Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks

1,2 JEREMY R BURT, MD, 2 NESLISAH TOROSDAGLI, MSc, 2 NAJI KHOSRAVAN, BSc, 2 HARISH RAVIPRAKASH, MSc, 2 ALIASGHAR MORTAZI, MSc, 1 FIONA TISSAVIRASINGHAM, MD, 2 SARFARAZ HUSSEIN, MSc and 2 ULAS BAGCI, PhD

1 Department of Radiology, Florida Hospital, Orlando, FL, USA
2 Department of Computer Science, Center for Research in Computer Vision, University of Central Florida (UCF), Orlando, FL, USA

Address correspondence to: Dr Ulas Bagci
E-mail: ulasbagci@gmail.com

ABSTRACT

Deep learning has demonstrated tremendous revolutionary changes in the computing industry and its effects in radiology and imaging sciences have begun to dramatically change screening paradigms. Specifically, these advances have influenced the development of computer-aided detection and diagnosis (CAD) systems. These technologies have long been thought of as “second-opinion” tools for radiologists and clinicians. However, with significant improvements in deep neural networks, the diagnostic capabilities of learning algorithms are approaching levels of human expertise (radiologists, clinicians etc.), shifting the CAD paradigm from a “second opinion” tool to a more collaborative utility. This paper reviews recently developed CAD systems based on deep learning technologies for breast cancer diagnosis, explains their superiorities with respect to previously established systems, defines the methodologies behind the improved achievements including algorithmic developments, and describes remaining challenges in breast cancer screening and diagnosis. We also discuss possible future directions for new CAD models that continue to change as artificial intelligence algorithms evolve.

INTRODUCTION

Computer-aided detection and diagnosis (CAD) systems have been employed for over three decades and have often been considered as “second-opinion” tools. CAD systems work by utilizing radiographic images with known diagnostic features to train highly specialized software solutions that are equipped with machine learning and pattern recognition algorithms. These systems can then recognize the imaging patterns they were trained with, on test images (i.e. unseen or not used in training), allowing them to participate in detection and diagnosis of various diseases. These systems have the potential to be very useful in the field of oncology, aiding with improved detection and diagnosis of a wide variety of tumor types.1,2

There have been numerous studies in the literature investigating the use of CAD systems for breast cancer detection and diagnosis. These studies have used various imaging modalities and machine learning algorithms, some of which have even gone through clinical workflow for feasibility tests.3,4 However, the success of these studies has been limited due to high phenotypic variations in tumors, large number of false positives, and poor diagnosis rates.5 For these reasons, many studies were dedicated to improving these systems. Lately, research in this field is moving towards a more favorable direction due to exciting new advances in machine learning, specifically “deep learning”.6,7

Deep learning, i.e. as deep neural networks, has been a rapidly growing subfield of machine learning. The main reasons behind this breakthrough over the past few years are increased availability of more advanced computer algorithms that are inspired by human intelligence, updates on contemporary hardware technology for processing and storing large data sets, and an increased availability of massive amounts of labeled data to train these algorithms with better precision. This revolutionary and cutting-edge approach to computer vision has had a broad spectrum of applications including graphics, genetics, medicine, the automotive industry, the Internet, and ultimately, radiology and imaging sciences.8–12

In this review, we aim to evaluate the impact of deep learning based diagnostic systems that can help clinicians with screening and diagnosing breast cancer. Not only do we summarize the details of recently developed deep
learning based CAD systems, but we also explain various deep neural network designs for image analysis, explore the benefits and limitations of recently developed decision support systems, and elucidate future perspectives that radiologists and clinicians can benefit from in their routine diagnostic tasks. To address the concerns associated with conventional imaging techniques, CAD and decision support systems have made considerable advancements allowing for precise characterization of various pathologies by recognizing imaging features that are not easily visible to human eyes. These enhanced diagnostic capabilities allow for a reduced number of missed tumor cases and, ultimately, assistance in the diagnostic decision-making process.

UNDERSTANDING BREAST CANCER DIAGNOSIS IN THE DEEP LEARNING ERA

Machine learning algorithms such as Naive Bayes, Genetic Algorithms, Fuzzy Logic, Clustering, Neural Networks, Support Vector Machines, Decision Trees and Random Forests etc. have been used for more than two decades for detection, diagnosis, classification, and risk assessment of breast cancer. Figure 1 shows a representative comparison of conventional machine learning CAD systems and deep learning based CAD systems, both of which utilize radiographic images for breast cancer diagnosis. The conventional machine learning approach for image classification is trained using carefully designed hand-engineered features (e.g. visual descriptions such as elongation, sphericity, or low gradients in borders) that are learned from radiologists and can be coded into algorithms. In contrast, deep learning employs high-level imaging features from large sets of images for training purposes. The literature pertaining to these machine learning methodologies, prior to the deep learning era, is vast. Interested readers may refer to the literature 13–26 for further description of conventional machine learning methods in breast cancer, which include a large number of methods that are beyond the scope of this review.

Literature review and search strategy

For the literature survey, we used Pubmed™, IEEEXplore™, Google Scholar™, and ScienceDirect™ to search for publications relating to deep learning applications towards breast cancer detection and diagnosis. Keywords searched included "deep learning", "breast cancer", "breast tissue", "convolutional neural network", "machine learning", "diagnosis", and "detection". Only papers that were published in peer-reviewed conferences and journals were selected for review. Our search yielded 28 research articles.

Figure 1. Comparison of conventional machine learning approach vs deep learning based approaches. ROI, Region of interest.

Figure 2. Statistical distribution of 28 deep learning papers selected for this review. The distribution regarding imaging modalities (a), year of publication (b), deep learning applications (c), and type of publication (d) are shown.
Deep learning is useful in effectively extracting features for certain tasks, the remaining challenge is deciding which specific features should be extracted to feed into the algorithms that will generalise data, by enabling the computer to make new and improved decisions. The basis of these algorithms is that they have much improved universal approximation properties (i.e. the ability to represent any non-linear function/associations) by comparison. This is largely due to the use of large numbers of layers, allowing flexibility of approximations of different function classes. Convolutional Neural Networks (CNNs) are a subclass of deep neural networks that employ a special function class, called “convolution”, in their layers instead of direct multiplication operations. CNNs are biologically-inspired variants of neural networks and the convolution operation in their layers provides invariance (i.e. output remains unchanged regardless of changes in input measurement), sparsity of learned features (i.e. most of the entries of the image features/signals have a value of zero; hence, the computer needs to store only non-zero entries), and sharing of parameters, allowing the entire system to work more accurately and efficiently.

Machine learning enables computers with the capacity to tackle real-world problems. These systems are trained with representative data, and when new input information is received, they utilize computer algorithms to identify regularities and make outcome decisions/predictions. The basis of these algorithms and learned relationships, called features, is widely varied; they can be as simple as detecting differences in intensity values of individual pixels, or as complex as recognizing advanced relationships between position, texture, and shape of the tumors. Although machine learning is useful in effectively extracting features for certain tasks, the remaining challenge is deciding which specific features should be extracted to feed into the algorithms for accurate diagnosis. Deep learning, in this regard, provides a basis for developing new and improved algorithms that are better equipped to generalize data, by enabling the computer to build complex concepts out of simple ideas. In the following section, we describe basic definitions that will help radiologists and clinicians understand the basic principles of deep learning as applied to radiology image analysis and CAD development.

**Figure 2a** shows the distribution of imaging modality used for breast cancer detection and diagnosis, among which mammography was the most commonly used modality. The search was conducted to include all previous publications up to and including December 2017.

The distribution of the deep learning papers published over time can be seen in **Figure 2b**, where half of the studies found were published in the year 2016. **Figure 2c** shows the distribution of the studies based on the application of deep learning with respect to diagnosis, detection, or segmentation of breast cancer, diagnosis being the most commonly studied field of investigation. Furthermore, **Figure 2d** shows the distribution of these studies between peer-reviewed journals and conferences.

Five authors worked on the literature survey. The topics were divided into subsections of “detection” and “diagnosis”. Two authors worked on detection and three on diagnosis. Papers corresponding to each subsection were found from the above mentioned sources by members of each group independently. The senior authors of the study (i.e. the first and last) verified and reviewed the pool of papers from both subsections as well as resolved any disagreements.

Deep learning in simplified details

The 28 publications, we selected for review utilized various approaches and applications of deep learning towards breast cancer detection and diagnosis. In this section, we briefly outline the inner workings of deep learning.

**Neural networks**, with respect to artificial intelligence, are inspired by the biological basis of neural networks, in which neurons can sense their environment and communicate information to surrounding neurons. In artificial intelligence, neural networks are typically represented by layers. These layers are, essentially, computational functions that process input information, as it compares to training data, to predict an outcome (i.e. \( f(x) = y \), where \( x \) is the input information, and \( y \) is the outcome prediction). Input neurons can sense new data and pass information onto neurons within different layers, processing this information. Connections between neurons are called “synaptic weights”, which are coefficients used to amplify or dampen the input signal by multiplication, assigning significance to the input to obtain the corresponding output. The computational power of these networks relies on the extent of training data that is available, allowing these neural networks to update weights of the connections. Simple network structures with only a few layers are known as “shallow” learning neural networks, whereas network structures which employ numerous and large layers are referred to as “deep” learning neural networks.

**Deep neural networks** are distinct from ancestral neural networks in that they have much improved universal approximation properties (i.e. the ability to represent any non-linear function/associations) by comparison. This is largely due to the use of large numbers of layers, allowing flexibility of approximations of different function classes. Convolutional Neural Networks (CNNs) are a subclass of deep neural networks that employ a specialized mathematical function, called “convolution”, in their layers instead of direct multiplication operations. CNNs are biologically-inspired variants of neural networks and the convolution operation in their layers provides invariance (i.e. output remains unchanged regardless of changes in input measurement), sparsity of learned features (i.e. most of the entries of the image features/signals have a value of zero; hence, the computer needs to store only non-zero entries), and sharing of parameters, allowing the entire system to work more accurately and efficiently.

**Figure 3** illustrates the relationship between CNNs and the human vision recognition pathway. With CNNs, a given image (e.g. a sample breast image as shown in **Figure 3**) is represented by edges, curves, or lines (i.e. local features) in the initial layers. Further layers combine these features to detect geometric shapes, or other midlevel feature representations; this continues until the system ends up detecting entire sections of the input image (i.e. global features). CNNs have been shown to be very successful in numerous practical applications.

**End-to-End learning** (or training) often refers to the joint training of all parameters in a network such as the approach taken in Jia et al31, Mortazi et al32 and Sukhbaatar et al33. In neural networks, the input is accepted from one end, and the network produces an output at the other end. Training of parameters between these two ends (input to output) is called End-to-End training or learning.

A pre-trained network, as the name implies, uses a network that has been previously trained with images and has optimized parameters for the task it will be performing. If a pre-trained network is used, then the parameters can be used for testing without the need for training the entire system, which can otherwise be a costly endeavor in terms of computation. Pre-training network will tend to work if the target task is similar to the base task (i.e. base task is the one that the network is trained and features are learned from). When the target data set is significantly
smaller than the base data set, and the tasks are considerably different from each other (e.g. base network is trained to classify natural images while the target network is aimed to classify tumor images from mammography data), then the phenomena called transfer learning can be used to transfer the knowledge from base task (i.e. features) into the target task.

Data augmentation, in the context of neural networks, refers to increasing the number of data points. With images, particularly radiologic images, data augmentation works in the following ways: (i) rotating the images in several different angles and saving those images as new training samples, (ii) adding different levels of noise, (iii) cropping the images, and (iv) geometrically transforming the images (e.g. rotation, translation, and scaling) without distorting its structural integrity. Data augmentation can increase the data size by more than 1000-fold by concatenating operations defined in (i–iv). It is effective when using insufficient quantities of data to train a generalizable model. This is often the case for radiology images and applications where data augmentation acts as a regularization. 34–36

LITERATURE REVIEW: DEEP LEARNING FOR BREAST CANCER DETECTION AND DIAGNOSIS

Tables 1 and 2 summarize the deep learning methods for detection and diagnosis of breast cancer, respectively. Evaluation of detection and diagnosis in those studies were performed using the standard metrics such as sensitivity (true-positive rate), specificity (true-negative rate), false-positive rate, false-negative rate, receiver operating characteristics (ROC) and free-response receiver operating characteristics curves. With respect to tissue segmentation evaluation, the Dice Similarity Coefficient metric is used (Tables 1 and 2: fifth and sixth columns).

Breast cancer detection and diagnosis using a pre-trained network

Mammography

Literature on deep learning-based breast cancer detection from mammography shares very similar network designs: pre-trained network, data augmentation (or transfer learning), and extracting features to be used with classifiers such as support vector machine, random forest, or others. In other words, deep networks were used only for extracting discriminative features (i.e. imaging features that are unique to tumor type). Advantages and disadvantages of detection methods are summarized in Table 1.

The use of pre-trained networks for breast cancer diagnosis started in 2015 (Table 2). Most pre-trained networks are trained on the ImageNet data set owing to a large number of images (>1 million) and thousands of classes. 57 Several studies showed that pre-trained models could be used to boost classification results for mammograms. 6 However, there is no consensus on what features should be used for classification. 58–60 The only study, to our knowledge, that puts some sort of feature interpretations into the diagnostic task is by Becker et al. 44 Uniquely, the authors studied the relationship of breast density to classification accuracy and found that low density was easier to classify. Despite the reported effectiveness of transfer learning, studies (see also Table 2) share also a concern: the differences in source task and target task in transfer learning should be solved to have...
Table 1. A summary of different detection methods with data sets and results

<table>
<thead>
<tr>
<th>Paper</th>
<th>Approach</th>
<th>Data set</th>
<th># Images used</th>
<th>AUC</th>
<th>ACC.</th>
<th>Pros (+)/Cons (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dhungel et al⁴³</td>
<td>m-DBN and R-CNN + SVM</td>
<td>DDSM-BCRP/INbreast</td>
<td>79/115 cases</td>
<td>0.75</td>
<td>77.0%</td>
<td>+Public data set, transparent results</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Complex system (four modules)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Performance gain of cascade of two R-CNNs is not justified</td>
</tr>
<tr>
<td></td>
<td>Faster R-CNN</td>
<td>In-house</td>
<td>850</td>
<td>0.72</td>
<td></td>
<td>+Includes finer layer information</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+Classification based on BI-RADS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Patch-based (suboptimal global information, slow testing time)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Impact of preprocessing performance not evaluated</td>
</tr>
<tr>
<td>Akselrod-Ballin et al⁴⁷</td>
<td>CNN + RF</td>
<td>In-house</td>
<td>62,272 images</td>
<td>0.941</td>
<td></td>
<td>+Hand crafted, deep and combined feature performances compared</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Suboptimal candidate detector</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Patch-based (suboptimal global information, slow testing time)</td>
</tr>
<tr>
<td>Kooi et al⁴⁸</td>
<td>CNN (AlexNet and VGG)+SVM</td>
<td>mini-MIAS</td>
<td>200</td>
<td></td>
<td>60.01 and 64.52%</td>
<td>+Simple system</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+Presented performance gain of data-augmentation and balanced data set</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−SVM is not justified</td>
</tr>
<tr>
<td>Jadoon et al⁷</td>
<td>CNN-DW and CNN-CT</td>
<td>IRMA</td>
<td>2796 patches</td>
<td>0.81</td>
<td>81.83 and 83.74%, respectively</td>
<td>+SVM and soft-max compared</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Patch-based (suboptimal global information, slow testing time)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Hand-crafted features used (not-deep, and sub-optimal)</td>
</tr>
<tr>
<td>Samala et al⁴⁹</td>
<td>Various CNN comparisons</td>
<td>In-house</td>
<td>127</td>
<td>0.89</td>
<td>0.93</td>
<td>+216 CNN architectures compared</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Exhaustive search of the architecture</td>
</tr>
<tr>
<td>Samala et al⁴¹</td>
<td>CNN (four convolutional layers and three fully connected layers)</td>
<td>In-house</td>
<td>2282 digitized film and digital mammograms and 324 DBT</td>
<td>0.90</td>
<td></td>
<td>+Transfer learning from mammogram to DBT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−DBT is acquired with a different geometry prototype</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Testing time five times slower than other systems</td>
</tr>
<tr>
<td>Fotin et al⁴²</td>
<td>CNN AlexNet</td>
<td>In-house</td>
<td>2607</td>
<td></td>
<td>ROI Sensitivity of 93.0%</td>
<td>+Conventional CNN approach</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Patch-based (suboptimal global information, slow testing time)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Testing time five times slower than other systems</td>
</tr>
<tr>
<td>Ertosun et al⁴³</td>
<td>Two back-to-back CNNs</td>
<td>DDSM</td>
<td>2420 images</td>
<td></td>
<td>85%</td>
<td>+Comparisons among different architectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Back-to-back CNNs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−Complex system</td>
</tr>
</tbody>
</table>

ANN, artificial neural network; BCRP, breast cancer research program (DoD); CNN, convolutional neural network; DBT, digital breast tomosynthesis; DDSM, digital dataset for screening mammography; IRMA, image retrieval in medical applications; RF, random forest; SVM, support vector machine. TPR, true positive rate.
### Table 2. A summary of published diagnostic methods with data sets and results

<table>
<thead>
<tr>
<th>Paper</th>
<th>Approach</th>
<th>Data set</th>
<th># Images used</th>
<th>AUC</th>
<th>Acc (%)</th>
<th>Pros (+)/Cons (−)</th>
</tr>
</thead>
</table>
| Dheeba et al<sup>44</sup> | WNN                                    | In-house                           | 216           | –       | 85.4    | + Use of texture features and WNN  
|                      |                                        |                                    |               |         |         | − No empirical/theoretical evidence of preferring WNN to CNN.                   |
| Dhungel et al<sup>4</sup>   | CNN + RF                               | INBreast                           | 116           | –       | 95      | + Incorporates low and high level features  
|                      |                                        |                                    |               |         |         | − Requires both training from scratch and fine-tuning                         |
| Jiao et al<sup>45</sup>   | 18 layer CNN + SVM                     | DDSM                               | 600           | –       | 96.7    | + Deeper network  
|                      |                                        |                                    |               |         |         | − Remains subjective as requires gray-level intensities                       |
| Arevalo et al<sup>46</sup> | Two conv layers and FC layers          | BCDR                               | 736           | 0.82    | –       | + Pre-processing and extensive data-augmentation  
|                      |                                        |                                    |               |         |         | − No regularizer is used, so optimality of the method is unknown.              |
| Becker et al<sup>47</sup> | ANN                                    | In-house                           | 286           | 0.82    | –       | + Relationship between breast density and accuracy was studied.  
|                      |                                        |                                    |               |         |         | − Model details missing                                                         |
| Sahiner et al<sup>48</sup> | Texture feature images + layer CNN     | In-house                           | 168           | –       | 86.0    | + Comparison among different texture feature for classification  
|                      |                                        |                                    |               |         |         | − Using shallow network                                                        |
| Huynh et al<sup>49</sup> | CNN (AlexNet)                          | In-house                           | 607           | 0.86    | –       | + Increasing training data by using augmented ROI  
|                      |                                        |                                    |               |         |         | − Insufficient network depth, use extra features for classification           |
| Peterson et al<sup>50</sup> | Unsupervised sparse autoencoder        | Dutch-breast cancer screening data set/MMHS/Dutch-breast cancer screening data set | 493/668/1576  | 0.65    | –       | + Using whole image for train network with autoencoder features  
|                      |                                        |                                    |               |         |         | − Using shallow network for supervisor learning                               |
| Qiu et al<sup>51</sup>    | Three conv-max pooling layers and two FC layers | In-house                           | 270           | –       | 71.4    | + Using bilateral mammographic tissue density as features  
|                      |                                        |                                    |               |         |         | − Using shallow network                                                        |
| Carneiro et al<sup>52</sup> | CNN-F                                  | INBreast/DDSM                      | 115/172       | 0.91/ 0.97 | 78.1    | + Using pre-trained model and multiview  
|                      |                                        |                                    |               |         |         | − No comparison with other methods                                             |
| Li et al<sup>53</sup>     | 3D-CNN                                 | In-house                           | 143           | 0.801   | –       | + 3D model  
|                      |                                        |                                    |               |         |         | − Small real training data set for 3D                                         |
| Zhou et al<sup>54</sup>   | CNN (AlexNet)                          | In-house                           | 463           | –       | 76      | + Faster classification  
|                      |                                        |                                    |               |         |         | − Non-3D approach                                                              |
| Zhang et al<sup>55</sup>  | PGBM                                   | In-house                           | 227           | –       | 93.4    | + Multistage architecture  
|                      |                                        |                                    |               |         |         | − No empirical advantage of PGBM over CNN shown                               |
| Cheng et al<sup>56</sup>  | SDAE                                   | In-house                           | 10,133 slices | –       | 82.4    | + Slice selection strategy to account for depth  
|                      |                                        |                                    |               |         |         | − Needs pre-training and training                                              |
| Kooi et al<sup>57</sup>   | CNN + gradient boosted tree classifier | In-house                           | 1804          | 0.8     | –       | + Large data set for training  
|                      |                                        |                                    |               |         |         | − Requires both training from scratch and fine-tuning                         |

AUC, area under the curve; CNN, Convolutional Neural Network; PGBM, PointwiseGated Boltzmann Machine; RF, random forest; ROI, region of interest; SDAE, Stacked Denoising Autoencoders; WNN, Wavelet Neural Network.
an effective diagnosis system. Till now, the only viable way to approach this problem is to use data augmentation.

### Ultrasound

Another modality used for classification of breast tissues is ultrasound. To perform automatic classification using shear-wave elastography, Zhang et al proposed Pointwise Gated Boltzmann Machines-based approach, where local and global features were combined to identify tissue types. In another work, Cheng et al used Stacked Denoising Autoencoders. These methods are able to extract higher-level features but they are still using not so deep networks compared to conventional CNN-based methods. Thus, their accuracies were limited by the discriminative power of the extract features affected by the network choices.

### MRI

Recently, the application of deep learning was explored for tissue classification using MRI. Li et al proposed a three-dimensional (3D) CNN-based architecture for breast tumor classification in dynamic contrast enhanced MRI. A 3D-CNN consisting of 10 layers was employed for feature learning and classification. Experimental evaluations demonstrated that the 3D-CNN-based approach outperformed the two-dimensional-CNN based approach by around 8% in terms of AUC, showing a promising future for deep learning with MRI in breast cancer diagnosis.

### Breast cancer detection and diagnosis using end-to-end training

#### Mammography

Another trend in deep learning based cancer detection is to use end-to-end training instead of using pre-trained models. Briefly, in end-to-end learning/training, post-processing steps are added to have one system learn its parameters jointly instead of connecting multiple individual parts. Studies showed that classification accuracies with end-to-end training-based networks were higher than a single network implementation. Other than that, network architecture, the use of alternative classifiers, and combination of handcrafted features with deep features were conceptually identical to each other. These studies used the three publicly available data sets, Digital Dataset for Screening Mammography-Breast Cancer Research Program [DoD (Department of Defense)], Image Retrieval in Medical Applications, and INbreast. Although promising results have been shown in end-to-end training approaches (Table 1), each module in the end-to-end training should be made sure contributing to the whole learning process because interactions of such modules with each other can slow down the learning significantly if learning dynamics of each module is different. Hence, this can lead into difficulties in converging into an optimal training model.

Within the breast cancer classification studies (Table 2), end-to-end training has been used for the first time by Sahiner et al in 1996, where gray-level difference statistics of mammogram images and spatial gray-level dependence images were used to train a three layer CNN. Similarly, since then, various studies have investigated different handcrafted features to train CNNs. These methods conceptually work in a similar way: combining multiple modules and jointly training all parameters to demonstrate the power of end-to-end learning as a single architecture.

#### Digital breast tomosynthesis

An in-depth comparison of different CNN architectures was performed by Samala et al where authors automatically differentiated micro-calciﬁcation in digital breast tomosynthesis scans from other tissues. A total of 216 unique deep learning CNN architectures were trained by varying the number of filters, filter kernel sizes and partial sums. Although this study is one of the few that evaluated the impact of architecture on a deep learning application, it presents an inefficient solution for finding the best architecture. Instead, an automated search algorithm such as genetic algorithms or simulated annealing would be more efficient.

Another comparison study was conducted by Fotin et al where authors found that the deep learning approaches outperform conventional approaches in terms of sensitivity and specificity (Table 1).

### Breast tumor segmentation

#### Mammography

Deep learning-based mammogram segmentation approaches are not significantly different than patch-based CNN-based detection approaches. In Dhungel et al, authors extracted the most discriminative imaging features to classify each voxel either as normal or abnormal. Although this increases the specificity rates, false-positive findings increase as well. The authors had to utilize the last layer as a combination of random field and structural support vector machine. Similarly in Chuquicusma et al, a CNN with overlapping patches was used for segmentation to increase segmentation accuracy. However, a drawback to this approach was the complicated model, which included two different network structures, causing instability when training. Furthermore, the patch-based classification was not able to incorporate spatial constraints and a post-processing step was required, which puts additional computational cost into the framework.

#### MRI

Earlier segmentation studies used patch-based systems (i.e., CNNs were trained using patches) to delineate tumors. Although the accuracy of such studies was lower than clinically acceptable, the algorithm design is promising for future improvements. Recently, a U-net architecture has been used for image segmentation applications on MRI. The U-net is an architecture based on CNN and takes its name because of the “U” shape of the network. This network is specifically tailored for successful biomedical image segmentation applications. Although promising results were obtained for segmentation (as in Dalmis et al where breast and fibroglandular tissues were delineated), it should be noted that success of the U-net strongly depends on data augmentation procedure, and precise labeling of the tissues. It should be also noted that U-net has a slow convergence rate, raising questions about training efficiency.

### DISCUSSION AND CONCLUDING REMARKS

With the help of deep neural networks, the diagnostic capabilities of learning algorithms are approaching levels of human
Lack of imaging data in big data era

One of the major problems in developing deep learning based CAD systems for breast cancer is the lack of the sufficient data for training models with millions of parameters. Some approaches that address this issue include: (1) building and training a model with a very shallow network (only a few thousand parameters), (2) data augmentation.

Each of these approaches has their own drawbacks. In the first approach, limiting the number of parameters will lead to potentially significant inaccuracies. On the other hand, data augmentation will either add noise to the images or require sampling of overlapping image patches. The augmented samples, however, can be highly correlated with each other resulting in overfitting. Overfitting is a well-known machine learning and statistical modeling problem, which occurs when a learning model memorizes the training data and is not able to generalize to the new data. Moreover, local patches cannot incorporate the global and spatial context of the image, which can lead to inaccuracies.

Technical challenges of deep learning

There are challenges associated with the use of transfer learning in deep learning including architecture selection, number of examples sufficient to fine-tune, as well as the numbers of layers used on top of the pre-trained model. Moreover, the effectiveness of transfer learning decreases when the target task (mammogram diagnosis) is different from the source task (pre-trained network’s task). 65

The success of deep learning methods currently relies on high capacity models requiring several iterative updates across many labeled examples. However, obtaining millions of labeled examples is not an easy task, especially in the medical imaging field. Nonetheless, once these systems have been perfected, deep learning can be used in prevention and treatment programs for optimal results, radically transforming clinical practice and public health. In the following, we envision the potential of deep learning to transform other imaging modalities in the context of breast cancer detection and diagnosis.

Potential role of PET/MRI

The increasing availability of PET/MRI will most certainly lead to continued improvements in the accuracy of diagnosing breast cancer. PET/CT is limited by relatively high radiation dose and low spatial resolution resulting in inadequate sensitivity for detection of cancers ≤ 2 cm. 66 In contrast, Pinker et al showed that fused multiparametric MRI and PET imaging had significant improvements in accuracy with an AUC of 0.935 (0.835–1), when compared with delayed contrast enhanced MRI or PET alone. 67 This technique has the potential to significantly decrease the number of unnecessary biopsies. There are also advantages in combining the sensitivity of MRI, to determine the extent of disease, and the sensitivity of PET, to detect axillary and chest nodal disease. 68 No published studies to date have evaluated the use of machine learning and PET/MR for breast cancer.

What would it take for radiologists to accept deep learning tools for daily use?

The use of this advanced technology has the potential to update breast imaging techniques that have changed very little over the past 40–50 years. Many of our current practices in breast screening and diagnosis suffer from limited specificity, requiring an image-guided biopsy to reach a definitive diagnosis. Computer assisted diagnosis using machine learning, including deep CNNs, has the ability to efficiently make accurate diagnoses of breast pathology, potentially without the need of biopsy. With the help of user-interface development and commercialization, these artificial intelligence algorithms will certainly be part of an exciting future in breast imaging.

However, there are more steps that must be taken before radiologists accept these decision-support systems in their daily routine workflows. First, a global real-life application should account for widespread geographic, ethnic, and genetic variations. 69 Common cases in certain regions of the world may be quite rare in other areas. However, from the current hardware point of view, it may not be possible to train the deep-learning systems with a large amount of worldwide data. Sampling from this large database requires the criteria of “best representative” positive and negative samples of an illness. Thus, with the current hardware limitations, it may be beneficial to use locally trained versions of the same application, and to integrate and adapt their outputs as needed for the worldwide stage. The commercialized real-life applications should also make clear how to interpret the “rare cases”. In general, these applications are going to miss the “rare cases” for which they are not trained. For the time being, a critical step will be to have radiologists provide a final verification of the outcomes of the real-life applications. In addition, real-life applications should allow radiologists to upload new data into the training system. However, it is not clear yet how training can be repeated, from scratch or as a pre-trained network, by each piece of uploaded data, and how training duration and overfitting can be handled by frequent training with slightly modified data sets. Finally, moving these systems to a virtual cloud environment and having them accessible at any time would be a very useful feature.

How deep learning tools will impact the practice of radiology in breast cancer diagnosis and evaluation?

Success in the development of accurate deep learning algorithms for breast cancer diagnosis has the potential to significantly
impact the practice of radiology. These algorithms can aid radiologists in detecting and discriminating normal and suspicious tissues. Deep learning tools can provide important cues, which may be left out in a single screening session by the radiologist. Moreover, these tools can be applied to reduce the interpretation time required radiologists, thus improving clinical efficiency.

Recently, deep learning is being applied to generate artificial but realistic images across different applications of computer vision. As it is both cumbersome and time-consuming to manually annotate radiological images, the use of Generative Adversarial Networks can help generate better synthetic images to train deeper architectures.6,27 Moreover, these images can also be used to train radiology residents and interns.

Funding
The study is partially funded by CRCV, UCF.

Funding
The study is partially funded by CRCV, UCF.

References


