", "mammography", "pathology images", "biopsy image", "histopathology image", "histopathological image", "ultrasound", "MRI", "magnetic resonance", "medical image", "CT", "computed tomography", "CAT", or "PET" to select studies that used at least one medical imaging modality. Group 3 key words include "classification", "multi-classification", "screening", "diagnosis", "prognosis", and "bounds" to extract studies related to classification. Meanwhile, Group 4 key words are "deep learning", "neural network", "stacked auto-encoder", "RNN", "CNN", "convolutional neural network", "machine learning", "transfer learning", "AlexNet", "VGG", "computer aided", "computer-aided", "multilayer perceptron", and "CAD" to isolate studies that used any ANN model. Group 5 of search criterion limited the duration of publication from January 2014 to June 2018. The trends of research publications showed that the use of DNN for breast image classification started in 2014. Finally, Groups 6 and 7 refined the studies to articles or proceedings published in English. Hence, application of the search criterion to the eight journal archives yielded 2446 studies on keyword search only. Furthermore, the studies were excluded and reduced by number from 2446 to 977 when overall seven groups of the search criterion were applied one after the other (for detailed flow, see Fig. 1).

2.3 Screening and selection criteria

In the first stage of searching strategy, 977 publications were collected. However, the search criterion was applied on the archives of eight journals individually. Therefore, one study could be shared by many journal archives simultaneously. Hence, duplicate studies were removed, and the remaining 537 studies were used for further screening and selection. Moreover, the process of paper screening and selection is based on two phases. In the first phase, paper selection was solely made on title and abstract reading. In the second phase, full article reading was performed. In phase-1, the paper screening and selection of the remaining 537 studies were analyzed by two reviewers individually. The aim of Phase-1 screening is to isolate studies that are highly related to the defined review scope. Thereafter, in phase 2, the 110 shortlisted studies were scrutinized by two more reviewers for further screening, and only 54 studies were retained. Finally, all reviewers discussed and compared the rejected or

Table 1 Selected search keywords in the different groups

Search groups	Search in	Criterion	Description
Group1	Title	"breast*" OR "breast cancer"	Disease-related search key words.
Group2	Topic	"mammogram*" OR "mammography" OR "pathology images" OR "biopsy image*" OR "histopathology image*" OR "histopathological image*" OR "ultrasound" OR "MRI*" OR "magnetic resonance" OR "medical image*" OR "CT" OR "computed Tomography" OR "CAT" OR "PET"	Medical imaging modality-related search key words
Group3	Title	"classification" OR "multi-classification" OR "screening" OR "diagnosis" OR "prognosis"	Classification-related key words
Group4	Title	"deep learning" OR "*neural network*" OR "stacked auto-encoder" OR "RNN" OR "CNN" OR "convolutional neural network*" OR "machine learning" OR "transfer learning" OR "AlexNet" OR "VGG*" OR "computer aided*" OR "computer-aided*" OR "multilayer perceptron" OR "CAD"	Artificial Neural Network-related key words
Group5	Publication Year	2014 to June 2018	Studies included in publication duration
Group6	Study Type	Article or Conference Proceedings	Study types Included
Group7	Language	English	Languages Included
Search Query	Group1 **AND** Group2 **AND** Group3 **AND** Group4 **AND** Group5 **AND** Group6 **AND** Group7		

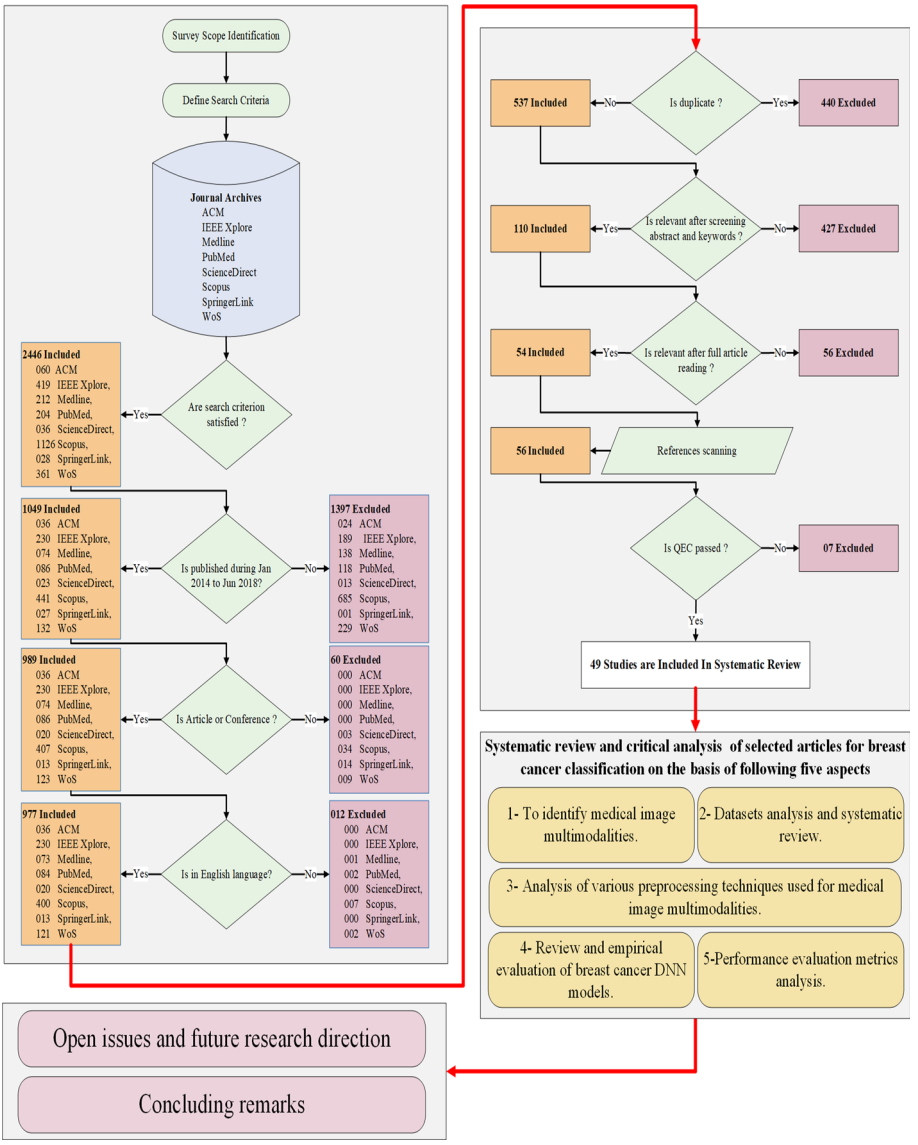


Fig. 1 Flow diagram illustrates the research methodology

selected papers until a consensus was reached. For any disagreements among reviewers, the majority rule was used to decide the inclusion or exclusion of the articles for final review. The screening criteria for the inclusion and exclusion of studies, adopted by the reviews, are as follows:

1. The manuscript must be published from 2014 to June 2018.
2. Manuscript sections should be in English.

   - Exclude papers that only have the abstract written in English (Aksebzeci and Kayaalti 2017; Tataroğlu et al. 2017)

3. Complete manuscript found.
   - Exclude if only the title or abstract is found (Antropova et al. 2018a, b; Mendel et al. 2018).
4. The manuscript should be an article or conference proceeding.
   - Exclude all other types, such as review paper, book sections (Selvathi and Aarthy Poornila 2018; Xu et al. 2017), or any other type of papers.
5. The manuscript should be related to BrC only.
   - Exclude all other diseases solely based on the heart (Hussain et al. 2015), lung, bladder, or prostate cancer.
6. Authorized or publicly available medical image datasets should be used.
7. The manuscript should focus on BrC classification.
   - Exclude papers focusing on detection (Cruz-Roa et al. 2017; Mina and Mat Isa 2015; Wang et al. 2014), segmentation (Chen et al. 2017a; Pan et al. 2017), cancer patient survivability (Jyh-Horng et al. 2014; Sivachitra and Vijayachitra 2015), or CAD analysis (Tan et al. 2013; van Zelst et al. 2018)-based publication
8. The manuscript must use at least one of the standard medical imaging modalities
   - Exclude other types of imaging techniques, such as microwave tomography imaging (Pack et al. 2016), synthetic breast imaging (Hassan et al. 2016), and signal imaging such as ultra wideband (Conceição et al. 2014) porotype imaging.

Moreover, the references of 54 papers were scanned, and two more studies (Bekker et al. 2016; Kumar et al. 2017a) matching the above inclusion criteria were found.

2.4 Quality evaluation strategy

To address the objectives of review, we assessed the quality of 56 studies on the basis of measurable quality evaluation criteria (QEC). The quantifiable QEC-adopted methodology guaranties that the chosen studies support the goals of this research. Therefore, all authors prepared a close-ended QEC check list individually, and the final list was approved unanimously after a rigorous discussion. Finally, a close-ended questions checklist (Appendix Table 11) was prepared and used as QEC for the unbiased quality evaluation of 56 previously chosen papers. The close-ended questions were answered as 1 or 0 assumed yes or no, respectively, and to easily sum up the total score for each selected study (Appendix Table 12). Two groups of reviews were selected to assess the chosen studies. Thereafter, the score of each study was calculated and evaluated by Cohen's kappa score for inter-rater agreement. The quality of the 56 selected studies was evaluated by measurable QEC. QEC was adopted to determine whether the selected primary study is appropriate to accomplish the intended review objectives. Disagreements between two reviewer groups had been addressed through the Delphi method (Dalkey and Helmer 1963) until a consensus was developed for the final selection of any study. Finally, all the reviews decided a cut-off value (i.e., 7) for QEC in including studies for systematic review. Hence, only 49 studies were filtered out by following the overall quality evaluation process. Therefore, this review included 49 studies.

2.5 Information abstraction

The abstraction of information from 49 chosen primary studies was organized in table form and comprised seven traits, namely, study objective(s), medical imaging modality used, DNN technique adopted, details of dataset(s) used, number of predicted classes, results, and future work suggested. Section 3 presents a critical and analytical review of these six aspects.

3 Breast cancer classification state of the art

This section covers the overall analysis of BrC classification by discussing almost all major studies. This review can assist researchers in BrC classification to gain a better, concise perspective of existing problems, solutions, and future directions. As mentioned in the methodology section (i.e., Sect. 2), 49 studies were scrutinized to achieve five goals: imaging modalities, datasets, pre-processing techniques, DNN applied, and performance evaluation metrics used for BrC Classification. The review of all these objectives is elaborated, starting from Sect. 3.1 to 3.5.

3.1 Medical imaging modalities

The review shows that the BrC classification is composed of five unique types of medical imaging modalities and their combinations known as multimodalities. The distribution of 49 chosen studies among various modalities along with publication type (article or conference paper) and number of studies is shown in Table 2. For clarity, imaging modalities can be bifurcated into colored images and grayscale images. Table 2 indicates that most of the work had been performed in either breast HP biopsy colored images or using breast X-ray grayscale images, also known as mammograms (MGs). Table 2 shows that 20 out of 49 publications (11 journal papers and 09 conference papers) are based on MG imaging modality. The main reason for the large number of publications using MGs may the availability of images. This imaging technology has been adopted for the last two decades. MG-based studies mostly explored the breast density grading or classification for two (binary) classes. Moreover, the second highest number of articles (11 journal papers and 9 conference papers) were published on HP images. In these studies, researchers usually classified BrC not only into two main cancer types (i.e., benign or malignant) but also into further subtypes of each benign and malignant BrC. However, the third highest number of papers were published for US images. By numbers, four papers (3 articles and 1 conference proceeding paper) were published using US images only. Fewer publications (one article and two conference proceeding papers) as compared with US images were found for MRI images. Moreover, very few publications used multimodalities for BrC classification. For instance, one paper was found for each combination, such as Mg with US and US with CT images. Unfortunately, none of the research publications used only CT or positron-emission tomography (PET). However, CT and PET have been used for BrC classification for many years and played a significant role (Ahn et al. 2013; Lebron et al. 2015). CT and PET images may be used if evidence shows that BrC has spread or reoccurred outside the breast. The detailed distribution of publication references, modality type, brief description of each modality used, and the number of publications as article papers or conference papers is shown in Table 2.

Table 2 Distribution of studies for various medical imaging modalities

Medical imaging modalities	Brief description	Studies		No. of studies
		Journal paper (JP)	Conference paper (CP)	
Mammogram (Mg)	Mammograms found in three forms, such as screen film mammograms (SFMs), digital mammograms (DMs), and digital breast tomography (DBT). SFMs and DMs are 2D grayscale in nature, but DBT provides multiple frames of 2D grayscale images that appear like a black-and-white video.	Arefan et al. (2015), Carneiro et al. (2017), Dhungel et al. (2017), Duraisamy and Emperumal (2017), Jaffar (2017), Kumar et al. (2017b), Qiu et al. (2017), Rouhi et al. (2015), Samala et al. (2017), Samala et al. (2018) and Sun et al. (2017)	Arevalo, González, Ramos-Pollán, Oliveira, and Lopez (2015), Bakkouri and Afdel (2017), Fonseca et al. (2015), Khan (2017), Kim et al. (2016), Kumar et al. (2017a), Leod and Verma (2016), Sert et al. (2017) and Zhang et al. (2017)	JP = 11 CP = 09
Ultrasound (US)	US images are also known as Sonograms. The US images are used in three combinations: simple 2D grayscale US images, US images along with additional additive features of shear-wave elastography (SWE) color images, and US images along with Nakagami colored images.	Han et al. (2017a), Nascimento et al. (2016) and Zhang et al. (2016)	Byra et al. (2017)	JP = 03 CP = 01
Magnetic Resonance Imaging (MRI)	MRI is used with pre and post contrast [Dynamic Contrast-enhanced (DCE) MRI] images to diagnose the BrC. Post contrast images are colored images but usually converted into grayscale to feed into ANN.	Rasti et al. (2017)	Amit et al. (2017) and Bevilacqua et al. (2016)	JP = 01 CP = 02

Table 2 continued

Medical imaging modalities	Brief description	Studies		No. of studies
		Journal paper (JP)	Conference paper (CP)	
Histopathology (HP) Images	HP Images are H&E stained colored images and subdivided into two categories: whole slide images (WSI) and image patches extracted from WSI by any expert pathologist.	Araujo et al. (2017), Bardou et al. (2018), Bejnordi et al. (2017b), Feng et al. (2018), Gandomkar et al. (2018), Han et al. (2017b), Nahid and Kong (2018), Nahid et al. (2018), Wan et al. (2017), Xu et al. (2016) and Zheng et al. (2017)	Abdullah-Al et al. (2017), Bayramoglu et al. (2017), Cao et al. (2016), Chang et al. (2017), Nahid and Kong (2017b), Nejad et al. (2017), Spanhol et al. (2017), Spanhol et al. (2016a) and Wu et al. (2016)	JP = 11 CP = 09
Multimodalities	Some studies used the combination of two modalities of grayscale images named as multimodalities for BrC classification. These combinations are Mg with MRI, and US with CT.	US with CT (Cheng et al. 2016)	Mg with MRI (Hadad et al. 2017)	JP = 01 CP = 01

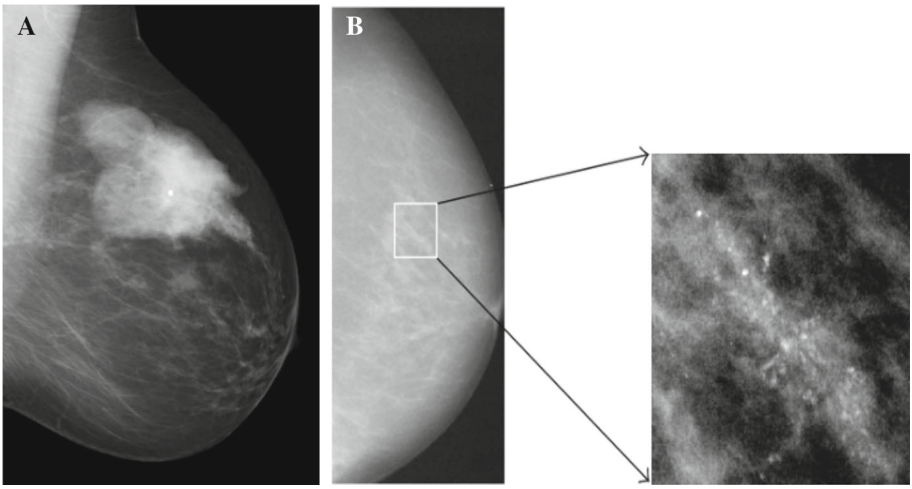


Fig. 2 a Mammogram screening: masses with areas of varying density reflecting the presence of elements which are of fat and soft-tissue density (James et al. 2016). **b** Left: a mammogram image view, right: a clustered micro-calcifications in magnified view (Jing et al. 2012)

3.1.1 Mammogram

MGs, also known as low-dose breast X-ray images, enable radiologists to investigate breast tissues for anomalies. MGs have been studied for the last two decades and usually suggested in early stages called MG screening (Fig. 2a). In MG analysis, a radiologist looks for the presence of mass (cyst or lump, Fig. 3a) and tiny deposits of calcium (specifically with irregular shape) called micro-calcifications that appear like small white spots or flecks (Fig. 2b). However, due to imaging technology advancement, MGs fall into three categories, namely, screen film mammography (SFM), full field digital mammograms (FFDM), and digital breast tomosynthesis (DBT). The traditional SFM images were used for BrC classification in many studies (Arevalo et al. 2015; Dhungel et al. 2017; Duraisamy and Emperumal 2017; Jaffar 2017; Khan 2017). Dhungel et al. (2017) proposed an integrated model for the detection, segmentation, and classification of BrC into benign or malignant masses using SFM. Similarly, Duraisamy and Emperumal (2017) proposed a novel method by using the Chan-Vese level set method to segment SFM images before classifying BrC into normal, benign, or malignant cases. The second category of MGs, FFDM (simply called digital MG or DM), is a well-adopted technology used by several researchers for BrC classification (Arefan et al. 2015; Carneiro et al. 2015, 2017; Hadad et al. 2017; Kumar et al. 2017b; Leod and Verma 2016; Qiu et al. 2017; Sert et al. 2017; Sun et al. 2017; Zhang et al. 2017). Carneiro et al. (2017) developed a holistic approach to classify unregistered DM and corresponding segmentation maps into normal, benign, or malignant breast lesions. Moreover, Qiu et al. (2017) proposed a model to classify between benign and malignant masses using DM without lesion segmentation, feature extraction, or feature selection. In the third category, the most advanced MG technology is 3D MG, known as DBT. The DBT machine takes many views by moving over the breast and integrates images together to look like a video. Nonetheless, due to limited availability of datasets, few studies used DBT for BrC classification. Kim et al. (2016) implemented a BrC classification model to discover the latent bilateral feature representations of masses using volume of interest in DBT. Similarly, Samala et al. (2018) developed an efficient

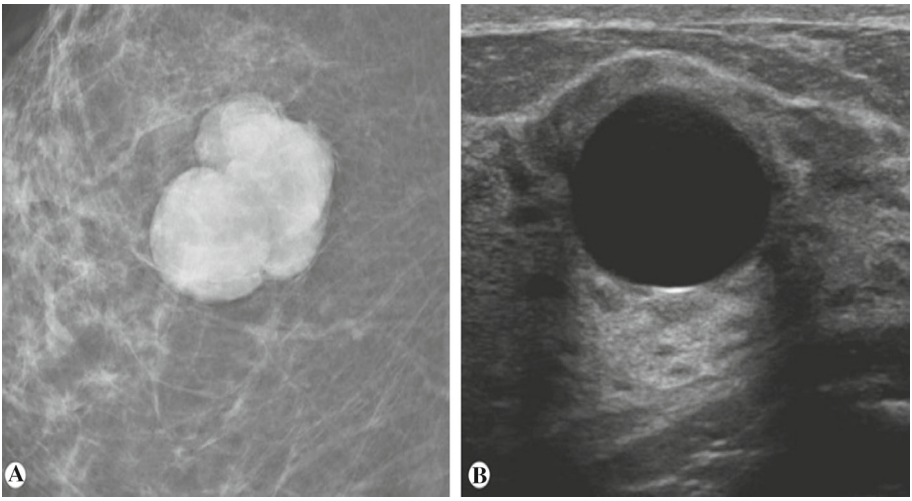


Fig. 3 **a** Well-defined rounded mass mammogram. **b** Absence of internal echoes and the posterior enhancement of the ultrasound beam are diagnostic of a cyst or lump or mass (James et al. 2016)

model by reducing the number of computations to perform BrC binary classification using all types of MGs, such as SFM, FFDM, and DBT. Apart from DBT image classification, most research used either SFM or DM. The prime advantage of the popularity of SFM is that the images are directly printed on large sheets of film; in addition, it is a more cost-effective and frequently available imaging technology than FFDM and DBT. By contrast, FFDM images are easier to view, stored, print, and manipulate using a desktop computer. Therefore, digital MG images can be viewed on a computer screen using many options, such as zooming, contrast enhancement, and highlighting the affected regions. Hence, due to efficient processing of digital images, most of the recent public datasets utilized by researchers are digital MGs instead of SFM. However, researches started to used DBT because of many reasons; for instance, DBT may give a clear view of the breast from multiple angles to diagnose cancer with higher confidence and reduce the chance of follow-up testing as compared with FM or DM (Radiological Society of North America 2018). Moreover, the availability of a large number of images per subject in video form provides better analysis opportunity to reduce the FNs in MGs. Table 3 lists the detailed advantages and limitations of MGs. Regardless of MG diagnosis popularity, some cases may have dense tissues (bulky patient) or thick breast skin, such as in younger women, rendering the cancerous area almost invisible. Hence, macro-classification can be overlooked or misinterpreted during image analysis and may increase the FN rate. When image analysis is suspicious, the doctor may suggest some complementary tests, such as US, CT, PET, MRI, or biopsy, to acquire a detailed view of suspicious breast regions.

3.1.2 Ultrasound

US images are also known as sonograms. Breast US (Fig. 3b) is an imaging test that sends high-frequency sound waves into breast and converts into images without radiation involvement unlike MGs and MRI. Apart from breast test, US test can help diagnose anomalies, such as pain, swelling, and infections into human body internal organs, including baby in mothers'

Table 3 Studies, imaging modality, strengths, weakness, and applications of various medical imaging modalities used in BrC classification

Imaging modality	Applications	Limitations
Mammogram (Mg)	Most studies employed SFM (Arevalo et al. 2015; Dhungel et al. 2017; Duraisamy and Emperumal 2017; Jaffar 2017; Khan 2017; Samala et al. 2018) or DM (Arefan et al. 2015; Carneiro et al. 2015; Carneiro et al. 2017; Fonseca et al. 2015; Hadad et al. 2017; Kumar et al. 2017b; Leod and Verma 2016; Qiu et al. 2017; Samala et al. 2018; Sert et al. 2017; Sun et al. 2017; Zhang et al. 2017) instead of DBT (Kim et al. 2016) for BrC diagnosis	Micro-calcifications are very small, isolated, with various sizes, shapes, dispersed, looks similar to their surroundings; thus, they cannot be identified in mammograms from high-frequency noise
	Relative to HP, DM technology is an efficient, highly standardized, and cost-effective method to capture, store, and process images	Several pre-processing tasks are needed before performing classification because the presence of many factors, artifacts, and structure, such as film emulsion error, digitization artifacts, fibrous strands, borders of breast, and hypertrophied lobules, causes misinterpretation
	Needs less expertise and professional knowledge to diagnose and categorize an image as compared with HP	High breast density complicates visualization of cancer in mammograms. However, the deeper breast are usually prone to cancer, and a radiologist can overlook or misinterpret the findings (Elmore et al. 2009). Hence, US or MRI can be preferred for a dense breast
	A large variety of computer-aided diagnostic (CAD) system are available to serve as a second opinion	
	DBT shows a significantly higher rate of screen-detected cancer compared with DM screening (Hofvind et al. 2018).	
Ultrasound (US)	Very few articles are found using US images (Byra et al. 2017; Cheng et al. 2016; De S. Silva et al. 2015; Han et al. 2017a; Nascimento et al. 2016; Zhang et al. 2016) for breast cancer diagnosis	Poor image quality is usually observed when a great amount of tissues is examined by US (Radiological Society of North America 2018; Ultrasound 2018)
	Images are taken in a real-time fashion. Hence, a breast lesion can be viewed from multiple angles, reducing FN rate in diagnosis	SWE images can cause misinterpretation if probe is pressed harder (Barr 2012; Youk et al. 2017)
	Widely available, extremely safe (noninvasive and no exposure to radiation) technology. Hence, preferred for a routine checkup among pregnant women	Solely single Nakagami parameters cannot distinguish between benign and malignant tissues (Tsui et al. 2008)

Table 3 continued

Imaging modality	Applications	Limitations
	Can detect invasive cancer areas that can be further used for biopsy, known as US-guided biopsy	The shadowing effect due to high attenuation makes the tumor contour unclear; thus, selecting the proper ROI and estimating tumor Nakagami parameters are difficult (Tsui et al. 2008)
	Additional features, such as color-coded SWE images and Nakagami parametric images, can be captured along with traditional US images to identify breast lesion ROI.	
Magnetic Resonance Imaging (MRI)	An MRI scan does not use potentially harmful ionizing radiation like CT scans and X-rays	MRI can still miss some tumors that a mammogram can detect. Thus, MRI is usually suggested in addition to a mammogram test
	MRI images show more details of tissues (e.g., soft tissues of breast) than CT scans (Tessa and Keith 2018)	An MRI is not generally recommended for women who are pregnant (MFMER 2018)
	MRI can identify suspicious areas that can be further used for biopsy, known as MRI-guided biopsy	May increase body temperature during long MRI (Tessa and Keith 2018)
	DCE MR imaging uses contrast agents to show clear and detailed view of affected breast regions	Contrast agents usually injected to enhance MRI images may create allergies or any complications, especially for kidney patients (MFMER 2018)
Histopathology (HP) Images	Many studies employed HP images (Abdullah-Al et al. 2017; Araujo et al. 2017; Bardou et al. 2018; Bayramoglu et al. 2017; Bejnordi et al. 2017b; Cao et al. 2016; Chang et al. 2017; Feng et al. 2018; Gandomkar et al. 2018; Han et al. 2017b; Nahid and Kong 2018; Nahid and Kong 2017b; Nahid et al. 2018; Nejad et al. 2017; Spanhol et al. 2017; Spanhol et al. 2016a; Wan et al. 2017; Wu et al. 2016; Xu et al. 2016; Zheng et al. 2017)	Breast biopsy is an invasive method and thus has higher risks than other modalities
	HP images can be used in two forms, namely, whole slide images or ROI extracted from WSI	Manual analysis of HP images is time intensive and requires high expertise; it depends on the professional experience and knowledge of a pathologist (Farahani et al. 2015)
	Images are colored, can diagnose multiple types of cancers (Han et al. 2017b) instead of detecting malignancy only (Qiu et al. 2017) through grayscale imaging modalities. Ultimately, it leads to better prognosis and treatment at early stage of BrC	Manual image inspections are tedious; thus, analysis reports are also affected by factors, such as fatigue and reduced pathologist attention (Spanhol et al. 2016b)

Table 3 continued

Imaging modality	Applications	Limitations
	In-depth study of BrC tissues is possible. HP images enable to provide more confident diagnosis results than any other imaging modalities	HP image appearance variability causes misdiagnosis due to variability, different lab protocols, fixation, sample orientation in the block, human expertise in tissue preparation, microscopy maintenance, and color variation due to differences in staining procedures (McCann et al. 2015)
	Multiple ROI images can be created from WSI, which results in less probability to miss the cancer tissue detection, especially in early stage, and reduces FN rate	For multiclass classification, traditional machine learning algorithms produce poor results because of high variability among images of the same cancer subtype (McCann et al. 2015). Hence, complex methods and high computational resources are required to improve computer-aided diagnosis.
	Images can be shared electronically to obtain opinion from experts, especially for borderline cases, where two cancer types are hard to characterize	
	Can be stored for a long time for future analysis or reference	

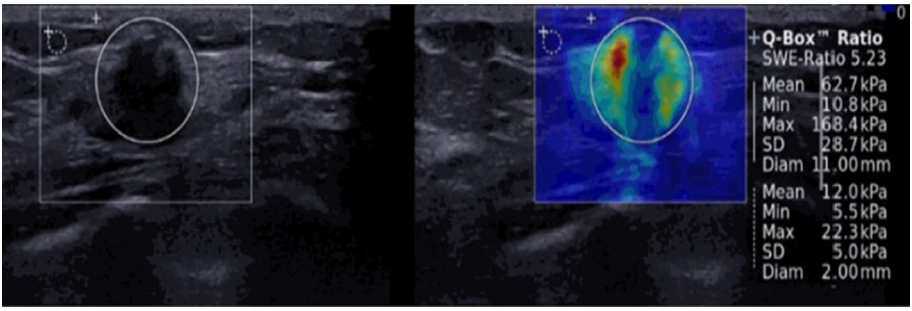


Fig. 4 Left side US image (B-Mode). Shear-wave elastography image on right side, shows an irregular mass in red color, known as heterogeneous elasticity. The statistical parameters (e.g. mean, minimum, maximum and etc.) of ROI (a large circle) are calculated (Youk et al. 2017)

womb, brain, lung, heart, and hips. In addition, US can help perform breast needle biopsy (Sect. 3.1.4) for the intrinsic analysis of breast tissues. As per common clinical practices, US is not used like MG as its own for only breast screening purpose. Therefore, US may be the best approach to find abnormalities in MG or physical examination (such as benign, a noninvasive cancer) in the form of solid lump (mass) or fluid-filled regions (cysts) (Cheng et al. 2016; De Silva et al. 2015; Han et al. 2017a; Nascimento et al. 2016). However, US cannot distinguish a cancerous mass from calcifications. Some researchers found that breast US is the better choice to diagnose BrC, especially when a MG is unable to highlight BrC lesions clearly, in young subjects with thick, fatty, or bulky breast skin. Detailed advantages and limitations of using US images are discussed in Table 3. Cheng et al. (2016) deployed a model to extract distinct features automatically from breast US images directly to perform accurate breast lesion classification as benign or malignant. Similarly, Nascimento et al. (2016) extracted hand-engineered morphological features from breast US images and fed them into ANN for BrC binary classification (benign or malignant). Moreover, due to new developing imaging technologies, US has been equipped with more advanced features, such as US with shear-wave elastography (SWE) (Fig. 4) and US with Nakagami images (Fig. 5). Elastography is a recently developed US technique used to visualize and measure tissue elasticity. Elastographic images are based on tissue stiffness or hardness (such as in liver or breast) and used to differentiate between benign and malignant lesions (Youk et al. 2017). It is a supportive parameter to US and adopted to quantify tumor grade by using a standardized color scheme. Hence, Zhang et al. (2016) used US SWE images to learn features directly by using a deep belief network to classify images (with higher accuracy) into benign or malignant BrC. Moreover, US images are used with Nakagami images for BrC analysis. US Nakagami parametric images are used with Nakagami distribution to model echo amplitude distribution to represent tissue characteristics (Tsui et al. 2016). These color-coded images can be captured along with traditional US images. The color-coded US images enable radiologists to quantify the stiffness or hardness of tissues. Hence, SWE and Nakagami features play an additional role to enhance BrC classification diagnosis. However, very few studies used the new US technology. Byra et al. (2017) developed a model and extracted the scattering properties of breast tissues from parametric maps of Nakagami images to perform BrC classification by using a convolutional neural network (CNN). Data collection, particularly the difficulty in collecting a large number of medical images from any medical institution, may be one of the reasons for the few publications.

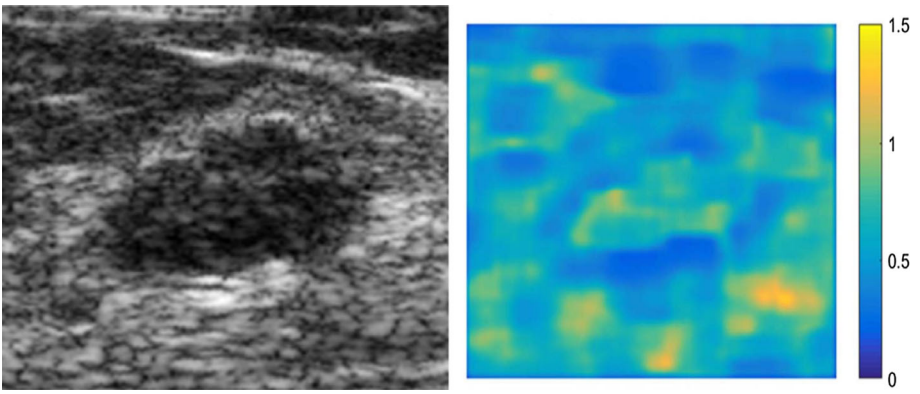


Fig. 5 Left US image (B-mode) of a lesion reconstructed using the RF data and on right side corresponding Nakagami map (Byra et al. 2017)

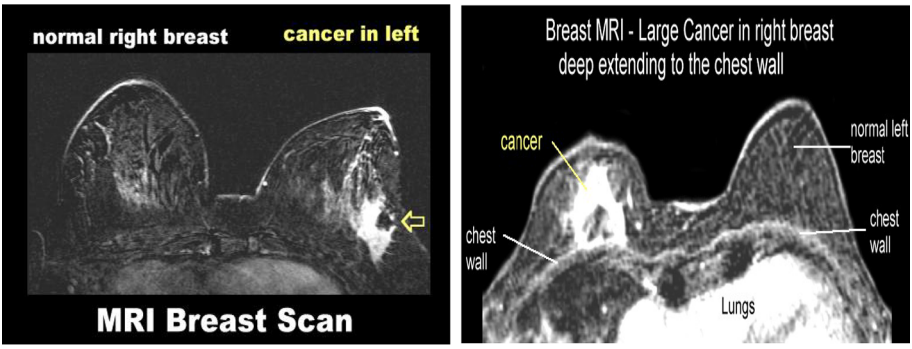


Fig. 6 Samples of breast MRI images (Breast Cancer Imaging 2018)

3.1.3 Magnetic resonance imaging

MRI is a diagnostic technology that uses magnetic fields and radio waves to capture a detailed image of the body's soft tissue, such as breast (Fig. 6), liver, or lung, and bones. Therefore, breast MRI images can show more clear views of breast soft tissues than MGs, US, or CT images (Tessa and Keith 2018). Table 3 lists the advantages and limitations of MRI. Furthermore, MRI can identify suspicious areas that can be used for breast biopsy, known as MRI-guided biopsy (Sect. 3.1.5). MRI machine captures many breast images of single subject and combines together as a detailed view. MRI is usually requested once the cancer has been diagnosed and the doctor wants to obtain detailed information about the extent of the disease (MFMER 2018). However, very few studies used MRI to classify BrC (Amit et al. 2017; Bevilacqua et al. 2016; Rasti et al. 2017) possibly because of the unavailability of public datasets. Bevilacqua et al. (2016) extracted features from MRI-segmented images and inputted them into an ANN for benign and malignant BrC identification. Analogously, Amit et al. (2017) extracted regions of interest (ROIs) from breast MRI images and inputted them into a CNN for multi-class classification. To enhance image quality, a contrast agent is usually injected into the human body before dynamic contrast enhanced MRI (DCE-MRI). This procedure can produce colored parametric views along with contrast enhanced grayscale

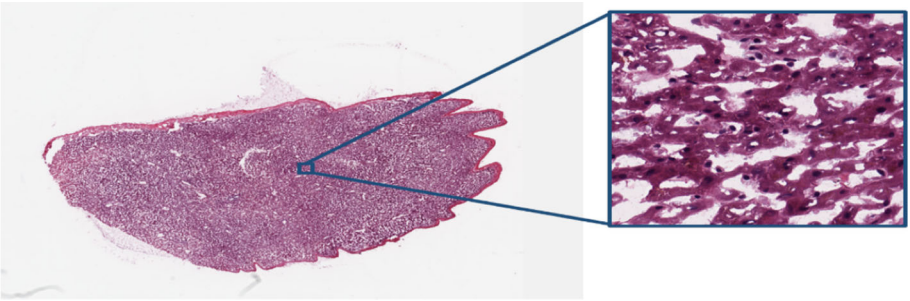


Fig. 7 Histopathology WSI is shown on the left at low magnification and a cropped region is shown on the right at high magnification (Liu et al. 2017)

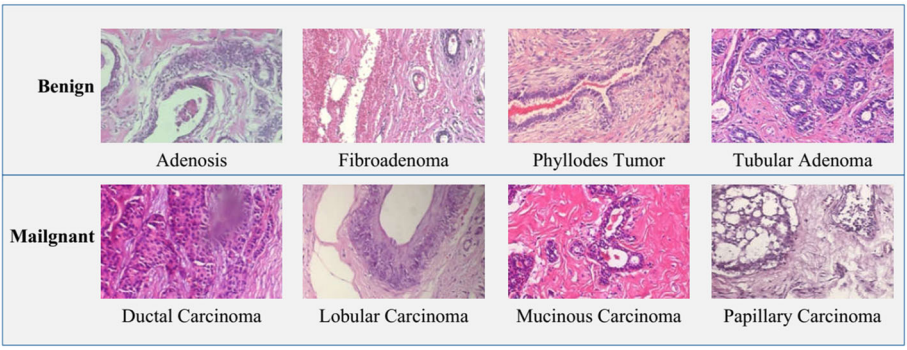


Fig. 8 Histopathology image patches showing eight subtypes of breast cancer (Spanhol et al. 2016b)

images to provide detailed information about cancerous tissues (Moon et al. 2009). However, only one study benefitted from DCE-MRI for BrC classification. Rasti et al. (2017) employed a deep learning ensemble CNN model to classify breast tumors using segmented DCE-MRI images of an exclusive dataset.

3.1.4 Histopathologic images

In HP biopsy imaging, tissue samples are collected from an abnormal region of the breast and fixed across glass microscope slides. These slides are stained by using hematoxylin–eosin (HE) and examined under a microscope by a pathologist for cancerous tissues diagnosis. Moreover, these stained slides are scanned and converted into digital colored images called WSIs (Fig. 7). Expert pathologists usually extract ROI patches from WSI with various zooming factors (Fig. 7) to diagnose multiple subtypes of noninvasive cancer (benign) or invasive BrC (malignant) (Fig. 8), which is impossible by using grayscale images. Due to tissue level image analysis, apart from BrC diagnosis, biopsy imaging is a gold standard for many types of cancers, including liver, lung, and bladder cancer (Rubin et al. 2008). Therefore, many researchers employed HP images to classify BrC multi-class accurately (Abdullah-Al et al. 2017; Araujo et al. 2017; Bardou et al. 2018; Bayramoglu et al. 2017; Bejnordi et al. 2017b; Cao et al. 2016; Chang et al. 2017; Feng et al. 2018; Gandomkar et al. 2018; Han et al. 2017b; Murtaza et al. 2019; Nahid and Kong 2017b, 2018; Nahid et al. 2018; Nejad et al. 2017; Spanhol et al. 2016a, 2017; Wan et al. 2017; Wu et al. 2016; Xu et al. 2016; Zheng

et al. 2017). For instance, Han et al. (2017b) used HP images to classify BrC into eight types. Araujo et al. (2017) used HP images to develop a model that classifies BrC into four subtypes. The above listed studies reported that the use of HP images is beneficial for specific subtypes of benign or malignant BrC. Automatic breast classification through HP images has several advantages over MGs and other imaging modalities (Table 3). For instance, HP images enable the classification of BrC into many subtypes instead of binary classes and the monitoring of treatment effects, whereas WSI images allow the creation of a large number of ROI images, which are required to train DNN models. Images can be shared electronically to obtain the opinion of any far distant expert pathologist and thus form an accurate diagnosis. Although HP images are authentic for automatic BrC classification, such images have some drawbacks in automatic image classification. For instance, biopsy is an invasive method. In addition, a long time is needed to create digital images from collected biopsy samples, and high expertise is needed to distinguish between subtypes of BrC. Moreover, color variation is high because of the staining process, lab protocols, and scanner brightness in the development of HP images, which complicate training a multi-class DNN model efficiently, especially when using borderline cases. Details of the imaging modalities used in previous studies are listed Table 3.

3.1.5 Multimodalities

Apart from classifying BrC by using a single medical imaging modality, few researchers preferred to use at least two different imaging modalities (Fig. 9). Hadad et al. (2017) trained various classification models by using two modalities, namely, MGs and MRIs. This study performed binary classification by identifying a breast image possessing either mass or non-mass regions. Moreover, images were classified as normal, benign, or malignant by (Khan 2017) through multimodalities, such as MGs with US images. Many imaging modalities for BrC classifications are usually adopted when the size of the collected exclusive dataset is small. Moreover, a model trained on multi-site, multi-datasets using multi-modalities is highly robust to classify real-life images. Eventually, the performance of the BrC classifier is unaffected by the images captured on various machines, different imaging protocols, and the environment for handling images. Hence, such type of models is trustworthy to be implemented in real life.

3.2 Breast cancer classification dataset analysis and review

This section elaborates a thorough analysis of public datasets that were utilized in various studies for BrC classification. Table 4 shows that eight public datasets were employed for BrC classification, namely, Breast Cancer Data Repository (BCDR), Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM), Digital Database for Screening Mammography (DDSM), INBreast, Mammographic Image Analysis Society (MIAS)/mini-MIAS, UCI Machine Learning Repository, Bio-Imaging Challenge 2015 Breast Histology (BICBH), and Breast Cancer Histopathological Image (BreakHis). Out of 49, 28 articles utilized public datasets, usually based on MG, US, or HP images. By contrast, 21 out of 49 studies employed exclusive datasets. In exclusive datasets, imaging modalities that are not publicly available similar to CT scan images were also used. Public datasets provided more annotated images than exclusive datasets. Hence, researchers can prepare BrC classification models by comparing the performance of developed classification models. Therefore, the model tested on public datasets is more reliable than the models tested

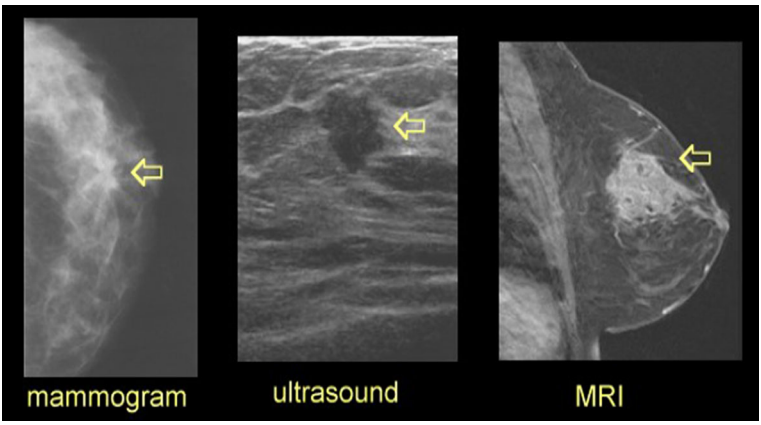


Fig. 9 Multimodalities used for BrC classification. Left image is a mammogram showing a solid mass. Center image is US showing stuff tissues as black. Right image is MRI providing a clear view of breast mass (Breast Cancer Imaging 2018)

Table 4 List of publically available datasets and corresponding URL

#	Dataset Name	URL
1	BCDR (Moura and López 2013)	https://bcdr.ceta-ciemat.es/information/about
2	CBIS-DDSM (Clark et al. 2013; Rebecca Sawyer Lee et al. 2016)	https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM
3	DDSM (Chris Rose 2006)	http://marathon.csee.usf.edu/Mammography/Database.html
4	INBreast (Moreira et al. 2012)	http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database
5	MIAS (Suckling et al. 2015)	https://www.repository.cam.ac.uk/handle/1810/250394
6	mini-MIAS (Suckling et al. 1994)	http://peipa.essex.ac.uk/info/mias.html
7	UCI (Dua 2017)	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
8	BICBH (Araújo et al. 2017)	https://rdm.inesctec.pt/dataset/nis-2017-003
9	BreakHis (Spanhol et al. 2016b)	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

on exclusive datasets. Regardless of database type (exclusive or public) at the abstract level, grayscale (e.g., Mg, US, and MRI) or colored images (e.g., HP images) are used for BrC classification. Moreover, most studies performed binary classification, and very few studies focused on multi-class problems for BrC classification. By contrast, some studies performed breast density grading (Cao et al. 2016; Wan et al. 2017) into three classes, namely, low, high, and medium grade. Detailed analysis of public datasets used for BrC classification is given in Fig. 5.

Table 5 shows the dataset name and type of imaging modality along with number of images, number of patients, number of classes, and class labels for each dataset. This table also shows the reference of studies in which a particular dataset was used and the number

of publications per dataset. The investigation of dataset reveals that most previous research used MG datasets and usually addressed either binary classification (benign or malignant) or tertiary classification (normal, benign, and malignant) of BrC. In this regard, 23 out of 28 studies used MG datasets. Out of 23 MG-based studies, nine (Bakkouri and Afdel 2017; Carneiro et al. 2015, 2017; Jaffar 2017; Kumar et al. 2017b; Leod and Verma 2016; Rouhi et al. 2015; Samala et al. 2017; Sert et al. 2017) used DDSM datasets and six (Arefan et al. 2015; Duraisamy and Emperumal 2017; Jaffar 2017; Khan 2017; Kumar et al. 2017a; Rouhi et al. 2015) used MIAS datasets. Moreover, the MGs of both INBreast and BCDR datasets each utilized a maximum of four studies (Arevalo et al. 2015; Bakkouri and Afdel 2017; Carneiro et al. 2017; Dhungel et al. 2017; Duraisamy and Emperumal 2017; Khan 2017; Kumar et al. 2017a). Meanwhile, only one study (Kumar et al. 2017a) classified MGs of CBIS-DDSM datasets. However, multimodality (US and MG)-based BCDR-F03 datasets were used by two studies (Arevalo et al. 2015; Duraisamy and Emperumal 2017) for BrC classification. Apart from MG-based limited (two or three class label) classification, HP images played a prominent role to solve multi-class (up to eight subtypes) problems for BrC classification. In this respect, 14 out of 27 studies performed classification by using BreakHis datasets, as shown in Table 5. Unfortunately, most studies performed binary classification, and very few obtained better results to solve multi-class problems. Moreover, only one study used Bio-Imaging Challenge 2015 Breast Histology datasets and tackled the multi-class BrC issue. For clarity, Table 5 shows that the total count of studies is greater than 27 because several studies employed more than one dataset. Thus, their count is added in more than one category. As per our review, the most widely used and authentic dataset in MG, US, and HP imaging modalities are DDSM, BCDR, and BreakHis, respectively, because these datasets contain a large number of images of many patients, which are required to train DNN classification models with confidence. Unlikely, no publicly available datasets have been employed for CT, MRI, PET modalities. Hence, the unavailability of online datasets might be a reason or publically available datasets may contain an insufficient number of images for training a DNN-based BrC classification model.

3.3 Pre-processing

This section covers the pre-processing techniques adopted for medical image multimodalities in BrC classification. In general, BrC image pre-processing tasks involve augmentation, ROI extraction, scaling, image normalization, and enhancement to remove artifacts or cropping, stain normalization, feature reduction, and image registration. However, the use of raw images (without pre-processing) usually distracts the classification model and may lead to high misclassification rate.

Figure 10 represents the distribution of pre-processing tasks performed in selected studies. The total count in Fig. 10 is more than 49 because one study may have performed more than one pre-processing task. Figure 10 shows that the majority of studies (32 out of 49) adopted image augmentation as a pre-processing technique to increase the number of images synthetically. Such a frequent use of image augmentation may be because the annotated medical images are mostly not found in a large quantity. Moreover, the second highest number of studies (21 out of 49) extracted ROI from BrC images; thus, DNNs can only learn representations related to the normal and abnormal regions instead of using the entire image, which usually contains irrelevant information. The same number of studies (20 out of 49) reduced the size of images before they are fed into DNNs. Rescaling is an essential task when images are directly fed into DNNs, such as CNNs. However, fewer studies (11 out of 49) employed

Table 5 Detailed analysis of public datasets used in breast cancer classification

Imaging modality	Dataset name	# Images	# Patients	# Classes	Class labels	Study references	# Studies**
Mammograms	BCDR	1734	–	3	Normal, Benign, Malignant	Khan (2017)	23
	BCDR-F03	736	344	2 or 10	(Benign, Malignant) or (Normal, Benign-calcification, Malignant-calcification, Benign-circumscribed masses, Malignant-circumscribed masses, Speculated masses, Ill-defined masses, Benign-architectural distortion, Malignant-architectural distortion, Asymmetry)	Arevalo et al. (2015), Duraisamy and Emperumal (2017)	
	CBIS-DDSM	4067	–	2	Benign, Malignant	Kumar et al. (2017a)	
	DDSM	10480	2620	2	Benign, Malignant	Bakkouri and Afdel (2017), Carneiro et al. (2015), Carneiro et al. (2017), Jaffar (2017), Kumar et al. (2017b), Leod and Verma (2016), Rouhi et al. (2015), Samala et al. (2017) and Sert et al. (2017)	
	INBreast	419	115	2 or 3	(Benign, Malignant) or (Normal, Benign, Malignant)	Carneiro et al. (2017), Dhungel et al. (2017), Kumar et al. (2017a)	
	MIAS/Mini-MIAS	322	161	2	Benign, Malignant	Arefan et al. (2015), Duraisamy and Emperumal (2017), Jaffar (2017), Khan (2017), Kumar et al. (2017a) and Rouhi et al. (2015)	

Table 5 continued

Imaging modality	Dataset name	# Images	# Patients	# Classes	Class labels	Study references	# Studies**
Mammograms and Ultrasound Images	BCDR	3703	1010	2	Benign, Malignant	Bakkouri and Afdel (2017)	01
Histopathology Images	BICBH	269	–	4	Normal, Benign, In situ carcinoma, Carcinoma	Araujo et al. (2017)	01
	BreakHis	7909	82	2 or 8	Four Benign Tumours (Adenosis, Fibroadenoma, Phyllodes tumor, Tubular adenoma), Four Malignant Tumours (Ductal carcinoma, Lobular carcinoma, Mucinous carcinoma and Papillary carcinoma	Abdullah-Al et al. (2017), Bardou et al. (2018), Bayramoglu et al. (2017), Chang et al. (2017), Feng et al. (2018), Gandomkar et al. (2018), Han et al. (2017b), Nahid and Kong (2018), Nahid and Kong (2017b), Nahid et al. (2018), Nejad et al. (2017), Spanhol et al. (2017) and Spanhol et al. (2016a)	14

**Here the total count of studies is greater than 28. This is because several studies employed more than one datasets thus their count is added in more than one categories

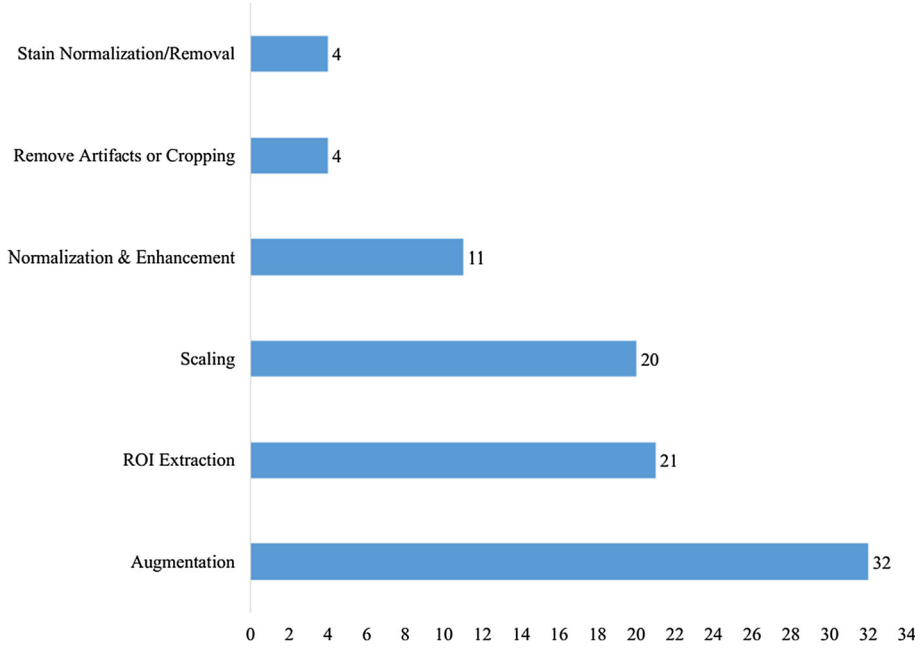


Fig. 10 Distribution of studies among pre-processing tasks performed for BrC classification

various pre-processing techniques to image normalization and enhancement before BrC classification. This approach reduces high- or low-intensity noise to make all images uniform and enables DNNs to learn accurate feature representations for normal and abnormal tissues of BrC. Conversely, few studies (4 out of 49) used pre-processing techniques to remove artifacts (e.g., labels, wedges, opaque markers, pectoral muscles, while strips, and border) from images. Hence, such pre-processing techniques eliminate non-breast regions from the image before performing BrC classification. These studies performed artifact removal because it is only required in specific imaging modalities, such as MG, US, and MRI. Finally, the same number of studies (4 out of 49) adopted stain normalization pre-processing techniques. A stain normalization technique helps reduce the inconsistencies commonly found in HP biopsy BrC images.

3.3.1 Augmentation

Augmentation increases the number of instances (BrC images) artificially. In general, a DNN model requires a large number of images to be trained to produce reliable results. Indeed, image augmentation is required when the target dataset does not contain enough number of images for training a DNN model properly. This review identified four types of augmentation techniques, of which geometric transforms, noise addition, and patch extraction were implemented over breast images directly and synthetic minority over-sampling technique was adopted for feature vector data (manually extracted from breast images) before feeding to any ANN. For instance, some studies (Amit et al. 2017; Arevalo et al. 2015; Bakkouri and Afdel 2017; Bardou et al. 2018; Bayramoglu et al. 2017; Bejnordi et al. 2017b; Byra et al. 2017; Carneiro et al. 2017; Chang et al. 2017; Dhungel et al. 2017; Duraisamy and Emperumal 2017; Gandomkar et al. 2018; Hadad et al. 2017; Han et al. 2017a, b; Jaffar

2017; Kim et al. 2016; Nejad et al. 2017; Rasti et al. 2017; Samala et al. 2017, 2018; Sert et al. 2017; Zhang et al. 2017; Zheng et al. 2017) utilized geometric transform (e.g., rotation at various angles, flip horizontally and vertically). Meanwhile, other studies (Araujo et al. 2017; Cheng et al. 2016; Duraisamy and Emperumal 2017; Feng et al. 2018; Gandomkar et al. 2018; Kumar et al. 2017a; Spanhol et al. 2016a, 2017; Xu et al. 2016; Zheng et al. 2017) extracted many patches from the original image. Moreover, patches are extracted by using three strategies, namely, random number of patches (Spanhol et al. 2016a, 2017), patches with 50% overlapping (Araujo et al. 2017; Spanhol et al. 2016a), and patches with no overlapping (fixed size window) (Cheng et al. 2016; Duraisamy and Emperumal 2017; Feng et al. 2018; Gandomkar et al. 2018; Kumar et al. 2017a; Xu et al. 2016; Zheng et al. 2017). Furthermore, augmentation by using noise addition or color variation was adopted in previous studies (Bejnordi et al. 2017b; Chang et al. 2017) to train a model robustly to handle noisy image while performing class label prediction. For instance, Chang et al. (2017) added a random distortion to original images while creating new synthetic images.

3.3.2 Image region of interest extraction

An original breast image may contain many regions of normal and abnormal tissues, and segregation of these regions is known as ROI extraction. ROI extraction has two major advantages. First, it increases the number of training and testing images required for DNNs. Second, it supports DNNs to learn only normal and abnormal regions instead of irrelevant regions. As mentioned in Sect. 3.3, many studies (Amit et al. 2017; Arefan et al. 2015; Arevalo et al. 2015; Bevilacqua et al. 2016; Cao et al. 2016; Cheng et al. 2016; Duraisamy and Emperumal 2017; Feng et al. 2018; Fonseca et al. 2015; Han et al. 2017a; Khan 2017; Kim et al. 2016; Kumar et al. 2017b; Leod and Verma 2016; Nascimento et al. 2016; Rasti et al. 2017; Rouhi et al. 2015; Samala et al. 2017, 2018; Wan et al. 2017; Zheng et al. 2017) extracted ROIs from the original image before BrC classification. For instance, Samala et al. (2018) extracted thousands of ROI from 3D Mg DBT images. Similarly, Rouhi et al. (2015) cropped the ROI of abnormal tissues and mass regions before BrC classification.

3.3.3 Scaling

Scaling or resizing is an important pre-processing task applied on images before they are fed directly into a DNN. Image scaling or interpolation occurs when an image is resized from one pixel grid to another. It increases or decreases the number of pixels by remapping. Most of the selected studies (Abdullah-Al et al. 2017; Arefan et al. 2015; Bakkouri and Afdel 2017; Bayramoglu et al. 2017; Carneiro et al. 2017; Chang et al. 2017; Cheng et al. 2016; Dhungel et al. 2017; Duraisamy and Emperumal 2017; Fonseca et al. 2015; Gandomkar et al. 2018; Han et al. 2017b; Jaffar 2017; Kim et al. 2016; Kumar et al. 2017a; Nejad et al. 2017; Spanhol et al. 2016a; Wan et al. 2017; Xu et al. 2016; Zhang et al. 2016) adopted interpolation methods, such as nearest neighborhood, bilinear, or bi-cubic. For instance, Dhungel et al. (2017) adopted the bi-cubic interpolation method to rescale images before feeding into a five-layered CNN for BrC binary classification. Zhang et al. (2016) utilized bilinear interpolation to resize US BrC images for binary classification. However, Bakkouri and Afdel (2017) adopted Gaussian pyramids to reduce and expand image size using MG images before classification.

3.3.4 Normalization and enhancement

Medical image acquisition and digitization are affected by involving color and light conditions. Hence, different color and light conditions affect all pixel values present in an image. To overcome these issues, researchers adopted many techniques, which can be broadly divided into two categories: global or local image normalization and enhancement techniques. Global image normalization and enhancement techniques perform the same operation on all pixels of images, such as histogram, mean, and median contrast/intensity normalization. By contrast, local image normalization and enhancement techniques perform an operation on any pixel depending on the contrast or intensity of the neighboring pixels. DNNs usually perform better when the input images are normalized and decorrelated because these properties help gradient-based optimization and learning (Jarrett et al. 2009). As mentioned in Sect. 3.3, some studies (Arefan et al. 2015; Arevalo et al. 2015; Bejnordi et al. 2017b; Duraisamy and Emperumal 2017; Han et al. 2017a; Jaffar 2017; Khan 2017; Nejad et al. 2017; Rasti et al. 2017; Sert et al. 2017) utilized the techniques to improve image quality before feeding into any type of DNN for BrC classification. For instance, some studies (Bejnordi et al. 2017b; Duraisamy and Emperumal 2017; Nejad et al. 2017) employed global contrast normalization by using mean filters to solve the multi-class BrC classification problem. Khan (2017) removed US image spackle noise and blurring effect by adopting Wiener and adaptive filters (e.g., mean, variance and spatial correlations). The author reduced the impulse noise usually found in US images by using mean filter and wavelet shrinkage. Moreover, image local contrast enhancement was performed by contrast limited adaptive histogram equalization (CLAHE). However, Jaffar (2017) adopted a hybrid of bilateral filter with log transformation to preserve edges while performing image normalization.

3.3.5 Removing artifacts

Artifacts are removed from breast images to eliminate all non-breast regions from the original raw image. Some imaging modalities, such as MG, US, and MRI, possess many artifacts (e.g., labels, wages, opaque markers, white strips/borders, thorax, lungs, chest wall, and pectoral muscle) (Fig. 11) that should be removed before starting the BrC classification task. Few studies (Arefan et al. 2015; Bayramoglu et al. 2017; Bevilacqua et al. 2016; Sert et al. 2017) adopted pre-processing techniques to remove non-breast regions because they may not use the entire raw image but breast image ROIs for classification. For instance, Arefan et al. (2015) extracted non-breast regions from MGs in two steps, namely, the creation of binary images created by pixel thresholding to detect connected areas and the deletion of small disconnected areas. Hence, the breast region is separated from the rest of the background before performing breast density multi-classification, such as fatty, glandular, or dense breast. Bevilacqua et al. (2016) classified breast US images after eliminating the thorax part by considering a geometric parabola that follows rib cage border. Moreover, Sert et al. (2017) removed white strips found at MG borders by thresholding the intensity value to 150.

3.3.6 Stain normalization

In digital pathology (DP) labs, the preparation of HP biopsy images involves different chemical, stains, lighting effects, and scanners to develop digital images from collected breast tissue samples. The inconsistencies in HP images may be introduced by using different chemicals for staining, concertation of colors, or different scanners from many vendors. Moreover,

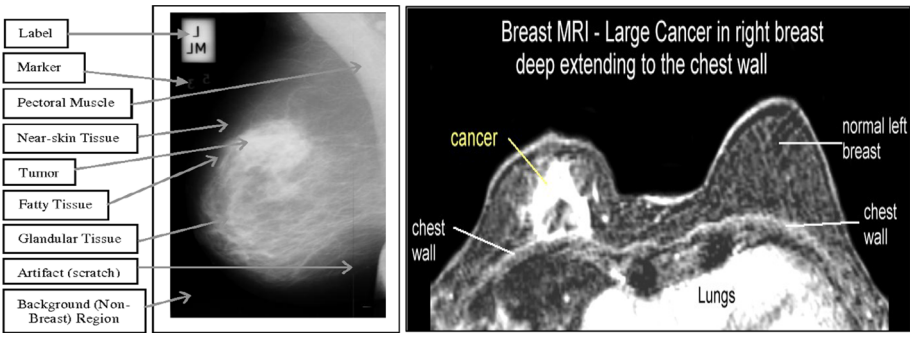


Fig. 11 Shows different artifacts in mammogram (left image) and in MRI (right image) (Breast Cancer Imaging 2018; Saidin et al. 2012)

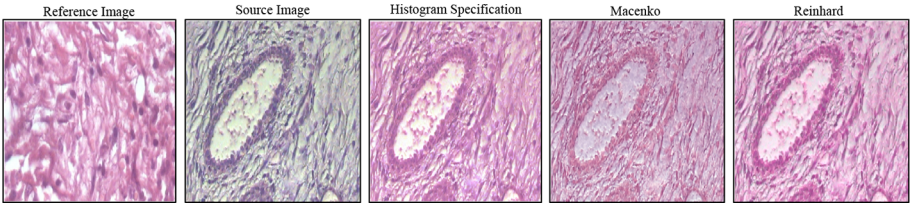


Fig. 12 Source image stain normalized by using reference image through three techniques

these factors may create major inconsistencies in images of two patients even if images are prepared in the same DP lab. To eliminate these inconsistencies, previous studies used RGB histogram specification, Reinhard's (Reinhard et al. 2001), Macenko's (Macenko et al. 2009), and Khan's methods (Khan et al. 2014) to normalize the HP images before classification (Fig. 12). Many studies (Abdullah-Al et al. 2017; Araujo et al. 2017; Cao et al. 2016; Gandomkar et al. 2018; Wan et al. 2017; Zheng et al. 2017) employed satin normalization or removal techniques before proceeding toward BrC classification. For instance, Abdullah-Al et al. (2017) used Retinex operation to perform a non-linear transform to normalize illumination. Wan et al. (2017) adopted Khan's method to perform a non-linear mapping-based stain normalization. Gandomkar et al. (2018) employed two stain normalization methods, namely, histogram specification-based method and Reinhard's method; the latter uses mean and standard deviation to match RGB channels with the reference image. Furthermore (Zheng et al. 2017) removed the color stain by using the color deconvolution method proposed by Ruifrok and Johnston (2001). This method separates the color information acquired by H&E staining. It determines the contribution of all applied stains according to the stain-specific RGB absorption.

3.4 Artificial neural network types used in BrC classification

The human brain consists of more than 10 billion interconnected neurons. Using chemical reactions, each neuron obtains information, processes it, and responds accordingly. Similarly, artificial neuron (AN) mimics the simple methods of mammal neuron, see AN in Fig. 13. The first simplified artificial neuron was introduced by (McCulloch and Pitts 1943). A group of ANs forms a layer, and a group of layers creates an ANN (Fig. 13). An ANN is an

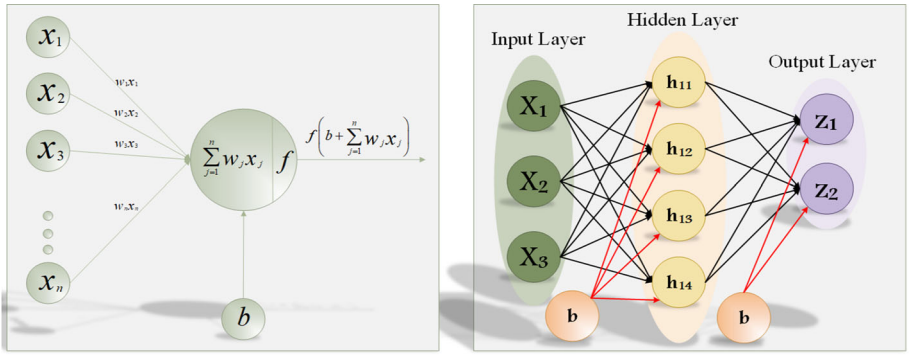


Fig. 13 Left: shows an artificial neuron. Right: sample of an artificial neural network

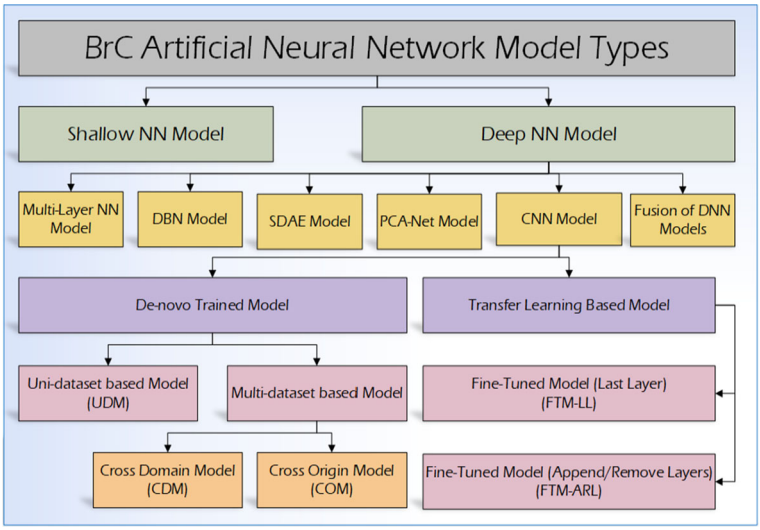


Fig. 14 Type of ANNs used for BrC classification and grading

ML technique that can learn and perform tasks, such as classification, prediction, decision-making, and visualization, by using sample data. Moreover, an ANN can perform multi-disciplinary tasks by using many types of real-life data, including structured (data in vector form), semi-structured (like emails), and unstructured data, such as BrC medical images. Many types of ANNs were developed to process different types of data. For the classification of BrC medical images, researchers mainly used two types of ANNs, namely, SNNs and DNNs (Fig. 14). Most researchers employed DNNs (also known as deep learning-based models) for BrC classification. In subsequent subsections, the types of ANNs used for BrC classification are discussed in the light of selected studies. Moreover, the pros and cons of each model are presented in Table 8.

3.4.1 Shallow neural network

An ANN with a single hidden layer is referred to as an SNN (Bebis and Georgiopoulos 1994). The basic building block (elementary unit) of an ANN is artificial neuron, simply referred as

neuron or node or hidden units. A simple AN is a mathematical function that works similar to a biological neuron. The output of an AN is represented by connection weights that update the effect of a given input, and the nonlinear characteristics are applied by any transfer function at a particular neuron. Afterward, neuron impulse is calculated by applying nonlinear function (i.e., activation function) on a weighted sum of input data. Simultaneously, a learning algorithm (e.g., backpropagation) is responsible for updating the weight to show the model's learning capability (Table 6). A simple AN and the basic structure of SNN are shown in Fig. 13.

In Fig. 13, a simple AN obtains unidirectional input, such as x_1, x_2, x_3, …, x_n, shown by arrows toward the activation function based on the weighted sum of input data. The neuron output is represented by $f(y)$ and has the following relationship:

$$f(y) = f\left(b + \sum_{j=1}^{n} w_j x_j\right), \tag{1}$$

where x_j, w_j represents the input and weight matrix, respectively, b is the bias neuron that allows a classifier to translate its decision boundary, *f(y)* is an activation function, and y is the sum of the scalar product of the weight matrix and input.

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n. \tag{2}$$

A nonlinearity function is also applied after the sum of the dot product of weighted inputs. The non-linearity function is also known as the activation function (Duch and Jankowski 1999). The most popular choices for activation functions are rectified linear unit (ReLU), tanh, and sigmoid as shown in Table 7.

As shown in Fig. 13, the nodes are distributed into the input, hidden, and output layers. The input signal flows from the input layer, passes toward the hidden layer, and ends at the output layer. Such type of input flow in a strict feed-forward fashion develops a feed forward ANN (FF-ANN). However, instead of using an ANN with single hidden layer, multiple hidden layers can be used, in FF-ANN. Noticeability, none of the nodes possess any connection within the same layer. This independency of neurons within a layer supports parallel computations while training an ANN. The training of an ANN is a learning process where patterns are learned from input data by changing the weights after applying some learning rules. Learning rules, such as backpropagation, delta rule, and perceptron rule, help modulate weights automatically while training the network. The trained ANN can then be used for prediction using real-life data. Recently, 3 out of 49 studies (Fig. 21) have created SNN models to classify BrC (Kumar et al. 2017b; Leod and Verma 2016; Rouhi et al. 2015). Kumar et al. (2017b), ensembled six binary ANNs for 4-class breast density grading classification using MGs. Rouhi et al. (2015) developed an SNN model to find the threshold for regions growing segmentation and classification of MGs into benign or malignant cases. These studies highlighted that using SNNs is beneficial for BrC classification. SNNs have some basic advantages owing to their simple structure. They possess a single hidden layer and work in feed forward fashion, thereby allowing to create, implement, and optimize BrC classification. SNNs consume the least computation resources and time among the different types of ANNs. Moreover, SNNs can produce better results than other types of ANNs even if the dataset is small. However, using SNN also has some limitations. For instance, an SNN used for structured data has a limited number of dimensions; otherwise, small networks are unable to show good generalization performance over high-dimensional data, especially when complex patterns need to be learned to solve multi-class problems. Moreover, the

Table 6 Distribution of studies among pre-processing methods and their advantages

Pre-processing method	Methodology	Advantages	References
Augmentation	Geometric Transform like rotation, flip	To avoid DNN model overfitting issue To overcome class imbalance training problem Network can learn lesions from many angles like a pathologist usually do in real life for better analysis of HP images.	Amit et al. (2017), Araujo et al. (2017), Arevalo et al. (2015), Bakkouri and Afdel (2017), Bardou et al. (2018), Bayramoglu et al. (2017), Bejnordi et al. (2017b), Bevilacqua et al. (2016), Byra et al. (2017), Carneiro et al. (2017), Chang et al. (2017), Cheng et al. (2016), Dhungel et al. (2017), Duraisamy and Emperumal (2017), Feng et al. (2018), Gandomkar et al. (2018), Hadad et al. (2017), Han et al. (2017a), Han et al. (2017b), Jaffar (2017), Kim et al. (2016), Kumar et al. (2017a), Nejad et al. (2017), Rasti et al. (2017), Samala et al. (2017), Samala et al. (2018), Sert et al. (2017), Spanhol et al. (2017), Spanhol et al. (2016a), Xu et al. (2016), Zhang et al. (2017) and Zheng et al. (2017)
	Add noise/Distortion (Gaussian noise, Barrel or Pin Cushion transforms)	Enables DNN to be trained robustly. Therefore, it can predict with higher accuracy even if images are noisy, as found in real life. Hence there will be improved class label prediction for noisy images DNN requires least pre-processing steps at the time of prediction.	
	Patch creation Methods (Patches with 50% overlapping, no overlapping or randomly selected patches)	Many images can be generated from the original images. Moreover, it can preserves the image aspect ratio, architecture or shape of lesion and subjective information. Hence, it increases the performance of classifier and reduces the chance of false negatives One can avoid synthetic images generated by geometric transform or noise addition methods	

Table 6 continued

Pre-processing method	Methodology	Advantages	References
		No need to rescale images before input to ANN. Hence, it may reduce the chance of information loss due to rescaling	
	Synthetic Minority Over-sampling Technique (SMOTE)	To increase the number of samples (vectors) to the minority class, in order to handle class imbalance problem before DNN training.	
ROI Extraction	Methods used like region growing, Nuclei Segmentation, Otsu Method, Markov Random Model	Enables to increase the number of positive and negative image samples Help DNN model to learn better representation related to abnormal and abnormal regions and reduces chances of overfitting Saves computation time and resourced	Amit et al. (2017), Arefan et al. (2015), Arevalo et al. (2015), Bevilacqua et al. (2016), Cao et al. (2016), Cheng et al. (2016), Duraisamy and Emperumal (2017), Feng et al. (2018), Fonseca et al. (2015), Han et al. (2017a), Khan (2017), Kim et al. (2016), Kumar et al. (2017b), Leod and Verma (2016), Nascimento et al. (2016), Rasti et al. (2017), Rouhi et al. (2015), Samala et al. (2017), Samala et al. (2018), Wan et al. (2017) and Zheng et al. (2017)
Scaling	Methods like Gaussian Pyramid, Bi-cubic interpolation, Bilinear interpolation	Required to resize the image before served as input to the DNN Carefully selected interpolation methods can avoid the loss of information in mapping to the new pixel grid Gaussian pyramid can help to increase the number of images along with resizing.	Abdullah-Al et al. (2017), Arefan et al. (2015), Bakkouri and Afdel (2017), Bayramoglu et al. (2017), Carneiro et al. (2017), Chang et al. (2017), Cheng et al. (2016), Dhungel et al. (2017), Duraisamy and Emperumal (2017), Fonseca et al. (2015), Gandomkar et al. (2018), Han et al. (2017b), Jaffar (2017), Kim et al. (2016), Kumar et al. (2017a), Nejad et al. (2017), Spanhol et al. (2016a), Wan et al. (2017), Xu et al. (2016) and Zhang et al. (2016)

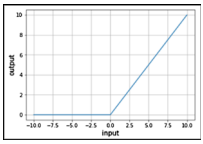
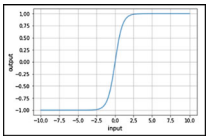
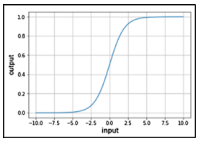
Table 6 continued

Pre-processing method	Methodology	Advantages	References
Normalization and Enhancement	Histogram equalization, adaptive Mean, Median filters, Log transforms, CLAHE method, Wiener Filter	Normalize the low-value and high-value intensity/contrast present in an image Adaptive filters remove noise by mean, variance and spatial correlations Reduces US image blurring effects and impulse noise DNN usually show better performance on normalized image, helps to minimize loss while backpropagation.	Arefan et al. (2015), Arevalo et al. (2015), Bejnordi et al. (2017b), Duraisamy and Emperumal (2017), Han et al. (2017a), Jaffar (2017), Khan (2017), Nejad et al. (2017), Rasti et al. (2017), Rouhi et al. (2015) and Sert et al. (2017)
Remove Artifacts	Using binary images and thresholding the pixel intensity, cropping border, Extracting Bigger regions, using geometric parabola around rib cage.	Help to eliminated non-breast regions (labels, wages, opaque markers, white strips/borders, thorax, lungs, chest wall and pectoral muscle) in mammogram, US and MRI.	Abdullah-Al et al. (2017), Cao et al. (2016), Gandomkar et al. (2018) and Wan et al. (2017)
Stain Normalization or Removal	Stain Normalization	To make variable color (due to H&E staining of HP images) uniform in all images of all patients. So that DNN will not distract due to brightness and color stain inconsistencies and show better classification results for multiclass BrC Contrast, intensity and color statistics of source images are almost alike the reference image Reinhard method preserves the structures of HP images. Therefore, suitable for BrC classification Khan's supervised method works at pixel level and thus achieves, a good result for stain separation.	Arefan et al. (2015), Bayramoglu et al. (2017), Bevilacqua et al. (2016) and Sert et al. (2017)

Table 6 continued

Pre-processing method	Methodology	Advantages	References
	Color Deconvolution	To extract intensities of hematoxylin–eosin (H&E) staining from HP images and convert into optical density space images without being significantly influenced. Hence it reduces the image dimensionality and uses least resources and enhances the performance of classification By adopting filtered and independent observations it reduces the signals impurity when estimating stain matrix It preserves texture information which is associated with stain colors in HP images.	

Table 7 Brief description of popular activation functions

Activation functions	ReLU	Tanh	Sigmoid
Equation	$\emptyset(x) = \max(x, 0)$	$\emptyset(x) = \tanh(x)$	$\emptyset(x) = \frac{1}{1+e^{-x}}$
Range	$(0, \infty)$	$(-1, 1)$	$(0, 1)$
Type	Discontinuous	Continuous	Continuous
Gradient	$if\ \emptyset > 0\ then\ 1,$ $else\ 0$	$1 - \emptyset(x)^2$	$\emptyset(x)(1 - \emptyset(x))$
Update suppressed near zero	Yes	Yes	No
Overcome vanishing gradient	Yes	No	No
Graph			

performance of the network depends on the designed features and the optimization of the network structure.

3.4.2 Deep neural network

DNNs are used for deep learning as an ML method and AI technique for automatic feature extraction. Usually, the word *deep* is referred when more than one hidden layer has been deployed between the input and output layers of any NN (Svozil et al. 1997). DNNs use representation learning to discover complex feature representation automatically (such as diagnosis of BrC using medical images) unlike traditional ML algorithms (e.g., support vector machine, random forest decision tree, and k-nearest neighborhood), which require HEFs to show optimum results. The empirical success of DNN is inherited by its mathematical formulas (Goceri 2018). Over the years, DNNs focused on applications such as speech recognition (Amodei et al. 2016; Hannun et al. 2014), fraud detection (Paula et al. 2016; Wang and Xu 2018), traffic sign detection (Islam et al. 2017), face recognition (Parkhi et al. 2015; Sun et al. 2014), emotion recognition (Jirayucharoensak et al. 2014; Kahou et al. 2016), natural language processing, medical image diagnosis (Lakhani and Sundaram 2017; Siddiqui et al. 2017; Wu et al. 2014), and human activity recognition (Nweke et al. 2018; Nweke et al. 2019). The upsurge in deep learning research is fueled by its ability to extract salient features from raw images of BrC without relying on laboriously extracted HEF. In recent years, an extensive number of DNNs have been proposed. The DNNs can be broadly categorized into multi-layer neural network (ML-NN), deep belief neural network, stacked denoising auto-encoders (SDAE), principal component analysis network (PCANet), and CNN. Furthermore, CNN models were either trained from scratch called De novo models or created through transfer learning (TL) by using pre-trained models (Fig. 14). In subsequent subsections, the types of DNNs used for BrC classification are discussed in the light of selected studies.

3.4.2.1 Multi-layer neural network An ML-NN is a type of DNN that is similar to an SNN. Nonetheless, an ML-NN possesses two or more hidden layers between the input and output layers, unlike an SNN (Bengio 2009; Deng and Yu 2014) (Fig. 15). However, ML-

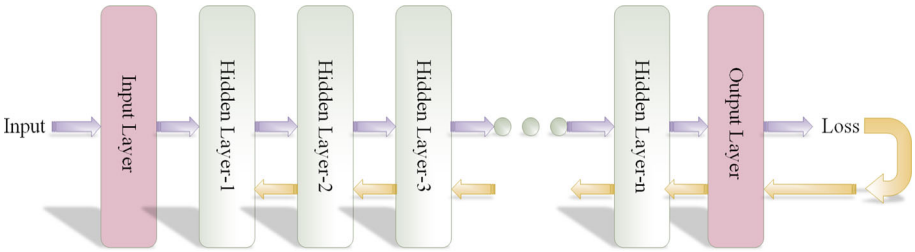


Fig. 15 A sample illustration of multi-layer neural network

NN training must be configured to obtain the desired results. Configuring an ML-NN is actually initializing and modulating the parameters to perform optimum training, such as initializing weights by generating any random number or by using prior domain knowledge before initiating the learning rule. Recently, the most popularly adopted learning rule is backpropagation (Abraham 2005). In backpropagation, the weights are automatically updated in each pass on the basis of error rate (loss) produced at the output layer by using gradient and chain-rule (Svozil et al. 1997). However, our survey revealed that very few studies (4 out of 49, Fig. 21) used ML-NN for BrC classification. Kumar et al. (2017b) proposed an ML-NN model with two hidden layers and optimized by different stopping criteria using 22 morphological features extracted from 100 US images to classify benign or malignant BrC.

Furthermore, Arefan et al. (2015) developed an ML-NN model using 2–20 hidden layers. They extracted nine statistical features from 43 Mg images to classify breast density as fatty, glandular, or dense. The afore-stated studies showed the urge of using ML-ANN. For instance, increasing the number of hidden layers can improve the generalization performance of the network. However, additional layers require more data instances for better training; otherwise, the network may be overfitted (good performance on validation data but unable to perform on target data). Furthermore, optimizing the number of hidden layers and training hyper-parameters for a larger size of ML-NN become crucial tasks (for further details, see Table 8).

3.4.2.2 Deep belief networks Deep belief network is a type of DNN (Hinton et al. 2006) that consists of several layers of restricted Boltzmann machines (RBMs), see Fig. 16a (Fischer and Igel 2012). An RBM is a generative model that serves as a building block in greedy layer-wise feature learning and training of DNN. RBM maps binary data-vectors using binary latent variables. Hence, the goal is to obtain abstract and distinct representation features. If the RBM network cannot directly be used for medical images (e.g., SWE images), then Point-wise gated Boltzmann machines (PGBM) (Fig. 16b) are adopted to model complex image data (e.g., BrC US-SWE images) while avoiding irrelevant patterns. Moreover, in unsupervised learning (performed by using unlabeled data), a DBN can learn to probabilistically reconstruct its inputs. Hence, a hidden layer works like a feature extracting entity. All the hidden layers are trained one after the other, i.e., one layer at a time. Finally, a DBN can be trained in a supervised fashion for classification (Fig. 16c). However, only one study utilized the advantages of DBN for BrC classification [28]. Zhang et al. (2016) deployed a two-layered DBN composed of PGBM and RBM for BrC binary classification by using breast US-based SWE colored images. PGBM was equipped to distinguish between relevant and irrelevant features from SWE images. Furthermore, relevant features were supplied to RBM to learn the relationship

Table 8 A summary of ANN models used in 49 studies for BrC classification

ANN types	Strengths	Weaknesses
SNN	Small size networks	Do not show good performance on high dimensional data
	Easy to develop, train and optimize the training parameters	Performance solely depends upon the designed features and the structure of ANN
	Small amount of data can obtain better generalization performance	
	Requires less training time, computational power, and memory to store weights	Difficult to generalize the predictions
ML-NN	It includes all advantages of SNN, additionally the increased hidden layers help to get better generalization performance	Includes all weaknesses of SNN, additionally higher number of hidden layers need more data to get better generalization performance
	High Dimensional data can be used for better feature extraction	
DBN	This efficient, greedy learning can be followed by, or combined with, other learning procedures that fine-tune all of the weights to improve the generative or discriminative performance of the whole network	Unable to track the loss while computing the log likelihood
	Can be deployed for high dimensional data that possess correlated features	
SADE	Automatic denoising form high dimensional data enhances the performance of BrC classification model, using real-life medical image	Denoising works better on high dimensional data as compared to low dimensions because of higher dependencies usually found among higher dimensions like BrC medical images
	Can track cross entropy which is what is being minimized by the model's learning algorithm like back-propagation	
PCA-Net	Due to large receptive field, it can extract overall observations of the objects in an image and captures more semantic level information	The use of simple hashing method cannot provide rich enough information to map the features. Hence effects the representation performance
	Due to binary hashing and block histogram PCANet is flexible for mathematical analysis and justification of its effectiveness	Preferred when data possess many irrelevant information.
CNN (De novo)	CNN(UDM): Customized deep CNN models can be created	CNN(UDM): usually difficult to train model for small number of images to solve multiclass problem
	Model can be created according to the type and number of images	Needs high expertise to design and optimize the deep network for specific data. May consumes lots of time and resources to get optimum results
	CNN(CDM): Includes same strengths as in CNN(UM)	CNN(CDM): Two times training of model will take longer time and may require higher resources

Table 8 continued

ANN types	Strengths	Weaknesses
	Additionally, model can be effective even if there are less number of target images to solve multiclass classification	Hard to optimize model training on two datasets of different domains like ImageNet and BreakHis
		Requires large number of instances (BrC images) with balanced distribution among classes
	CNN(COM): Customized deep CNN models can be created	CNN(COM): Medical images collected from different sites always have different image acquisition protocols. Hence needs extra and carefully adopted pre-processing methodologies to get a reliable generalized model
	The training, validation and testing performed on larger number of images of same modality. Usually show better performance	
	Preferred when source images (usually exclusive dataset images) are not enough for training	
	It allows to use all target images for testing purpose only	
CNN (pre-trained)	Deep CNN model can be trained quickly using least resources as compared to de novo training	If target dataset is very small (like 100 images) then results may be not reliable
	Can show comparable performance even if target data is smaller in size like HP BrC images	Retraining also requires class wise balance data to produce unbiased results, usually not found in real-life medical images
	CNN(FTM-ARL) possess fusion of new layers to be trained from scratch, so flexible to learn more generalized and unbiased weights from small amount of target data like BrC images as compared to FTM-LL	Limitations are same as in CNN(FTM-LL) except CNN(FTM-ARL): Training time may increase due to the introduction of new layers to be trained from scratch
		The optimization of newly appended layers needs to be addressed carefully to get desired results

among the BrC relevant features. Finally, SVM was used to classify benign or malignant BrC cases by using features extracted through RBM. The main advantage of using a DBN for image classification is that it is mostly trained layer by layer, allowing each layer to be optimized easily for improved feature generalization. In addition, the layers, except the last one, can be trained in an unsupervised fashion. The last hidden layer is usually trained in supervised manner to fine tune the network output. Hence, a DBN provides an opportunity to perform better training using a small number of annotated images, also called semi-supervised learning. Semi-supervised learning is useful for medical image classification because finding labelled images for different types of cancers is difficult. However, using RBMs layered deep networks also has some limitations. For instance, a DBN cannot track the loss while computing the log likelihood for which we care about as the best trained model.

3.4.2.3 Stacked denoising autoencoder A stacked denoising autoencoder (SDAE) is a type of stacked autoencoder that helps eliminate noisy features (Fig. 17). SDAE networks can automatically extract discriminant representative hidden patterns from data using intrinsic

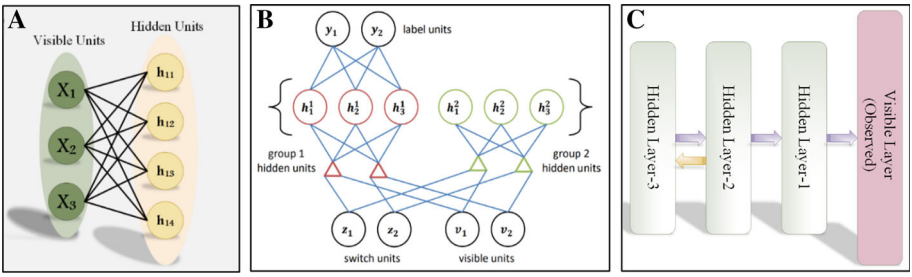


Fig. 16 A restricted Boltzmann machine (RBM) with fully connected visible and hidden units (**a**), a sample diagram of supervised PGBM shown (**b**) (Sohn et al. 2013) and **c** a sample diagram of supervised DBN

data reconstruction mechanism. The SDAE network can hypothetically address the issues of high variations in either shape or appearances of lumps. As the inherent benefit of automatic feature extraction along with noise tolerance, SDAE-based models can conceivably minimize issues related to image processing inaccuracies, which ultimately lead to non-reliable feature extraction. Due to noise tolerance nature, 2 out of 49 (Fig. 21) studies (Cheng et al. 2016; Feng et al. 2018) developed an SDAE-based model to classify BrC images. Cheng et al. (2016) developed a model by two-phased training. In the first phase, two-layered SDAE is trained using image ROIs. In the second phase, the pretrained model is refined by supervised learning with additional neurons to preserve the original image size and aspect ratio. Softmax was used for benign or malignant classification for both breast US and lung CT images, with an accuracy and area under the ROC curve (AUC) of $94.4\% \pm 3.2\%$ and $98.4\% \pm 1.5\%$, respectively. Similarly, Feng et al. (2018) deployed SDAE consisting of three layers along with softmax. An SDAE extracts features layer by layer from breast HP image ROIs in an unsupervised manner, and the model is fine-tuned by using labels to train softmax for benign or malignant BrC classification. The authors obtained $98.28\% \pm 0.12\%$ and $90.54\% \pm 0.45\%$ accuracies for the two classes. These results indicate that the performance of the SDAE-based model is comparable to that of any other type of DNN model because of its integral ability of noise reduction, especially when real-life medical images usually possess noise from different sources. Hence, auto noise reduction for medical images helps the network to learn more relevant features. Furthermore, layer-by-layer training facilitates easy optimization and regulation of training parameters. Regardless of its major advantages, SDAE also has some limitations. For instance, SDAE shows poor performance on low-dimensional data or data possessing poor correlation among the dimensions (Vincent et al. 2010a). High-dimensional data, such as medical images, usually inherit very high correlation.

3.4.2.4 Principal component analysis network Principal component analysis network (PCANet) is an easily implementable, two-staged, unsupervised deep learning technique for image classification (Chan et al. 2015). The two-staged network basically performs three tasks, namely, cascade PCA, binary hashing, and block wise histogram. PCA is used to learn multi-stage weights (filter banks), followed by binary hashing and block histogram for indexing and pooling. Binary hashing simply encodes the quantized binary code mapping to the sequence of principal components (Fig. 18). According to our survey, only one study [34] employed PCANet with some variation of kernel for breast and liver cancer analysis. Wu et al. (2016) created a PCANet-based model to classify breast/liver cancer HP images in binary classes. The author used random binary hashing in PCANet instead of simple sequence binary hashing to generate multiple random codes for information extraction. Finally, a low-rank

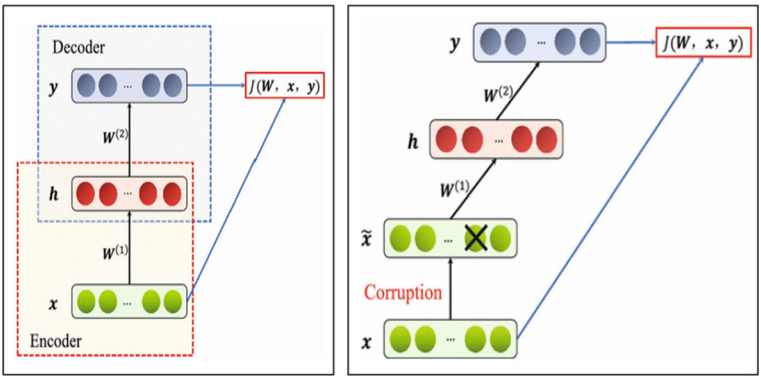


Fig. 17 Left side figure, a sample network diagram of traditional autoencoder. Right side figure, a network diagram of stacked denoising autoencoder (Vincent et al. 2010b)

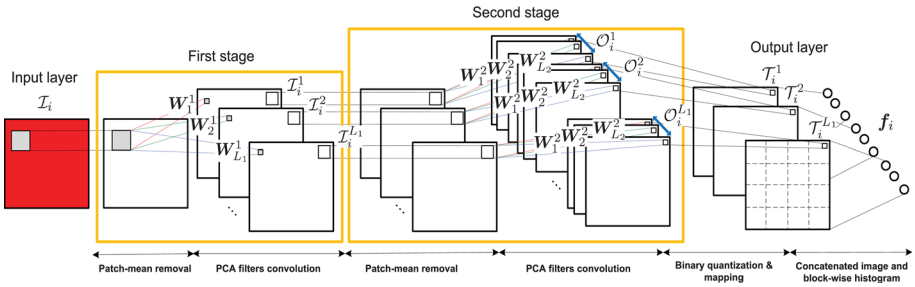


Fig. 18 A sample of two-staged PCANet block diagram (Chan et al. 2015)

bilinear classifier is used to classify images of two datasets. Compared with other deep learning networks, PCANets are easier to design, implement, and to train by using different types of high-dimensional data. Due to binary hashing and block histogram, PCANet is flexible for mathematical analysis and justification of its effectiveness. Moreover, PCANet has a large receptive field, so that it can extract overall observations of the objects in an image and learn invariance from it. Hence, PCANet can capture pixel-level information.

3.4.2.5 Convolutional neural network CNN is a type of deep learning-based ANN technique. This technique has gained attention after the work of (Hinton and Salakhutdinov 2006). Moreover, the history of CNN for medical image classification is a long one. Initially, a CNN-based "Neocognitron" model was proposed by (Fukushima and Miyake 1982). Recently, image classification has been revolutionized after the birth of AlexNet (Krizhevsky et al. 2012) (Fig. 19). A deep CNN model usually consists of some primary layers, such as an input layer, one or more convolution layers, one or more fully connected (FC) layers, and an output layer using softmax to compute label probabilities. Convolution layers are responsible for learning high-level features, such as edges and bobs, whereas FC layers learn pixel-level features. Apart from primary layers, some other layers including a normalization layer (increases network stability) and a pooling layer (progressively reduces the spatial size of the representation to reduce the amount of parameters and computation in the network)

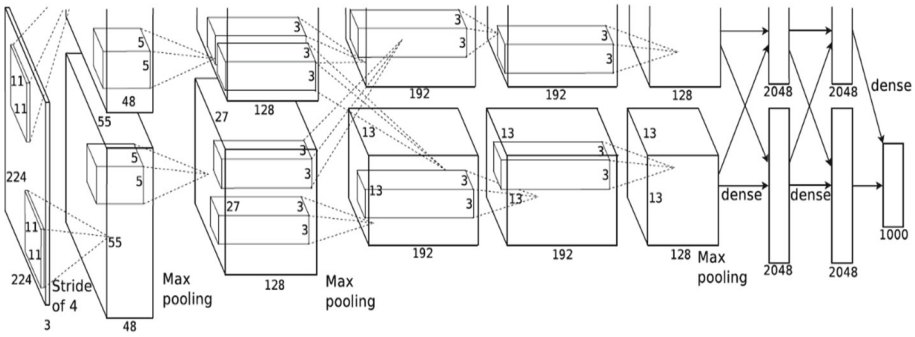


Fig. 19 An illustration of deep CNN based AlexNet model (Krizhevsky et al. 2012)

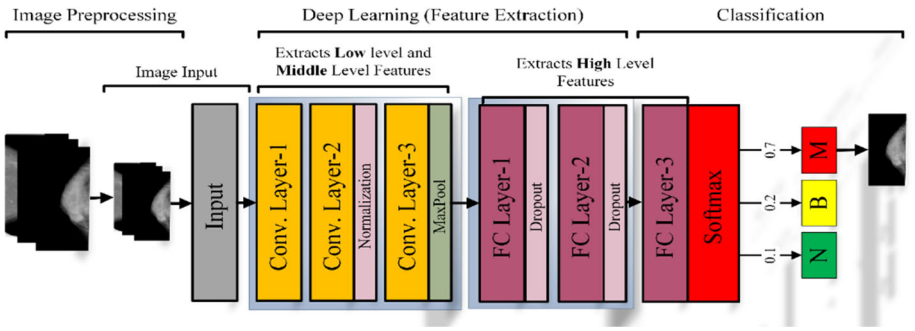


Fig. 20 An illustration of deep CNN based model for BrC classification using

may be used after convolution layers, and a dropout layer (reduces network overfitting) is usually deployed after the FC layer (Fig. 20). However, training is performed in a supervised manner using backpropagation. In addition, hyper parameters such as input image size and batch size (Goceri and Gooya 2018) need to be carefully adjusted to obtain optimum results. In brief, the concept of Deep CNN is to make a hierarchical model to represent data at multiple levels of abstraction and enable the model to obtain accurate representations from data in a self-taught manner (Shen et al. 2017).

The CCN used for breast classification is divided into two broad categories, namely, de novo trained model and TL-based model (Fig. 14). CNN models that were created and trained from scratch are called "de novo models" (Hadad et al. 2017). Conversely, CNN models that exploited previously trained networks (e.g., AlexNet, VGG-Net, GoogLeNet, and ResNet) are called "TL-based models."

This survey on BrC classification revealed that 28 out of 51 (Fig. 21) studies (Abdullah-Al et al. 2017; Amit et al. 2017; Araujo et al. 2017; Arevalo et al. 2015; Bakkouri and Afdel 2017; Bardou et al. 2018; Bayramoglu et al. 2017; Bejnordi et al. 2017b; Byra et al. 2017; Cao et al. 2016; Dhungel et al. 2017; Fonseca et al. 2015; Hadad et al. 2017; Han et al. 2017b; Kim et al. 2016; Kumar et al. 2017a; Nahid and Kong 2017b, 2018; Nejad et al. 2017; Qiu et al. 2017; Rasti et al. 2017; Spanhol et al. 2016a; Sun et al. 2017; Wan et al. 2017; Xu et al. 2016; Zheng et al. 2017) used de novo training (see Fig. 20). Conversely, 11 out of 51 studies (Bejnordi et al. 2017b; Dhungel et al. 2017; Han et al. 2017b; Kumar et al. 2017a; Zheng et al. 2017) employed pre-trained CNN for BrC classification. In this review, the de novo CNN models are further categorized into two subtypes, namely, uni-dataset

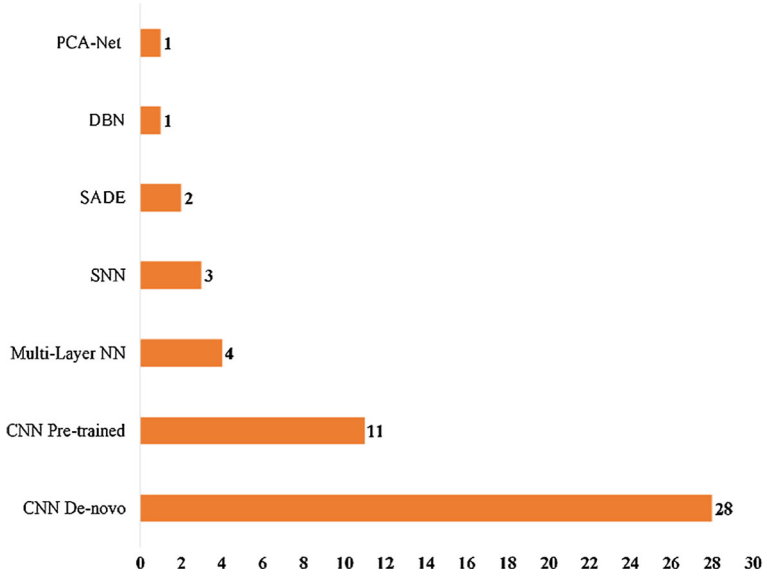


Fig. 21 Distribution of articles among various ANNs for BrC classification

and multi-dataset models. Uni-dataset models (UDM) are de novo models that are trained and tested on a single dataset, whereas cross-domain models (CDM) or cross-origin models (COM) are those trained and tested on multiple datasets (Fig. 14). CDM-type models are created from scratch, trained on a dataset of different domains (e.g., nonmedical images), and ultimately retrained (after fine tuning) for target datasets, such as BrC images. By contrast, COMs are trained on images of the same domain but collected from different sites, followed by fine tuning and retraining for the target dataset. However, CDMs are usually smaller in size (possess less number of layers) and created with some special layers to enhance the classification performance as compared with pretrained models such as AlexNet (Han et al. 2017b).

Apart from models trained from scratch, pre-trained models were also adopted in many studies (11 out of 49, Fig. 21) for BrC classification. The pretrained models were trained on natural images and mostly possess highly deep structure to learn many class labels; for instance, AlexNet trained for 1000 class labels and contain five convolution layers along with three FC layers (Fig. 19). Two strategies were adopted by researchers to perform TL for BrC classification. First, only the last layer was fine-tuned followed by the retraining of the adopted model, named here as the fine-tuned model (last layer) (FTM-LL). Second, one or more layers of the pre-trained network were replaced with newly created layers before retraining the network using target data, named here as the fine-tuned model (append/remove layer) (FTM-ARL) (Fig. 14).

3.4.2.6 Fusion of deep neural networks The review of selected studies showed that that most of the CNN-based models use a single type of CNN and are not used in a fused fashion. Some studies (Bejnordi et al. 2017b; Nahid and Kong 2018) deployed models by embedding some residual blocks along with convolutional layers on the basis of pretrained models, such as ResNet. However, CNN models that were used along with residual block and were trained from scratch produced good results. For instance, Nahid and Kong (2018) developed a model

using residual block with convolution layer and obtained an accuracy of 92.19%. By contrast, a fusion of CNNs was prepared by Nahid et al. (2018). The authors deployed three types of model, namely, CNN, long short-term memory (LSTM), and a fusion of CNN and LSTM. The CNN-based model outperformed the other models. Hence, the failure of fused models may be due to the small number of images that are to be fed into a larger fused network. In particular, training from scratch using a small number of images with a large (fused) network may produce unreliable results. Hence, considerable effort is required to assess confidently the effective use of fused CNN type of networks.

3.4.3 Empirical evaluation of BrC deep neural network models using different datasets

This section presents an empirical evaluation of different types of DNN on publicly available datasets. Table 9 shows the study-wise DNN models that have been employed on various datasets related to BrC classification. Here, majority of the studies employed CNN instead of multi-layer NN and SNN to classify BrC. Moreover, most of the studies used MGs (22 out of 49) followed by HP images. However, the most common datasets utilized for MG classification are DDSM, INBreast, BCDR-F03, and mini-MIAS.

Carneiro et al. (2017) developed a CNN(FTM-ARL)-based model and achieved the best performance (0.96 ± 0.05 VUS, 0.96 ± 0.05 AUC) by using the DDSM dataset for three classes (normal, benign, or malignant) of BrC. However, using the same DDSM dataset, Rouhi et al. (2015) and Leod and Verma (2016) deployed SNN and reported 0.94 AUC and 86% accuracy for a binary classification problem. These studies show that the CNN(FTM-ARL) model outperforms the SNN model using the same dataset. This finding can be attributed to the fact that CNN pre-trained models along with some new layers are more capable of learning better generalized activations as compared with shallow learning from scratch for BrC classification. Bakkouri and Afdel (2017) and Abdullah-Al et al. (2017) also used the DDSM MG dataset to distinguish between benign or malignant breast lesions. However, a former study adopted a CNN(UDM)-based model and showed a higher accuracy of 97.28% as compared with that obtained in a later study (i.e., 93.35%) that adopted a CNN(COM)-based model. The reason behind the success of CNN(UDM) may be because the former study extracted the image ROIs by using Gaussian pyramids, which may enhance the model performance. Moreover, both Dhungel et al. (2017) and Kumar et al. (2017a) used the INBreast MG dataset to distinguish between benign and malignant breast tumors. Although both studies used CNN(COM) models, the former study reported better performance (i.e., Sn = 98%, Sp = 70%) than the latter study (i.e., Ac = 75%, AUC = 0.57). Hence, the former study performed better than the latter possibly because of the use of a small network that is more likely to be overfitted instead of a deep-layered network. Similarly, Duraisamy and Emperumal (2017) and Arevalo et al. (2015) used the BCDR-F03 MG dataset to classify BrC. Here, the first study used a CNN(FTM-LL)-based model, whereas the second study created a CNN(UDM) model. The first study model outperformed the second one because TL-based models usually perform better on a small number of images (BCRDR-F03 possesses only 736 images) than models trained from scratch. Similarly, Jaffar (2017) and Nahid and Kong (2018) utilized mini-MIAS MGs for two (benign/malignant) and three (normal/benign/malignant) types of BrC predictions. Moreover, the former study created a ML-NN network, whereas the latter study employed a CNN(COM) model type. However, the latter study showed better performance (i.e., Sn = 97%, Sp = 100%) than the former (i.e., Sn = 93.25%, Sp = 90.50%). The better performance of the former study might be due to the smaller size of the network instead of using deep-layered convolutional networks, especially when dealing with a small number of images, such as the mini-MIAS dataset with only 322 images of 161 patients.

Table 9 Study-wise performance of ANNs for breast cancer classification

Ref	ANN Type	Dataset	# Classes	Performance
Kumar et al. (2017b)	SNN	DDSM	4	Ac = 79.5%
Rouhi et al. (2015)	SNN	DDSM, MIAS	2	Avg [(Ac = 86.66%, Sn = 87.91%, Sp = 85.40%, AUC = 0.8825) (MIAS), (Ac = 95.01%, Sn = 96.25%, Sp = 93.78%, AUC = 0.9499) (DDSM)]
Leod and Verma (2016)	SNN	DDSM, UCI	2	Ac = 86% (DDSM), Ac = 89.175% (UCI)
Feng et al. (2018)	SDAE	ED(HP Image)	2	Ac = 98.28 ± 0.12, 90.54 ± 0.45, Pr = 97.88, 90.04
Cheng et al. (2016)	SDAE	ED(US)	2	Ac = 94.4 ± 3.2, Sn = 90.8 ± 5.3, Sp = 98.1 ± 2.2, AUC = 98.4 ± 1.5
Wu et al. (2016)	PCA-Net	ED(HP Image)	2	Ac = 78.46 ± 3.92, Sn = 71.00 ± 4.18, Sp = 83.23 ± 5.30
Bevilacqua et al. (2016)	Multi-Layer NN	ED(MRI)	2	Avg Ac = 89.77 ± 5.84, Min Ac = 73.08 ± 0.43, Sn = 0.89 ± 0.10, Sp = 0.90 ± 0.09
Nascimento et al. (2016)	Multi-Layer NN	ED(US)	2	Ac = 96.98%, AUC = 0.98
Arefan et al. (2015)	Multi-Layer NN	mini-MIAS	3	Ac = 97.66%
Khan (2017)	Multi-Layer NN	mini-MIAS, BCDR	3	Sn = 97%, Sp = 100% (mini-MIAS)Mg, Sn = 98%, Sp = 97% (BCDR)Mg, Sn = 99%, Sp = 100% (BCDR)US
Zhang et al. (2016)	DBN	ED(US-SWE)	2	Ac = 93.4%, Sn = 88.6%, Sp = 97.1%, AUC = 0.947
Arevalo et al. (2015)	CNN(UDM)	BCDR-F03	2	AUC = 0.86
Araujo et al. (2017)	CNN(UDM)	BICBH	4, 2	Ac = 77.8% (4 classes), Ac = 83.3% (2 classes), Sn = 95.6%
Bardou et al. (2018)	CNN(UDM)	BreakHis	8, 2	Ac = 83.31–88.23% (8 Classes), Ac = 96.15%,98.33% (2 Classes)
Bayramoglu et al. (2017)	CNN(UDM)	BreakHis	3	Avg Ac = 80.10%

Table 9 continued

Ref	ANN Type	Dataset	# Classes	Performance
Spanhol et al. (2016a)	CNN(UDM)	BreakHis	2	Ac = 90.0 ± 6.7(Image level), Ac = 85.6 ± 4.8(Patient level)
Abdullah-Al et al. (2017)	CNN(UDM)	BreakHis	2	Ac = 85.36%, Sp = 70.36%, Rc = 91.36%, Pr = 89%
Nahid and Kong (2017b)	CNN(UDM)	BreakHis	2	Max Sp = 97.18%, Max Sn = 99%
Nejad et al. (2017)	CNN(UDM)	BreakHis	2	Ac = 77.5%
Nahid and Kong (2018)	CNN(UDM)	BreakHis	2	Ac = 92.19%, Sn = 94.94%, Rc = 98.20%, Pr = 98%
Nahid et al. (2018)	CNN(UDM)	BreakHis	2	Ac = 91%, Pr = 96%
Bakkouri and Afdel (2017)	CNN(UDM)	DDSM, BCDR	2	Ac = 97.28%, Sn = 99.79%, Sp = 94.78%.
Kim et al. (2016)	CNN(UDM)	ED(DBT)	2	Avg AUC = 0.847 ± 0.012
Rasti et al. (2017)	CNN(UDM)	ED(DCE-MRI)	2	Ac = 96.39%, Sn = 97.73%, Sp = 94.87%
Wan et al. (2017)	CNN(UDM)	ED(HP Image)	3	Avg. Ac = 69%
Cao et al. (2016)	CNN(UDM)	ED(HP Image)	2	Ac = 90%, 74%, 76%. AUC = 0.93
Xu et al. (2016)	CNN(UDM)	ED(HP Image)	2	Max AUC = 0.93163
Fonseca et al. (2015)	CNN(UDM)	ED(Mg)	4	Max Ac = 78.35%, Avg Ac = 73.05%,
Qiu et al. (2017)	CNN(UDM)	ED(Mg)	2	Avg AUC = 0.790 ± 0.019, Max AUC = 0.836 ± 0.036
Sun et al. (2017)	CNN(UDM)	ED(Mg)	2	Ac = 82.43%, AUC = 0.8818
Hadad et al. (2017)	CNN(UDM)	ED(Mg, MRI)	2	Ac = 94%, AUC = 0.98 (MRI)
Amit et al. (2017)	CNN(UDM)	ED(MRI)	3	Ac = 83%, AUC = 0.91
Byra et al. (2017)	CNN(UDM)	ED(US, Nakagami)	2	Ac = 83%, Sn = 82.4, Sp = 83.3, AUC = 0.912 ± 0.005
Duraisamy and Emperumal (2017)	CNN(FTM-LL)	BCDR-F03, MIAS	10	Ac = 99%, Sn = 98.75%, Sp = 1.0%, AUC = 0.9815

Table 9 continued

Ref	ANN Type	Dataset	# Classes	Performance
Gandomkar et al. (2018)	CNN(FTM-LL)	BreakHis	4, 2	Max Ac = 95.69% (Benign 4 subclasses), Max Ac = 97.89% (Malignant 4 subclasses) Ac = 96.25(Patient level), Max Ac = 98.52% (2 Classes),
Chang et al. (2017)	CNN(FTM-LL)	BreakHis	2	Ac = 83%,89%, AUC = 0.93
Spanhol et al. (2017)	CNN(FTM-LL)	BreakHis	2	Max Ac = 86.3%
Sert et al. (2017)	CNN(FTM-LL)	DDSM	2	Ac = 94.1%, Pr = 95%, Sn = 94%
Samala et al. (2017)	CNN(FTM-LL)	DDSM, ED(Mg)	2	AUC = 0.82 ± 0.02,
Zhang et al. (2017)	CNN(FTM-LL)	ED(Mg)	2	AUC = 0.73
Han et al. (2017a)	CNN(FTM-LL)	ED(US)	2	Ac = 91%, Sn = 0.86, Sp = 93%, AUC > 0.9
Samala et al. (2018)	CNN(FTM-ARL)	DDSM, ED(DBT)	2	ED AUC = 0.90 ± 0.4
Carneiro et al. (2017)	CNN(FTM-ARL)	DDSM, INBreast	3, 2	VUS = 0.96 ± 0.05(DDSM), 3-class, VUS = 0.94 ± 0.05 (INBreast), 3-class, AUC = 0.96 ± 0.05(DDSM), 2-class, AUC = 0.94 ± 0.05 (INBreast), 2-class
Kumar et al. (2017a)	CNN(COM)	CBIS-DDSM, MIAS, INBreast	2	Ac = 75%, AUC = 0.57 (INBreast)
Jaffar (2017)	CNN(COM)	DDSM, mini-MIAS	2	Avg Ac = 93.35%, Sn = 93% (DDSM), Avg Ac = 92.85%, Sn = 93.25% Sp = 90.50%, AUC = 0.92 (mini-MIAS).
Zheng et al. (2017)	CNN(COM)	ED(HP Image)	15, 2	Ac = 96.4% (15 classes), Ac = 95.9%, AUC = 0.86306 (2 classes)
Bejnordi et al. (2017b)	CNN(COM)	ED(HP Image-WSI)	3	Ac = 81.3%, AUC = 0.962
Dhungel et al. (2017)	CNN(COM)	INBreast	2	Sn = 98%, Sp = 70%
Han et al. (2017b)	CNN(CDM)	BreakHis, ImageNet	8	Avg Ac = 93.2%

ED exclusive dataset, *Ac* accuracy, *Sn* sensitivity, *Sp* specificity, *Pr* precision, *Avg* average, *Max* maximum

Apart from MG datasets, many studies (20 out of 49) used HP image datasets, especially for multi-class BrC classification. In addition, the dataset was commonly used for HP images in BreakHis followed by BICBH (Han et al. (2017b). Bardou et al. (2018) utilized the BreakHis dataset for multi-class (eight classes) BrC classification. The first study implemented a CNN(CDM) model, whereas the CNN(UDM) network was used by Bardou et al. (2018). Comparative analysis of both studies showed that the first study outperformed (Avg. Ac = 93.2%) the other study because of the pre-training of the newly created model using the ImageNet dataset. However, these studies improved the diagnosis of the eight subtypes of breast lesions. Similarly, other studies (Abdullah-Al et al. 2017; Nahid and Kong 2017b; Nahid et al. 2018; Nejad et al. 2017; Spanhol et al. 2016a) employed the BreakHis dataset by using the same type of network, such as CNN(UDM), for binary classification. However, the first study showed the highest accuracy of 92.19% among all the studies. The author possibly deployed many residual blocks using CNN (for global feature extraction) along with contourlet transform and histogram features (for local feature extraction).

Alongside MG or HP image datasets for BrC classification, some studies used US (Byra et al. 2017; Cheng et al. 2016; De S. Silva et al. 2015; Han et al. 2017a; Khan 2017; Nascimento et al. 2016; Zhang et al. 2016), MRI (Amit et al. 2017; Bevilacqua et al. 2016; Hadad et al. 2017; Rasti et al. 2017), or more than one modality (Hadad et al. 2017). Moreover, most of the datasets used for US and MRI images are exclusive because these modalities are rarely found in publicly available datasets. Zhang et al. (2016) employed a two-layered DBN for the extraction of features from breast US-SWE images for malignancy detection. The author narrated an accuracy of 93.4% (AUC = 0.94). Similarly, Nascimento et al. (2016) developed a ML-NN model to classify breast US images into benign or malignant lesions. The author reported a higher accuracy of 96.98% (AUC = 0.98). Furthermore, Byra et al. (2017) employed a CNN(UDM) model by using US-based Nakagami images. This study reported 83% accuracy (AUC = 0.912 ± 0.005) for binary classes of BrC. Few researchers adopted breast MRI modality (Amit et al. 2017; Bevilacqua et al. 2016; Hadad et al. 2017; Rasti et al. 2017) for cancer diagnosis using exclusive datasets. For instance, Bevilacqua et al. (2016) reported an accuracy of $89.77\% \pm 5.84\%$ for binary classes by deploying ML-NN for breast MRI classification. Similarly, Rasti et al. (2017) implemented a CNN(UDM) model from scratch for benign or malignant breast DCE-MRI classification. They reported the highest accuracy of 96.39% for malignancy diagnosis. Instead of using the single modality, the authors maximized multi-modality to train the NN model. Khan (2017) developed a CNN(COM) model by using two exclusive datasets of different modalities, such as MGs and breast MRI, to perform binary classification. However, model training was performed on MG images, whereas testing results were obtained by using breast MRI. The reported accuracy was 94% (AUC = 0.98) for benign and malignant classes of breast MRI image. Hence, this review shows that the fusion of multimodalities can improve the performance of DNN models.

3.5 Evaluation metrics analysis and review

After training the DNN model followed by image pre-processing, training, and validation of BrC images, the test images are then served as input to the trained DNN model for classification to evaluate its performance. In general, the evaluation metrics are computed from the confusion matrix. In the confusion matrix, the actual (input) classes are represented with rows, whereas the column represents the predicted (output) class labels. Therefore, the BrC can be classified as true positive or true negative when correctly classified and false

positive or false negative when incorrectly classified. Based on the confusion matrix, the most popularly adopted evaluation measures for BrC classification are accuracy, sensitivity, specificity, precision, FMeasure, AUC, and volume under the ROC surface (VUS) (Landgrebe and Duin 2008). These metrics are briefly defined in subsequent paragraphs.

Accuracy (Ac) This measure represents how many of the total instances are correctly classified. It simply shows how much normal patients are correctly predicted and how much abnormal (BrC) patients are correctly diagnosed. It can be expressed by Eq. (3):

$$Ac = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \tag{3}$$

Sensitivity (Sn) or Recall (Rc) This measure indicates how much of the total positive instances are predicted correctly. In simple words, it represents how much BrC patients are correctly predicted from overall abnormal (BrC) patients. Thus, it should be as high as possible. Low Sn means many cancer patients are misdiagnosed and will be treated as normal. Hence, Sn is highly important in medical image diagnosis. It can be computed by using Eq. (4):

$$Sn = \frac{TP}{(TP + FN)}. \tag{4}$$

Specificity (Sp) This measure shows how much of the total negative predictions are correct. It simply represents how much of the normal (BrC) prediction is correct. It should be high as possible but is less important in medical diagnosis than Sn. It can be denoted by Eq. (5):

$$Sp = \frac{TN}{(TN + FP)}. \tag{5}$$

Precision (Pr) This measure denotes how much of the total positive predictions are correct. It simply represents how much of the abnormal (BrC) prediction is correct. Both Sn and Pr should be high for medical image diagnosis to avoid misdiagnosis of cancerous patients. It can be calculated by Eq. (6):

$$Pr = \frac{TP}{(TP + FP)}. \tag{6}$$

FMeasure This measure reflects the simultaneous impact of both Sn and Pr through harmonic means by applying more penalty over extreme values. It helps to compare two models with high Sn and low Pr and vice versa. It can be measured by Eq. (7). where β is penalty and its value can be ½, 1, or 2.

$$FMeasure = \frac{\left(1 + \beta^2\right)(Pr \times Sn)}{\left(\beta^2 \times Pr + Sn\right)}. \tag{7}$$

AUCROC The ROC plots the curve of precision against sensitivity. AUC is a common evaluation measure that helps to choose optimal models and ignore suboptimal ones (Fig. 22a), showing the performance comparison of four classification models for BrC. The figure shows that model-1 outperforms the three other models. By contrast, model-4 shows the lowest performance. The AUC value can be computed by using Eq. (8). An AUC value lies between 1 and 0. However, an AUC value of 1 represents a perfect model and an area of 0.5 or below reflects an ineffective model.

$$\text{AUC} = \frac{\sum_i R_i\left(I_p\right) - I_p(I_p + 1)/2}{I_p + I_n}, \tag{8}$$

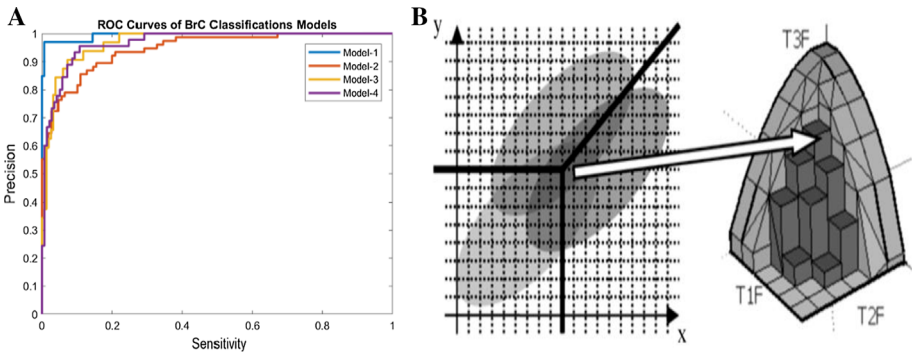


Fig. 22 **a** A sample ROC diagram, comparing the performance of four classification models of breast cancer. **b** Illustration of sample VUS diagram for three classes

where I_p and I_n denote the number of positive and negative BrC images, respectively, and R_i is the rank of the ith positive image in the ranked list.

Volume under the ROC surface (VUS) The ROC is a standard tool to evaluate two-class classification problems. It was extended and enabled to evaluate multi-class problems named as VUS (for three class VUSs, see Fig. 22b) (He and Frey 2008). Furthermore, in multi-classes, the independent and dependent (of same type) classes are grouped, and many ROCs are created. Finally, the decomposed ROCs are interrogated by using cost-sensitive and Neyman–Pearson optimization along with volume under the curve (Ferri et al. 2003).

Table 10 shows the frequency of studies that used particular performance evaluation measures to compute the performance of BrC classification models. The majority of the studies (12 out of 50) evaluated the performance by employing the accuracy metric. Moreover, studies (Bardou et al. 2018; Bayramoglu et al. 2017; Gandomkar et al. 2018; Han et al. 2017b; Nejad et al. 2017; Spanhol et al. 2016a, 2017) calculated accuracy at four magnifications ($40\times$, $100\times$, $200\times$, and $400\times$) based on two levels, such as accuracy at the image level and accuracy at the patient level, by using BreakHis HP images. However, the patient-level accuracy is more important than the image-level accuracy in medical science. For instance, previous studies (Bayramoglu et al. 2017; Gandomkar et al. 2018; Han et al. 2017b; Spanhol et al. 2016a, 2017) used the BreakHis dataset and showed accuracy at the four magnifications at both levels. Moreover, the second highest number of studies (9 out of 50) used accuracy along with AUC. The AUC evaluation measure is usually adopted to analyze the behavior of a model for each class (or for multiple model performance comparison). It reveals the authenticity of overall predicted accuracy and whether a model is biased toward any particular class. However, the studies that created multiple deep CNN de novo models used exclusive datasets and tried to solve the multi-class BrC problem by reporting the AUC along with accuracy to ensure that the newly trained model is unbiased and better than other models. For instance, four studies (Amit et al. 2017; Bejnordi et al. 2017b; Wan et al. 2017; Zheng et al. 2017) developed de novo models to classify BrC into more than two classes. Furthermore, an equal number of studies (6 out of 50) reported either the AUC or AUC along with Acc, Sn, and Sp. Few of the selected studies (Byra et al. 2017; Cheng et al. 2016; Han et al. 2017a; Zhang et al. 2016) that used exclusive datasets of breast US images considered the Ac, Sn, Sp, and AUC metrics to test the performance of trained CNN models before deploying their commercial usage.

Table 10 Frequency count of performance metrics used in each selected primary study

Study reference	Performance metrics	# of studies
Arefan et al. (2015), Bardou et al. (2018), Bayramoglu et al. (2017), Fonseca et al. (2015), Gandomkar et al. (2018), Hadad et al. (2017), Han et al. (2017b), Kumar et al. (2017b), Leod and Verma (2016), Nejad et al. (2017), Spanhol et al. (2017) and Spanhol et al. (2016a)	Accuracy	12
Amit et al. (2017), Bejnordi et al. (2017b), Cao et al. (2016), Chang et al. (2017), Kumar et al. (2017a), Nascimento et al. (2016), Sun et al. (2017), Wan et al. (2017) and Zheng et al. (2017)	Accuracy, AUC	9
Byra et al. (2017), Cheng et al. (2016), Han et al. (2017a), Jaffar (2017), Rouhi et al. (2015) and Zhang et al. (2016)	Accuracy, sensitivity, specificity, AUC	6
Arevalo et al. (2015), Kim et al. (2016), Samala et al. (2017), Samala et al. (2018), Xu et al. (2016) and Zhang et al. (2017)	AUC	6
Bakkouri and Afdel (2017), Bevilacqua et al. (2016), Duraisamy and Emperumal (2017), Rasti et al. (2017) and Wu et al. (2016)	Accuracy, sensitivity, specificity	5
Nahid and Kong (2018) and Sert et al. (2017)	Accuracy, sensitivity, precision, F-measure	2
Carneiro et al. (2017)	AUC, VUS	1
Khan (2017) and Nahid and Kong (2017b)	Sensitivity, specificity	2
Nahid et al. (2018)	Accuracy, precision	1
Feng et al. (2018)	Accuracy, precision, F-measure	1
Araujo et al. (2017)	Accuracy, sensitivity	1
Abdullah-Al et al. (2017)	Accuracy, sensitivity, specificity, precision, F-measure	1
Dhungel et al. (2017)	Sensitivity	1
Qiu et al. (2017)	Sensitivity, specificity, AUC	1

Apart from some basic evaluation measures, few studies used more sophisticated evaluation measures, such as FMeasure and VUS, for multi-class BrC classification. For instance, Carneiro et al. (2017) used the VUS metric to show the performance of a TL-based CNN model for three classes of BrC using the INBreast and DDSM datasets. Furthermore, some recent studies (Abdullah-Al et al. 2017; Feng et al. 2018; Nahid and Kong 2018; Sert et al. 2017) have reported FMeasure with few other evaluation metrics, such as Ac, Sn, Sp, and Rc.

4 Discussion

This review comprehensively studied the academic articles on BrC image classification published from January 2014 to June 2018. In specific, the current research enhanced the confidence level to make better decisions for BrC image analysis in five aspects, namely, use

of BrC imaging modalities, datasets, image pre-processing techniques adopted, creation of DNN models, and the performance metrics utilized to compare the results.

Most of the studies used public datasets instead of exclusive datasets. Noticeability, to solve BrC classification (especially multi-class) problems, deep learning algorithms require a large number of annotated medical images. However, the collection of annotated medical images has many limitations, such as the availability of required images in a large quantity, the availability of expert doctors to select and label the images, and the reliability of manually annotated images, especially when dealing with multi-class labeling, such as HP images of BrC. Hence, it is an extremely difficult and time-consuming process. Therefore, to avoid exclusive dataset critical issues, most researchers preferred to use publically available datasets. Public datasets can easily be downloaded from websites, which usually provide a large number of BrC images of many patients/cases along with labels and other related information. However, public datasets are usually preprocessed. For instance, DDSM dataset MGs are cropped, and non-breast regions are removed and converted into a computer-readable format with lossless compression (Heath et al. 2000). Hence, the model developed on public datasets may become less robust to handle real-life images.

Among all the imaging modalities, most researchers adopted MG images instead of US or MRI for BrC classification based on two (benign or malignant) or three classes (normal, benign, or malignant). The reason may be most available public datasets are based on MGs and servicing for many years. Moreover, MGs are usually used by researchers for binary classification. Hence, modeling is easier using MGs as compared with eight class classification, such as HP images in the BrakHis dataset. Moreover, preprocessed MGs usually allow to extract better intensity as well as texture features for BrC classification. However, finding the soft tissue (e.g., breast) density in MGs is difficult. In addition, MGs provide very poor shape-based features of breast lesion or calcification, and no thermal feature extraction is possible for BrC classification. Hence, US images can be adopted in addition to MGs to overcome these issues. Breast US images can extract texture- and shape-based features but not intensity-based features. Hence, breast US images are weak in detecting small nodules and accurate borders of breast lesion. Apart from MG and US images, MRI images are also used for BrC classification. However, MRI images are similar to MGs, except MRI images use low-intensity radiations and create many images that look like a video stream. Hence, MRI can enhance BrC visibility to provide more opportunity to extract features. Apart from grayscale images, color images (e.g., HP images) are also utilized by many researches to solve the BrC classification problem to identify two, four, and eight subtypes of BrC. HP images enable the CAD system to solve the multi-class BrC problem with more confidence as compared with any expert doctor (Vestjens et al. 2012). However, HP images, even those (among the images of two patients) developed in the same digital laboratory, possess high inconsistencies due to variations in color, intensity, and brightness. Hence, HP images need advanced pre-processing techniques to normalize without losing color-, texture-, intensity-, and shape-related information of breast lesion and its surrounding. Otherwise, it can lead to poor classification results, especially when dealing with multi-class (up to eight subtypes) BrC classification.

This review identified two major types of ANNs, such as SNNs and DNNs, for BrC classification. However, few researchers employed SNNs because their simple network can learn tasks better for both practical and theoretical reasons. In addition, they require less training time, computational power, and memory to store intermediate computational results (e.g., weights). Thus, they can be implemented economically with ease by using a normal desktop machine. Moreover, SNNs can show better generalization performance on a small amount of data than DNNs. SNNs also provide quicker reposes than DNNs at the time of

testing as required in real time. However, using SNNs has some limitations. For instance, they may not show better performance on high-dimensional data such as BrC images. Usually, SNNs use structured data; hence, their performance depends on the designed features and the number of neurons used in hidden layers. Therefore, to avoid the limitations of SNNs, most researchers employed DNN-based approaches for BrC classification. This review indicates that DNN-based BrC classification approaches are based on either ML-NN or CNN. In ML-NN, the increased number of hidden layers is supported to improve the generalization performance for BrC image classification. However, it requires a larger number of images as compared with SNNs. Furthermore, the performance of the network depends on the optimization of parameters, number of hidden layers used, and number of neurons per layer employed in the creation of ML-NN for BrC classification. Such type of network is difficult to optimize, especially in the multi-class classification of BrC images. Alternatively, the majority of researchers used CNN-based approaches to deal with high-dimensional data for multi-class BrC classification. CNN approaches used by researchers are often of two types: establishment of a de novo model that is trained from scratch or adoption of a pre-trained model also known as TL-based model. However, the majority of DNN-based models are based on CNN de novo models because de novo models are created and optimized according to the size, nature, and type of specific data, such as BrC images. Hence, a small CNN de novo model can produce better BrC classification results if designed and trained with proper optimization. Conversely, employment and training of deeper layers on a small amount of data may face more overfitting issues. Furthermore, de novo training the parameter optimization is difficult and can be achieved by trial-and-error methods. Hence, multiple models may be created and trained, which requires a long time and computational resources. Therefore, to overcome de novo CNN training issues, many researchers deployed pre-trained models, such as AlexNet. These models are already trained on millions of nonmedical images (natural images) to classify ten hundred natural objects, such as a pen, a tree, and a cap. Moreover, TL-based models are retrained on medical images after fine tuning. Fine tuning may involve the removal of last layers, the use of a small learning rate, and freezing the weights of the first few layers. The analysis of selected studies reveals that the TL-based models show comparable performance while using a small number of medical images. Moreover, these models can be trained without using high-computational resources, such as GPU. However, if the dataset is too small (like less than 1000 images), then the pretrained network may lead to overfitting. Otherwise, researchers performed augmentation (rotation, translation, and flipping) to increase the number of images.

The "no free lunch" theorem of Wolpert and Macready (1997) inferred that no any single ML algorithm performs optimally in all domains. Hence, a variety of DNN-based techniques should be employed to evaluate which algorithm outperforms on a specific type of data, such as BrC images. The selected primary studies implemented their own customized data set and different experimental settings. Thus, statistically comparing the performance values across the studies is infeasible. Nonetheless, comparison of the performance of different studies shows that the CNN model outperforms other DNN models for BrC classification.

5 Open issues and future research direction

This section presents new research directions that can be further exploited in BrC classification. This section gives prominence to future research directions. Considerable effort is required to improve the performance of BrC multi-class classification problems by using

medical image multimodalities. The open issues and future research directions are discussed as follows.

1. *Multiple imaging modalities for BrC classification* Exiting studies mostly employed single modalities, such as MG, US, or MRI images, for BrC classification. However, multiple imaging modalities of the same patient can be utilized during the construction of the BrC model to increase the reliability and correctness of the automated BrC classification model. For instance, during the training phase, the MG and US images are combined to construct the BrC model. In addition, the multi-view stacking approach can be used by utilizing a variety of imaging modalities to construct and evaluate the BrC model. In this manner, we are actually trying to utilize features of both modalities. In this study, MG represents better features related to image intensity and texture, whereas US images possess texture and shape features more prominently. Hence, the combination of various types of features taken from a single case can help enhance the classification ability of the BrC model.
2. *Multiple HEFs and imaging modalities for BrC classification* Recently, most studies used only breast imaging modalities to develop a DNN classification model. However, only Byra et al. (2017) used US-Nakagami statistical parameters for BrC classification along with US images. Hence, breast images may be used with other types of modalities, such as DNA sequences or physical examination findings, such as change in breast size or shape, skin dimpling, thickening, swelling, or redness. Reducing FNs in BrC classification is beneficial for researchers.
3. Adoption of rarely used imaging modalities for BrC classification Almost all studies used well-known (e.g., MG, US, and HP) breast imaging modalities to develop DNN classification models. However, only few studies (Amit et al. 2017; Bevilacqua et al. 2016; Cheng et al. 2016; Rasti et al. 2017; Zhang et al. 2016) used some other breast imaging modalities, such as US-SWE (Zhang et al. 2016), MRI (Amit et al. 2017; Bevilacqua et al. 2016; Rasti et al. 2017), or CT images (Cheng et al. 2016), for BrC classification. To the best of our knowledge, no study has used digital infrared breast images (thermal images) for BrC classification. Hence, models that can benefit from many other types of breast imaging modalities should be introduced instead to depend on the limited number of breast imaging technologies. Such type of rarely used image (e.g., CT, PET and thermal images) modalities may enhance the performance of BrC classification models.
4. *Publicly available dataset for BrC classification* A standardized public dataset (SPD) is required for each specific cancer type, such as BrC. To reduce the dependency of single modality, an SPD should be created by using multimodalities for each case/patient. For instance, multiple views of MGs along with some other modalities, such as US, MRI, or CT images, should be provided for the same patient. Furthermore, images should be collected from all types of BrC cases. Images of borderline cases should be marked for multi-class BrC classification because they enable researchers to analyze the robustness of the newly created model for multi-class BrC classification. Moreover, such type of dataset provides an opportunity for the researchers to learn more generalized representation features by using DNNs for BrC classification.
5. *Medical image-based TL approach for BrC classification* TL-based models adopt pre-trained models that are already trained on a huge number of non-medical images for multi-class classification. Medical images are usually used to retrain the pretrained models followed by a fine tuning step. However, TL usually faces overfitting issue if a small number of images is used for retraining. Hence, models that are trained on medical

images are needed. Such type of domain-specific pretrained models is cost effective and consumes less resources and time to be trained and tested efficiently for any type of cancer.

6. *Domain aware TL approach for BrC classification* The collected number of images is insufficient to train a model from scratch for BrC classification. In this case, in the first phase, the model should be trained from scratch by using any public dataset of similar type of modality. In the second phase, TL is performed, and the pretrained model of the first phase is retained for the target (exclusive) dataset. Such type of models usually overcomes the overfitting issues and can show promising performance.
7. *Unsupervised clustering approaches for BrC classification* The majority of the selected primary studies used BrC classification in a supervised learning fashion. These approaches produced better results by utilizing labelled images for training. However, in real life, BrC images are difficult to collect along with proper labels tagged by expert doctors. In the majority of cases, a large number of medical images are available without labels. The large number of unlabeled images is an important source of information and cannot be used in supervised learning. Therefore, a BrC classification model that can be trained in an unsupervised fashion by using a variety of clustering techniques is urgently needed.
8. *Active learning approaches for BrC classification* The review of 49 studies revealed that only two studies (Cheng et al. 2016; Feng et al. 2018) implemented active learning approach to classify breast images. Active learning is a semi-supervised learning approach where BrC classification can obtain optimum results with few labeled images. This type of approach becomes highly effective when a large number of breast images is collected but very few of these images are annotated. Nonetheless, the labelling of collected BrC images is a difficult, time-consuming, and cumbersome task. Thus, new researchers may explore various active learning algorithms for BrC classification.
9. *Reinforcement learning approach for BrC classification* Enabling an ML model so that it can learn from its environment concurrently is a major challenge. The main issue is to obtain enough samples of BrC images to represent all types of BrC distinctly. The model can learn from experience and predict a specific class label from data (Sutton and Barto 1998). Moreover, the automatic multi-class classification of BrC images is facing many challenges because of high intra-class similarity and low inter-class similarity issues. Thus, the development of a reinforcement learning-based model may convincingly improve the efficiency and performance of BrC classification models by using medical images.
10. *Case-based reasoning (CBR) for BrC classification* CBR is an approach to solve new problems by recalling the solution of past problems. It retains and updates the current solution made by humans and is applied on future problems. To the best of our knowledge, none of the 49 studies implemented the CBR approach to classify BrC into multiple classes by using images. Due to the advent of new imaging technologies, medical science is able to diagnose cancer in a detailed and sophisticated manner. Hence, a CBR-based CAD system that can learn from routine decisions/experiences made by expert doctors for BrC classification should be developed.

6 Conclusion

This thorough review presented a critical analysis of BrC classification by analyzing collectively the major research endeavours presented by current scholars to assist the new researchers in this domain. Articles on BrC classification published in 2014–2018 were extensively reviewed. Overall, 49 academic studies were carefully selected from eight unique academic repositories. The review was performed on the basis of selected primary studies from five aspects, namely, datasets used, various medical imaging modalities exploited, image pre-processing techniques, types of DNNs, and performance metrics used to construct and evaluate the BrC classification model. In BrC classification, various types of public and exclusive datasets were used. However, exclusive datasets are usually smaller in size than public datasets. Thus, more researchers preferred to use public datasets over exclusive ones. However, public datasets that contain multimodality images of the same patient along with some other information, such as DNA sequence, are urgently needed. Such type of dataset can help reduce FPs using automated systems. Furthermore, among all the datasets, MG and HP imaging modalities were widely adopted, followed by US images, and very few used MRI and CT breast images. Thus, other modalities (e.g., PET, CT and thermal images) that may provide different types of lesion characteristics should be explored to improve BrC classifications results. Furthermore, in pre-processing tasks, image augmentation, scaling, image intensity/contrast normalization, stain normalization, and stain removal techniques were mostly adopted to remove image inconsistences before feeding to any DNN model. However, pre-processing techniques should be adopted carefully so that important information, such as lesion texture-, shape-, and illumination-based information, can be preserved. In this review, several types of DNN architecture were identified to classify BrC. Among these, CNN was the most popular choice of researchers for BrC classification. Of these CNN-based models, de novo and TL-based models were employed by the researchers, and results showed that de novo models showed better results. By contrast, pre-trained models were also tested on small datasets after fine tuning by using augmented images for BrC multi-class classification. To evaluate the DNN models, various performance metrics were used, such as AUC, accuracy, sensitivity, specificity, FMeasure, and VUS. Among these, the first three were more common and essential in medical image classification. Finally, this review revealed various new research challenges that require extensive efforts to improve BrC classification models. We believe that this comprehensive review will provide a profound understanding of the BrC classification domain and valuable insights to researchers in this field.

Acknowledgements This work was supported by University Malaya Research Grant Program—AFR (Frontier Science) (RG380-17AFR).

Compliance with ethical standards

Conflict of interest The authors have no conflict of interests to declare.

Appendix

See Tables 11 and 12.

Table 11 List of questions used as quality evaluation criteria (QEC)

Sr. #	QEC questions
QEC 1	Is the aim of study clearly defined?
QEC 2	Is research methodology complete and well-defined?
QEC 3	Are the adopted pre-processing techniques justified?
QEC 4	Is the number of training images and testing images specified?
QEC 5	Is the class imbalance and DNN model overfitting issues for training are addressed?
QEC 6	Is the DNN model and classifiers used clearly defined?
QEC 7	Is the DNN model tested on more than one datasets?
QEC 8	Are Multiple performance metrics of study compared with existing baseline papers?
QEC 9	Are performance metric results properly interpreted and discussed

Table 12 Quality evaluation criteria applied on 56 studies

Studies	QEC 1	QEC 2	QEC 3	QEC 4	QEC 5	QEC 6	QEC 7	QEC 8	QEC 9	Total
Samala et al. (2018)	1	1	1	1	1	1	1	0	1	8
Samala et al. (2017)	1	1	1	1	1	1	1	1	1	9
Carneiro et al. (2017)	1	1	1	1	1	1	1	1	1	9
Hadad et al. (2017)	1	1	1	1	1	1	1	1	1	9
Jiang et al. (2017)	1	1	1	0	1	1	0	0	1	6
Chang et al. (2017)	1	1	1	1	1	1	0	1	1	8
Haarburger et al. (2018)	1	1	1	0	1	1	0	0	1	6
Arefan et al. (2015)	1	1	1	1	1	1	0	1	0	7
Arevalo et al. (2015)	1	1	1	1	1	1	0	1	0	7
Fonseca et al. (2015)	1	1	1	1	1	1	1	1	1	9
Rouhi et al. (2015)	1	1	1	1	1	1	1	1	1	9
Bekker et al. (2016)	1	1	1	0	1	1	0	0	1	6
Bevilacqua et al. (2016)	1	1	1	1	1	1	0	1	1	8
Cao et al. (2016)	1	1	1	1	1	1	1	0	0	7
Cheng et al. (2016)	1	1	1	1	0	1	1	1	1	8
Kim et al. (2016)	1	1	1	1	1	1	0	1	1	8
Leod and Verma (2016)	1	1	1	1	0	1	1	1	0	7
Spanhol et al. (2016a)	1	1	1	1	1	1	0	1	0	7
Wu et al. (2016)	1	0	1	1	0	1	1	1	1	7
Xu et al. (2016)	1	1	1	1	0	1	1	1	0	7
Zhang et al. (2016)	1	1	1	1	0	1	0	1	1	7
Abdullah-Al et al. (2017)	1	1	1	1	1	1	0	0	1	7
Amit et al. (2017)	1	1	1	0	1	1	0	1	1	7
Araujo et al. (2017)	1	1	1	1	1	1	0	1	1	8

Table 12 continued

Studies	QEC 1	QEC 2	QEC 3	QEC 4	QEC 5	QEC 6	QEC 7	QEC 8	QEC 9	Total
Bakkouri and Afdel (2017)	1	1	1	1	1	1	1	1	1	9
Bayramoglu et al. (2017)	1	1	1	1	1	1	0	1	0	7
Bejnordi et al. (2017a)	1	1	0	1	0	1	0	1	1	6
Bejnordi et al. (2017b)	1	1	1	1	1	1	0	1	1	8
Byra et al. (2017)	1	1	1	1	1	0	0	1	1	7
Dhungel et al. (2017)	1	1	1	1	1	1	0	1	1	8
Duraisamy and Emperumal (2017)	1	1	1	1	1	1	1	1	1	9
Gardezi et al. (2017)	1	1	0	1	0	1	0	1	1	6
Han et al. (2017b)	1	1	1	1	1	1	0	1	0	7
Khan (2017)	1	1	1	0	0	1	1	1	1	7
Kumar et al. (2017a)	1	1	1	1	0	1	1	0	1	7
Nahid and Kong (2017b)	1	1	1	1	0	1	0	1	1	7
Nejad et al. (2017)	1	1	1	1	1	1	0	1	0	7
Qiu et al. (2017)	1	1	1	1	0	1	0	1	1	7
Rasti et al. (2017)	1	1	1	1	1	1	0	1	1	8
Sert et al. (2017)	1	1	1	1	1	1	0	1	1	8
Sun and Binder (2017)	1	1	1	1	1	1	0	0	0	6
Sun et al. (2017)	1	1	1	1	1	1	0	0	1	7
Ting et al. (2017)	1	1	1	0	0	1	0	0	1	5
Wan et al. (2017)	1	1	1	1	0	1	0	1	1	7
Zhang et al. (2017)	1	1	1	1	1	1	0	1	0	7
Zheng et al. (2017)	1	1	1	1	1	1	1	1	1	9
Bardou et al. (2018)	1	1	1	1	1	1	0	1	1	8
Feng et al. (2018)	1	1	1	1	1	1	1	0	0	7
Gandomkar et al. (2018)	1	1	1	1	1	1	0	1	0	7
Nahid and Kong (2018)	1	1	1	1	0	1	0	1	1	7
Nahid et al. (2018)	1	1	1	1	0	1	0	1	1	7
Spanhol et al. (2017)	1	1	1	0	1	1	0	1	1	7
Nascimento et al. (2016)	1	1	1	1	1	1	0	1	1	8
Jaffar (2017)	1	1	1	1	1	1	1	1	1	9
Han et al. (2017a)	1	1	0	1	1	1	0	1	1	7
Kumar et al. (2017b)	1	1	1	1	1	1	0	1	1	8

References

Abraham A (2005) Artificial neural networks. In: Peter H, Sydenham RT (eds) Handbook of measuring system design. Wiley, London, pp 901–908

Adoui ME, Drisis S, Benjelloun M (2017) Analyzing breast tumor heterogeneity to predict the response to chemotherapy using 3D MR images registration. Paper presented at the Proceedings of the 2017 international conference on smart digital environment, Rabat, Morocco. http://delivery.acm.org/10.1145/3130000/3128137/p56-el_adoui.pdf?ip=103.18.2.251&id=3128137&acc=ACTIVE%20SERVICE&key=69AF3716A20387ED%2EE7759EC8BE158239%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1527212708_ca160ac108047af6b06969e33a701e4f. Accessed 15 Aug 2018

Ahn SJ, Kim YS, Kim EY, Park HK, Cho EK, Kim YK et al (2013) The value of chest CT for prediction of breast tumor size: comparison with pathology measurement. World J Surg Oncol 11:130. https://doi.org/10.1186/1477-7819-11-130

Aksebzeci BH, Kayaalti Ö (2017) Computer-aided classification of breast cancer histopathological images. Paper presented at the 2017 Medical Technologies National Congress (TIPTEKNO)

Amit G, Ben-Ari R, Hadad O, Monovich E, Granot N, Hashoul S (2017) Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches. Paper presented at the progress in biomedical optics and imaging—proceedings of SPIE

Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C et al (2016) Deep speech 2: end-to-end speech recognition in English and mandarin. Paper presented at the International conference on machine learning

Antropova N, Abe H, Giger ML (2018a) Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. J Med Imaging 5(1):6. https://doi.org/10.1117/1.jmi.5.1.014503

Antropova N, Huynh B, Giger M (2018) Recurrent neural networks for breast lesion classification based on DCE-MRIs. Paper presented at the progress in biomedical optics and imaging—proceedings of SPIE

Araujo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C et al (2017) Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 12(6):14. https://doi.org/10.1371/journal.pone.0177544

Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C et al (2017) Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 12(6):e0177544

Arefan D, Talebpour A, Ahmadinejhad N, Asl AK (2015) Automatic breast density classification using neural network. J Instrum. https://doi.org/10.1088/1748-0221/10/12/t12002

Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Lopez MAG (2015) Convolutional neural networks for mammography mass lesion classification. Paper presented at the 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

Bakkouri I, Afdel K (2017) Breast tumor classification based on deep convolutional neural networks. Paper presented at the Proceedings—3rd international conference on advanced technologies for signal and image processing, ATSIP 2017

Bardou D, Zhang K, Ahmad SM (2018) Classification of breast cancer based on histology images using convolutional neural networks. IEEE Access. https://doi.org/10.1109/access.2018.2831280

Barr RG (2012) Sonographic breast elastography: a primer. J Ultrasound Med 31(5):773–783

Bayramoglu N, Kannala J, Heikkila J (2017) Deep learning for magnification independent breast cancer histopathology image classification. Paper presented at the Proceedings—international conference on pattern recognition

Bebis G, Georgiopoulos M (1994) Feed-forward neural networks. IEEE Potentials 13(4):27–31

Bejnordi BE, Lin J, Glass B, Mullooly M, Gierach GL, Sherman ME et al (2017a) Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. Paper presented at the 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)

Bejnordi BE, Zuidhof G, Balkenhol M, Hermsen M, Bult P, van Ginneken B et al (2017b) Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. J Med Imaging 4(4):8. https://doi.org/10.1117/1.jmi.4.4.044504

Bekker AJ, Greenspan H, Goldberger J (2016) A multi-view deep learning architecture for classification of breast microcalcifications. Paper presented at the Proceedings—international symposium on biomedical imaging

Bengio Y (2009) Learning deep architectures for AI. Found Trends® Mach Learn 2(1):1–127. https://doi.org/10.1561/2200000006

Beutel J, Kundel HL, Van Metter RL (2000) Handbook of medical imaging, vol 1. SPIE Press, Bellingham

Bevilacqua V, Brunetti A, Triggiani M, Magaletti D, Telegrafo M, Moschetta M (2016) An optimized feed-forward artificial neural network topology to support radiologists in breast lesions classification. Paper presented at the GECCO 2016 companion—proceedings of the 2016 genetic and evolutionary computation conference

Breast Cancer Imaging (2018) Breast Cancer Imaging. Retrieved from http://www.aboutcancer.com/breast_cancer_imaging.htm. Accessed 20 Aug 2018

Byra M, Piotrzkowska-Wroblewska H, Dobruch-Sobczak K, Nowicki A (2017) Combining Nakagami imaging and convolutional neural network for breast lesion classification. Paper presented at the IEEE international ultrasonics symposium, IUS

Cao J, Qin Z, Jing J, Chen J, Wan T (2016) An automatic breast cancer grading method in histopathological images based on pixel-, object-, and semantic-level features. Paper presented at the 2016 IEEE 13th international symposium on biomedical imaging (ISBI)

Carneiro G, Nascimento J, Bradley AP (2015) Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp. 652–660

Carneiro G, Nascimento J, Bradley AP (2017) Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans Med Imaging 36(11):2355–2365. https://doi.org/10.1109/TMI.2017.2751523

Chan T, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017–5032. https://doi.org/10.1109/TIP.2015.2475625

Chang J, Yu J, Han T, Chang H, Park E (2017) A method for classifying medical images using transfer learning: a pilot study on histopathology of breast cancer. Paper presented at the 2017 IEEE 19th international conference on e-health networking, applications and services (Healthcom)

Chen H, Qi X, Yu L, Dou Q, Qin J, Heng P-A (2017a) DCAN: deep contour-aware networks for object instance segmentation from histology images. Med Image Anal 36:135–146. https://doi.org/10.1016/j.media.2016.11.004

Chen JM, Li Y, Xu J, Gong L, Wang LW, Liu WL, Liu J (2017b) Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: a review. Tumor Biol 39(3):12. https://doi.org/10.1177/1010428317694550

Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C et al (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 6:24454. https://doi.org/10.1038/srep24454

Chris Rose DT, Williams A, Wolstencroft K, Taylor C (2006) DDSM: digital database for screening mammography. Retrieved from http://marathon.csee.usf.edu/Mammography/Database.html. Accessed 26 Aug 2018

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P et al (2013) The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057. https://doi.org/10.1007/s10278-013-9622-7

Conceição RC, Medeiros H, Halloran MO, Rodriguez-Herrera D, Flores-Tapia D, Pistorius S (2014) SVM-based classification of breast tumour phantoms using a UWB radar prototype system. Paper presented at the 2014 XXXIth URSI general assembly and scientific symposium (URSI GASS)

Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC et al (2017) Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci Rep. https://doi.org/10.1038/srep46450

Dalkey N, Helmer O (1963) An experimental application of the Delphi method to the use of experts. Manag Sci 9(3):458–467

De S Silva SD, Costa MGF, De A Pereira WC, Filho CFFC (2015) Breast tumor classification in ultrasound images using neural networks with improved generalization methods. Paper presented at the Proceedings of the annual international conference of the IEEE Engineering in Medicine and Biology Society, EMBS

Deng L, Yu D (2014) Deep learning: methods and applications. Found Trends® Signal Process 7(3–4):197–387. https://doi.org/10.1561/2000000039

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

Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 31(4–5):198–211

Dua D (2017) KT UCI machine learning repository. University of California, School of Information and Computer Science, Irvine. Retrieved from http://archive.ics.uci.edu/ml

Duch W, Jankowski N (1999) Survey of neural transfer functions. Neural Comput Surv 2(1):163–212

Duraisamy S, Emperumal S (2017) Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier. IET Comput Vision 11(8):656–662. https://doi.org/10.1049/iet-cvi.2016.0425

Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM et al (2009) Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 253(3):641–651

Ertosun MG, Rubin DL (2015) Probabilistic visual search for masses within mammography images using deep learning. Paper presented at the 2015 IEEE international conference on bioinformatics and biomedicine (BIBM)

Farahani N, Parwani AV, Pantanowitz L (2015) Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol Lab Med Int 7:23–33

Feng Y, Zhang L, Yi Z (2018) Breast cancer cell nuclei classification in histopathology images using deep neural networks. Int J Comput Assist Radiol Surg 13(2):179–191. https://doi.org/10.1007/s11548-017-1663-9

Ferri C, Hernández-Orallo J, Salido MA (2003) Volume under the ROC surface for multi-class problems. Paper presented at the European conference on machine learning

Fischer A, Igel C (2012) An introduction to restricted Boltzmann machines. Paper presented at the Iberoamerican Congress on pattern recognition

Fonseca P, Mendoza J, Wainer J, Ferrer J, Pinto J, Guerrero J, Castaneda B (2015) Automatic breast density classification using a convolutional neural network architecture search procedure. Paper presented at the Progress in biomedical optics and imaging—proceedings of SPIE

Fukushima K, Miyake S (1982) Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. Competition and cooperation in neural nets. Springer, Berlin, pp 267–285

Gandomkar Z, Brennan PC, Mello-Thoms C (2018) MuDeRN: multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med. https://doi.org/10.1016/j.artmed.2018.04.005

Gardezi SJS, Awais M, Faye I, Meriaudeau F (2017) Mammogram classification using deep learning features. Paper presented at the 2017 IEEE international conference on signal and image processing applications (ICSIPA)

Goceri E (2017) Advances in digital pathology. Paper presented at the international conference on applied analysis and mathematical modeling. Istanbul, Turkey

Goceri E (2018) Formulas behind deep learning success. Paper presented at the international conference on applied analysis and mathematical modeling. Istanbul, Turkey

Goceri E, Goceri N (2017) Deep learning in medical image analysis: recent advances and future trends. Paper presented at the international conferences computer graphics, visualization, computer vision and image processing. Istanbul, Turkey

Goceri E, Gooya A (2018) On the importance of batch size for deep learning. Paper presented at the international conference on mathematics. Istanbul, Turkey

Goceri E, Songul C (2018) Biomedical information technology: image based computer aided diagnosis systems. Paper presented at the international conference on advanced technologies. Antalya, Turkey

Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B (2009) Histopathological image analysis: a review. IEEE Rev Biomed Eng 2:147

Haarburger C, Langenberg P, Truhn D, Schneider H, Thüring J, Schrading S et al (2018) Transfer learning for breast cancer malignancy classification based on dynamic contrast-enhanced MR images. Paper presented at the Informatik aktuell

Hadad O, Bakalo R, Ben-Ari R, Hashoul S, Amit G (2017) Classification of breast lesions using cross-modal deep learning. Paper presented at the proceedings—international symposium on biomedical imaging

Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, Seong YK (2017a) A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 62(19):7714–7728. https://doi.org/10.1088/1361-6560/aa82ec

Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S (2017b) Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep. https://doi.org/10.1038/s41598-017-04075-z

Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E et al (2014) Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567

Hassan NA, Yassin AH, Tayel MB, Mohamed MM (2016) Ultra-wideband scattered microwave signals for detection of breast tumors using artifical neural networks. Paper presented at the 2016 3rd International conference on artificial intelligence and pattern recognition, AIPR 2016

He X, Frey EC (2008) The meaning and use of the volume under a three-class ROC surface (VUS). IEEE Trans Med Imaging 27(5):577–588

Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP (2000) The digital database for screening mammography. Paper presented at the Proceedings of the 5th international workshop on digital mammography
Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
Hofvind S, Hovda T, Holen ÅS, Lee CI, Albertsen J, Bjørndal H et al (2018) Digital breast tomosynthesis and synthetic 2D mammography versus digital mammography: evaluation in a population-based screening program. Radiology 287(3):787–794
Hussain A, Farooq K, Luo B, Slack W (2015). A novel ontology and machine learning inspired hybrid cardiovascular decision support framework. Paper presented at the 2015 IEEE symposium series on computational intelligence
Islam KT, Raj RG, Mujtaba G (2017) Recognition of traffic sign based on bag-of-words and artificial neural network. Symmetry 9(8):138
Jaffar MA (2017) Deep learning based computer aided diagnosis system for breast mammograms. Int J Adv Comput Sci Appl 8(7):286–290
Jalalian A, Mashohor S, Mahmud R, Karasfi B, Saripan MIB, Ramli ARB (2017) Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. Exclus J 16:113–137. https://doi.org/10.17179/excli201-701
James JJ, Wilson ARM, Evans AJ (2016) The breast. Retrieved from https://radiologykey.com/the-breast-2/. Accessed 28 Aug 2018
Jarrett K, Kavukcuoglu K, LeCun Y (2009) What is the best multi-stage architecture for object recognition? Paper presented at the 2009 IEEE 12th international conference on computer vision
Jiang F, Liu H, Yu S, Xie Y (2017) Breast mass lesion classification in mammograms by transfer learning. Paper presented at the ACM international conference proceeding series
Jing H, Yang Y, Nishikawa RM (2012) Regularization in retrieval-driven classification of clustered microcalcifications for breast cancer. J Biomed Imaging 2012:3
Jirayucharoensak S, Pan-Ngum S, Israsena P (2014) EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. Sci World J 2014:1–10
Jyh-Horng C, Jinn Tsong T, Tung-Kuan L, Kao-Shing H, Hon-Yi S (2014) Predictive models for 5-year mortality after breast cancer surgery. Paper presented at the 2014 International conference on machine learning and cybernetics
Kahou SE, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K et al (2016) Emonets: multimodal deep learning approaches for emotion recognition in video. J Multimodal User Interfaces 10(2):99–111
Kasban H, El-Bendary M, Salama D (2015) A comparative study of medical imaging techniques. Int J Inf Sci Intell Syst 4:37–58
Keele S (2007) Guidelines for performing systematic literature reviews in software engineering. In: Technical report, Ver. 2.3 EBSE Technical Report. EBSE
Khan MHM (2017) Automated breast cancer diagnosis using artificial neural network (ANN). Paper presented at the 2017 3rd Iranian conference on signal processing and intelligent systems, New York
Khan AM, Rajpoot N, Treanor D, Magee D (2014) A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 61(6):1729–1738
Kim DH, Kim ST, Ro YM (2016) Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. Paper presented at the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)
Abdullah-Al N, Bin Ali F, Kong YN, IEEE (2017) Histopathological breast-image classification with image enhancement by convolutional neural network. Paper presented at the 2017 20th International conference of computer and information technology, New York
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Paper presented at the Advances in neural information processing systems
Kumar D, Kumar C, Shao M (2017a) Cross-database mammographic image analysis through unsupervised domain adaptation. Paper presented at the 2017 IEEE international conference on big data (big data)
Kumar I, Bhadauria HS, Virmani J, Thakur S (2017b) A classification framework for prediction of breast density using an ensemble of neural network classifiers. Biocybern Biomed Eng 37(1):217–228. https://doi.org/10.1016/j.bbe.2017.01.001
Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284(2):574–582
Landgrebe TC, Duin RP (2008) Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. IEEE Trans Pattern Anal Mach Intell 30(5):810–822

Lebron L, Greenspan D, Pandit-Taskar N (2015) PET imaging of breast cancer: role in patient management. PET Clinics 10(2):159–195. https://doi.org/10.1016/j.cpet.2014.12.004
Lee H, Chen Y-PP (2015) Image based computer aided diagnosis system for cancer detection. Expert Syst Appl 42(12):5356–5365. https://doi.org/10.1016/j.eswa.2015.02.005
Leod PM, Verma B (2016) Polynomial prediction of neurons in neural network classifier for breast cancer diagnosis. Paper presented at the Proceedings—international conference on natural computation
Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
Liu F, Hernandez-Cabronero M, Sanchez V, Marcellin MW, Bilgin A (2017) The current role of image compression standards in medical imaging. Information 8(4):131
Lo C, Shen Y-W, Huang C-S, Chang R-F (2014) Computer-aided multiview tumor detection for automated whole breast ultrasound. Ultrason Imaging 36(1):3–17. https://doi.org/10.1177/0161734613507240
Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X et al (2009) A method for normalizing histology slides for quantitative analysis. Paper presented at the IEEE international symposium on biomedical imaging: from nano to macro, 2009. ISBI'09
McCann MT, Ozolek JA, Castro CA, Parvin B, Kovacevic J (2015) Automated histology analysis: opportunities for signal processing. IEEE Signal Process Mag 32(1):78–87
McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115–133. https://doi.org/10.1007/BF02478259
Mehdy MM, Ng PY, Shair EF, Saleh NIM, Gomes C (2017) Artificial neural networks in image processing for early detection of breast cancer. Comput Math Methods Med. https://doi.org/10.1155/2017/2610628
Mendel KR, Li H, Sheth D, Giger ML (2018) Transfer learning with convolutional neural networks for lesion classification on clinical breast tomosynthesis. Paper presented at the Progress in biomedical optics and imaging—proceedings of SPIE
MFMER (2018) Breast MRI. Retrieved from https://www.mayoclinic.org/tests-procedures/breast-mri/about/pac-20384809. Accessed 30 Aug 2018
Mina LM, Mat Isa NA (2015) Breast abnormality detection in mammograms using artificial neural network. Paper presented at the I4CT 2015—2015 2nd international conference on computer, communications, and control technology, art proceeding
Moon M, Cornfeld D, Weinreb J (2009) Dynamic contrast-enhanced breast MR imaging. Magn Reson Imaging Clin N Am 17(2):351–362
Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) Inbreast: toward a full-field digital mammographic database. Acad Radiol 19(2):236–248
Moura DC, López MAG (2013) An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. Int J Comput Assist Radiol Surg 8(4):561–574
Murtaza G, Shuib L, Mujtaba G, Raza G (2019) Breast cancer multi-classification through deep neural network and hierarchical classification approach. Multimed Tools Appl. https://doi.org/10.1007/s11042-019-7525-4
Nahid AA, Kong Y (2017a) Involvement of machine learning for breast cancer image classification: a survey. Comput Math Methods Med. https://doi.org/10.1155/2017/3781951
Nahid AA, Kong YA (2017b) Local and global feature utilization for breast image classification by convolutional neural network. Paper presented at the 2017 International conference on digital image computing—techniques and applications, New York
Nahid AA, Kong Y (2018) Histopathological breast-image classification using local and frequency domains by convolutional neural network. Information (Switzerland). https://doi.org/10.3390/info9010019
Nahid AA, Mehrabi MA, Kong Y (2018) Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. Biomed Res Int. https://doi.org/10.1155/2018/2362108
Nascimento CDL, Silva SDS, da Silva TA, Pereira WCA, Costa MGF, Costa Filho CFF (2016) Breast tumor classification in ultrasound images using support vector machines and neural networks. Revista Brasileira de Engenharia Biomedica 32(3):283–292. https://doi.org/10.1590/2446-4740.04915
Nejad EM, Affendey LS, Latip RB, Ishak IB (2017) Classification of histopathology images of breast into benign and malignant using a single-layer convolutional neural network. Paper presented at the ACM international conference proceeding series
Nweke HF, Teh YW, Alo UR, Mujtaba G (2018) Analysis of multi-sensor fusion for mobile and wearable sensor based human activity recognition. Paper presented at the Proceedings of the international conference on data processing and applications
Nweke HF, Teh YW, Mujtaba G, Al-garadi MA (2019) Data fusion and multiple classifier systems for human activity detection and health monitoring: review and open research directions. Inf Fusion 46:147–170

Pack C, Shin S, Choi HD, Jeon SI, Kim J (2016) Optimized multilayer perceptron using dynamic learning rate based microwave tomography breast cancer screening. Paper presented at the proceedings of the ACM symposium on applied computing

Pan X, Li L, Yang H, Liu Z, Yang J, Zhao L, Fan Y (2017) Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks. Neurocomputing 229:88–99. https://doi.org/10.1016/j.neucom.2016.08.103

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. Paper presented at the BMVC

Paula EL, Ladeira M, Carvalho RN, Marzagão T (2016) Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. Paper presented at the 2016 15th IEEE international conference on machine learning and applications (ICMLA)

Qiu Y, Yan S, Gundreddy RR, Wang Y, Cheng S, Liu H, Zheng B (2017) A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. J X-Ray Sci Technol 25(5):751–763. https://doi.org/10.3233/XST-16226

Radiological Society of North America, I. R. (2018) RadiologyInfo for patients. Retrieved from https://www.radiologyinfo.org/en/info.cfm?pg=genus. Accessed 2 Sep 2018

Rasti R, Teshnehlab M, Phung SL (2017) Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. Pattern Recogn 72:381–390. https://doi.org/10.1016/j.patcog.2017.08.004

Rebecca Sawyer Lee FG, Hoogi A, Rubin D (2016) Curated breast imaging subset of DDSM dataset. The Breast Cancer Imaging Archieve. Retrieved from https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM#4413fe70f2bb4159b326a3f07fa6e6a9. Accessed 10 Sep 2018

Reinhard E, Adhikhmin M, Gooch B, Shirley P (2001) Color transfer between images. IEEE Comput Graphics Appl 21(5):34–41

Rouhi R, Jafari M, Kasaei S, Keshavarzian P (2015) Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Syst Appl 42(3):990–1002. https://doi.org/10.1016/j.eswa.2014.09.020

Rubin R, Strayer DS, Rubin E (2008) Rubin's pathology: clinicopathologic foundations of medicine. Lippincott Williams & Wilkins, Philadelphia

Ruifrok AC, Johnston DA (2001) Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 23(4):291–299

Sadaf A, Crystal P, Scaranelo A, Helbich T (2011) Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. Eur J Radiol 77(3):457–461

Saidin N, Sakim HAM, Ngah UK, Shuaib IL (2012) Segmentation of breast regions in mammogram based on density: a review. arXiv preprint arXiv:1209.5494

Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Cha KH, Richter CD (2017) Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 62(23):8894–8908. https://doi.org/10.1088/1361-6560/aa93d4

Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Richter C, Cha K (2018) Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. Phys Med Biol 63(9):8. https://doi.org/10.1088/1361-6560/aabb5b

Sathish D, Kamath S, Rajagopal KV, Prasad K (2016) Medical imaging techniques and computer aided diagnostic approaches for the detection of breast cancer with an emphasis on thermography—a review. Int J Med Eng Inf 8(3):275–299. https://doi.org/10.1504/IJMEI.2016.077446

Schneider M, Yaffe M (2000) Better detection: improving our chances. Paper presented at the Digital mammography: 5th international workshop on digital mammography IWDM

Selvathi D, Aarthy Poornila A (2018) Deep learning techniques for breast cancer detection using medical image analysis. In: Hemanth J, Balas VE (eds) Biologically rationalized computing techniques for image processing applications. Springer, Cham, pp 159–186

Sert E, Ertekin S, Halici U (2017) Ensemble of convolutional neural networks for classification of breast microcalcification from mammograms. Paper presented at the Proceedings of the annual international conference of the IEEE engineering in medicine and biology society, EMBS

Shan J, Alam SK, Garra B, Zhang YT, Ahmed T (2016) Computer-aided diagnosis for breast ultrasound using computerized Bi-rads features and machine learning methods. Ultrasound Med Biol 42(4):980–988. https://doi.org/10.1016/j.ultrasmedbio.2015.11.016

Shen D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Siddiqui MF, Mujtaba G, Reza AW, Shuib L (2017) Multi-class disease classification in brain MRIs using a computer-aided diagnostic system. Symmetry 9(3):37

Sivachitra M, Vijayachitra S (2015) Classification of post operative breast cancer patient information using complex valued neural classifiers. Paper presented at the 2015 International conference on cognitive computing and information processing (CCIP)
Sohn K, Zhou G, Lee C, Lee H (2013) Learning and selecting features jointly with point-wise gated Boltzmann machines. Paper presented at the Proceedings of the 30th international conference on international conference on machine learning—volume 28, Atlanta, GA, USA
Sophie Softley Pierce, P. M., Breast Cancer Care (2017) Three quarters of NHS Trusts and Health Boards say 'not enough' care for incurable breast cancer patients
Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2016a) Breast cancer histopathological image classification using convolutional neural networks. Paper presented at the Proceedings of the international joint conference on neural networks
Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2016b) A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng 63(7):1455–1462
Spanhol FA, Oliveira LS, Cavalin PR, Petitjean C, Heutte L (2017) Deep features for breast cancer histopathological image classification. Paper presented at the 2017 IEEE international conference on systems, man, and cybernetics (SMC)
Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C et al (1994) The mammographic image analysis society digital mammogram database. Paper presented at the Exerpta Medica. International Congress series
Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C et al (2015) Mammographic Image Analysis Society (MIAS) database v1, p 21
Sun J, Binder A (2017) Comparison of deep learning architectures for H&E histopathology images. Paper presented at the 2017 IEEE conference on big data and analytics (ICBDA)
Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. Paper presented at the Advances in neural information processing systems
Sun W, Tseng TB, Zhang J, Qian W (2017) Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput Med Imaging Graph 57:4–9. https://doi.org/10.1016/j.compmedimag.2016.07.004
Sutton RS, Barto AG (1998) Introduction to reinforcement learning, vol 135. MIT Press, Cambridge
Svozil D, Kvasnicka V, Pospichal J (1997) Introduction to multi-layer feed-forward neural networks. Chemometr Intell Lab Syst 39(1):43–62
Tan T, Platel B, Twellmann T, van Schie G, Mus R, Grivegnee A et al (2013) Evaluation of the effect of computer-aided classification of benign and malignant lesions on reader performance in automated three-dimensional breast ultrasound. Acad Radiol 20(11):1381–1388. https://doi.org/10.1016/j.acra.2013.07.013
Tataroğlu GA, Genç A, Kabakçı KA, Çapar A, Töreyin BU, Ekenel HK et al (2017) A deep learning based approach for classification of CerbB2 tumor cells in breast cancer. Paper presented at the 2017 25th Signal processing and communications applications conference (SIU)
Tessa S, Keith JFM (2018) The difference between an MRI and CT scan. Retrieved from https://www.healthline.com/health/ct-scan-vs-mri. Accessed 13 Sep 2018
Ting FF, Sim KS, IEEE (2017) Self-regulated multilayer perceptron neural network for breast cancer classification. Paper presented at the 2017 International conference on robotics, automation and sciences, New York
Tsui P-H, Yeh C-K, Chang C-C, Liao Y-Y (2008) Classification of breast masses by ultrasonic Nakagami imaging: a feasibility study. Phys Med Biol 53(21):6027
Tsui P-H, Ho M-C, Tai D-I, Lin Y-H, Wang C-Y, Ma H-Y (2016) Acoustic structure quantification by using ultrasound Nakagami imaging for assessing liver fibrosis. Sci Rep 6:33075
Ultrasound (2018) General ultrasound. Retrieved from https://www.radiologyinfo.org/en/info.cfm?pg=genus. Accessed 17 Sep 2018
van Zelst JCM, Tan T, Clauser P, Domingo A, Dorrius MD, Drieling D et al (2018) Dedicated computer-aided detection software for automated 3D breast ultrasound; an efficient tool for the radiologist in supplemental screening of women with dense breasts. Eur Radiol. https://doi.org/10.1007/s00330-017-5280-3
Vestjens JHMJ, Pepels MJ, de Boer M, Borm GF, van Deurzen CHM, van Diest PJ, Tjan-Heijnen VCG (2012) Relevant impact of central pathology review on nodal classification in individual breast cancer patients. Ann Oncol 23(10):2561–2566. https://doi.org/10.1093/annonc/mds072
Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010a) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408
Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010b) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408

Wan T, Cao J, Chen J, Qin Z (2017) Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features. Neurocomputing 229:34–44. https://doi.org/10.1016/j.neucom.2016.05.084
Wang Y, Xu W (2018) Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst 105:87–95
Wang J, Yang Y (2018) A context-sensitive deep learning approach for microcalcification detection in mammograms. Pattern Recogn 78:12–22. https://doi.org/10.1016/j.patcog.2018.01.009
Wang H, Cruz-Roa A, Basavanhally A, Gilmore H, Shih N, Feldman M et al (2014) Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. J Med Imaging. https://doi.org/10.1117/1.jmi.1.3.034003
Wang D, Wu K, Gu C, Guan X (2017) Time efficient cell detection in histopathology images using convolutional regression networks. Paper presented at the 2017 36th Chinese control conference (CCC)
Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82
World Health Organization (2018) Cancer. Retrieved from http://www.who.int/en/news-room/fact-sheets/detail/cancer. Accessed 20 Sep 2018
Wu K, Chen X, Ding M (2014) Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. Optik Int J Light Electron Opt 125(15):4057–4063
Wu J, Shi J, Li Y, Suo J, Zhang Q (2016) Histopathological image classification using random binary hashing based PCANet and bilinear classifier. Paper presented at the 2016 24th European signal processing conference (EUSIPCO)
Xu J, Luo X, Wang G, Gilmore H, Madabhushi A (2016) A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 191:214–223. https://doi.org/10.1016/j.neucom.2016.01.034
Xu J, Zhou C, Lang B, Liu Q (2017) Deep learning for histopathological image analysis: towards computerized diagnosis on cancers. In: Lu L, Zheng Y, Carneiro G, Yang L (eds) Deep learning and convolutional neural networks for medical image computing: precision medicine, high performance and large-scale datasets. Springer, Cham, pp 73–95
Yassin NIR, Omran S, El Houby EMF, Allam H (2018) Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. Comput Methods Programs Biomed 156:25–45. https://doi.org/10.1016/j.cmpb.2017.12.012
Youk JH, Gweon HM, Son EJ (2017) Shear-wave elastography in breast ultrasonography: the state of the art. Ultrasonography 36(4):300–309. https://doi.org/10.14366/usg.17024
Yousefi M, Krzyżak A, Suen CY (2018) Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. Comput Biol Med 96:283–293. https://doi.org/10.1016/j.compbiomed.2018.04.004
Zhang Q, Xiao Y, Dai W, Suo JF, Wang CZ, Shi J, Zheng HR (2016) Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics 72:150–157. https://doi.org/10.1016/j.ultras.2016.08.004
Zhang X, Zhang Y, Han EY, Jacobs N, Han Q, Wang X, Liu J (2017) Whole mammogram image classification with convolutional neural networks. Paper presented at the 2017 IEEE international conference on bioinformatics and biomedicine (BIBM)
Zheng Y, Jiang Z, Xie F, Zhang H, Ma Y, Shi H, Zhao Y (2017) Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. Pattern Recogn 71:14–25. https://doi.org/10.1016/j.patcog.2017.05.010

Affiliations

Ghulam Murtaza[1,2] · Liyana Shuib[1] · Ainuddin Wahid Abdul Wahab[1] · Ghulam Mujtaba[1] · Ghulam Mujtaba[3] · Henry Friday Nweke[1,4] · Mohammed Ali Al-garadi[5] · Fariha Zulfiqar[1] · Ghulam Raza[6] · Nor Aniza Azmi[7]

Ainuddin Wahid Abdul Wahab
ainuddin@um.edu.my

Ghulam Mujtaba
mujtaba@iba-suk.edu.pk

Ghulam Mujtaba
mujtaba1974@gmail.com

Henry Friday Nweke
henrynweke@siswa.um.edu.my

Mohammed Ali Al-garadi
mohammed.g@qu.edu.qa

Fariha Zulfiqar
farihazulfiqar@siswa.um.edu.my

Ghulam Raza
ghulam.raza@ymail.com

Nor Aniza Azmi
noraniza.azmi@ukm.edu.my

[1] Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

[3] Department of Computer Science, PAF-KIET, Karachi, Pakistan

[4] Computer Science Department, Ebonyi State University, P.M.B 053, Abakaliki, Ebonyi State, Nigeria

[5] Qatar University, Ibn Khaldoon Hall, Doha, Qatar

[6] St. James's Hospital, James's Street, Ushers, Dublin 8, Ireland

[7] Diagnostic Imaging and Radiotherapy Programme, Faculty of Health Sciences Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia