ORIGINAL ARTICLE

Inconsistent Performance of Deep Learning Models on Mammogram Classification

Xiaoqin Wang, MD, MS[a,b], Gongbo Liang, MS[c], Yu Zhang, BS[c], Hunter Blanton, BS[c], Zachary Bessinger, PhD[c], Nathan Jacobs, PhD[c]

Abstract

Objectives: Performance of recently developed deep learning models for image classification surpasses that of radiologists. However, there are questions about model performance consistency and generalization in unseen external data. The purpose of this study is to determine whether the high performance of deep learning on mammograms can be transferred to external data with a different data distribution.

Materials and Methods: Six deep learning models (three published models with high performance and three models designed by us) were evaluated on four different mammogram data sets, including three public (Digital Database for Screening Mammography, INbreast, and Mammographic Image Analysis Society) and one private data set (UKy). The models were trained and validated on either Digital Database for Screening Mammography alone or a combined data set that included Digital Database for Screening Mammography. The models were then tested on the three external data sets. The area under the receiver operating characteristic curve (auROC) was used to evaluate model performance.

Results: The three published models reported validation auROC scores between 0.88 and 0.95 on the validation data set. Our models achieved between 0.71 (95% confidence interval [CI]: 0.70-0.72) and 0.79 (95% CI: 0.78-0.80) auROC on the same validation data set. However, the same evaluation criteria of all six models on the three external test data sets were significantly decreased, only between 0.44 (95% CI: 0.43-0.45) and 0.65 (95% CI: 0.64-0.66).

Conclusion: Our results demonstrate performance inconsistency across the data sets and models, indicating that the high performance of deep learning models on one data set cannot be readily transferred to unseen external data sets, and these models need further assessment and validation before being applied in clinical practice.

Key Words: Deep learning, mammogram, performance inconsistency

J Am Coll Radiol 2020;■:■-■.

INTRODUCTION

Recently, deep learning, fueled with high-quality data, has demonstrated revolutionary potential in imaging recognition and has been rapidly applied in medical image research [1-6]. Numerous deep learning models have been developed for automatic lesion detection and classification on radiological images with performance that surpasses that of human radiologists [7-12]. For example, deep learning models on mammogram classifications often report promising results with the area under the receiver operating characteristic curve (auROC) above 0.90 [13-18], which is significantly better than the performance of human experts (0.71) [19]. These research breakthroughs have led to a faster than ever pace of product development for artificial intelligence (AI)-based radiology image analysis tools, and many commercial products are on the market. However, the superhuman or near-expert performance of these models, which are trained and tested on a particular crafted data set, may not be achieved on unseen external data sets [20,21]. This

[a]Department of Radiology, University of Kentucky, Lexington, Kentucky.
[b]Markey Cancer Center, University of Kentucky, Lexington, Kentucky.
[c]Department of Computer Science, University of Kentucky, Lexington, Kentucky.
Corresponding author and reprints: Xiaoqin Wang, University of Kentucky, Department of Radiology, 800 Rose Street, Lexington, KY 40506; e-mail: Xiaoqin.wang@uky.edu.

This work was supported by grant No. IRG 16-182-28 from the American Cancer Society (PI-Xiaoqin Wang) and grant No. IIS-1553116 from the National Science Foundation (PI- Jacob Nathan).
The authors state that they have no conflict of interest related to the material discussed in this article.

1546-1440/20/$36.00 ■ https://doi.org/10.1016/j.jacr.2020.01.006

potential performance uncertainty raises the concern of model generalization and validation, which needs to be addressed before the models are rushed to real-world clinical practice.

The goal of this study is to determine whether the high performance of deep learning models achieved on one data set persists when exposed to data sets that the models have not "seen" before. When designing this study, we wanted to include commonly used deep learning methods for medical image analysis, published models that claim to have excellent performance, and different image annotation methods used in the model training stage and to compare models in an intuitive and straightforward way.

We chose the task of breast cancer classification because breast cancer is commonly diagnosed in women throughout the world [22-25]. Also, a large number of standardized mammograms with structured reports have accumulated in clinical practice due to the 1992 Mammography Quality Standards Act and the 2003 ACR's BI-RADS. This wealth of high-quality image data with ground truth provide a rich base for research on deep learning of mammograms [26-29].

MATERIALS AND METHODS

In this study, we investigated the performance consistency of six deep learning models in four mammogram data sets. The mammograms containing calcifications, masses, or both were classified as benign or malignant. The models are two *end2end* [30], one *frcnn_cad* [6], and three new models we designed. First, we trained all models on the Digital Database for Screening Mammography (DDSM) data set because it is the largest publicly available mammogram database. Then, we validated the accuracy of each model and recorded the results on DDSM. Finally, we tested each model on three different data sets and compared these results with the validation results to assess performance consistency.

Data Sets

Four data sets from different patient populations were used in the study, namely, DDSM [31] from the United States, INbreast from South America [32], Mammographic Image Analysis Society (MIAS) from the United Kingdom [33], and UKy, a private data set recently collected in the United States (Table 1). The data used in this study include digital screen film mammogram and full-field digital mammogram from 1994 to 2017. Mammogram from contralateral breast is not included in all the data sets except UKy.

DDSM. The DDSM data set initially constructed in 1999 contains 2,620 labeled cases, including 10,480 digital screen film mammography images [31]. The DDSM cases come from the Massachusetts General Hospital mammography program and Wake Forest University School of Medicine mammography program. The films were scanned nonuniformly with scanners at different institutions with different gray levels to optical density mapping [34]. DDSM, as the largest publicly available mammography data set, is widely used for developing deep learning models.

We chose to use the Curated Breast Imaging Subset of DDSM [35], which is an updated and standardized version of the original DDSM, for training and testing of our three models. After converting the image pixel values into optical density values, we partitioned the data set into training and testing sets with a ratio of 4:1, respectively. The *end2end* model was also trained and validated using DDSM; however, the details of how the training and testing sets were split were not available to us. The whole DDSM data set was used for training for the *frcnn_cad* model.

MIAS. This data set is a digitalized mammography data set that was generated by the Mammographic Image Analysis Society in 1994 [33]. The films were taken from the United

Table 1. Data set details

Data set	DDSM	INbreast	MIAS	UKy
Origin	United States	South America	United Kingdom	United States
Use in this study	Training and validation	Testing	Testing	Testing
Year(s)	1999	2008-2010	1994	2014-2017
Type of data	DSFM	FFDM	DSFM	FFDM
Total no. of cases	2,620	115	161	507
Malignant; benign	35%; 65%	26%; 72%	16%; 84%	58%; 42%
No. of images	10,480 (Total)	387 of 410	322	1,872

DDSM = Digital Database for Screening Mammography; DSFM = digital screen film mammogram; FFDM = full field digital mammogram; MIAS = Mammographic Image Analysis Society.

FLA 5.6.0 DTD ■ JACR5065_proof ■ 7 March 2020 ■ 12:12 am ■ ce

Kingdom National Breast Screening Program and were digitized to a 50-μm pixel edge. The database contains 322 digital films with radiologist labeling at the locations of abnormalities and is available on 2.3-GB 8-mm (exabyte) tape. The database was reduced to a 200-μm pixel edge so all the images are 1,024 × 1,024 [33]. We used MIAS as the external testing data set for all the models except *frcnn_cad* because *frcnn_cad* had been trained with MIAS [6].

INbreast. The INbreast data set contains 115 full-field digital mammogram cases (410 images) with BI-RADS assessment scores and histological findings for cancers [32]. Because no pathological result for each image is available, like in the literature, we categorized images that have BI-RADS assessment of 1 and 2 as benign and images that have BI-RADS assessment of 4, 5, 6 as positive. Therefore, 387 images were used in this study. Again, we used the INbreast data set as the external testing set for all models except *frcnn_cad* because it has been exposed to INbreast during the model training [6].

UKy. The UKy data set is a private clinical data set. The mammogram data were retrospectively collected from patients seen at a comprehensive breast imaging center in the United States from January 2014 to December 2017. The data set contains 1,093 benign and 779 malignant full-field digital mammogram images from 507 patients. The images were acquired with a Hologic device (Marlborough, Massachusetts) in 12-bit Digital Imaging and Communications in Medicine (DICOM) format at a resolution of 33,280 × 4,096, which was down sampled to 832 × 832. Each image was reviewed by breast radiologists and proved with biopsy. The data set contains many complicated cases, such as patients with history of benign biopsy, markers, or prior surgery, including lumpectomy. We used UKy as the external testing data set for all evaluated models.

Models

We evaluated six deep learning architectures, three designed by us and three published by other researchers These deep learning models have various but commonly used model design strategies, training techniques, and annotation methods.

The *end2end* and three new models we designed used a transfer learning technique, which pretrains models on the natural image domain (ImageNet data set) [36] and transfers the models to another imaging domain later. This technique has been proven to alleviate the problem of small data size [37,38]. The *frcnn_cad* model used the instance-based learning method, which is a widely used deep learning method for the object detection with proven success in multiple image domains [39,40].

To train a breast cancer classifier, we needed the training samples with accurate labels. Two labeling methods are often used for image samples: bounding box for lesions and coarse annotation for whole images (an image-level label). The three models designed by us used image-level labels for training; the other models used bounding box for training.

The three deep convolutional neural network (CNN) models that we designed contain two components: a deep feature extractor and a CNN classifier (Fig. 1). We selected three popular CNN models as the feature extractors in this study: AlexNet [41], VGG16 [42], and ResNet50 [43]. The feature extracting CNN models were pretrained on the ImageNet data set [36] and used to extract the image features from mammograms. The classifier then used the mammographic feature maps as input to predict whether the mammogram image was benign or malignant. More details about the CNN feature extractors and classifiers, as well as model implementation, can be found in the e-only appendix.

The *end2end* models were released by Shen [30] and used VGG16 and ResNet50 as their backbone. The models were pretrained on DDSM with bounding box. The author reported that the best single model achieved a 0.88 auROC on the DDSM validation set [30].

The *frcnn_cad* model is a model that used the Region (R)-CNN approach. The base CNN was a VGG16 network with 16 layers [6]. The model was proposed by Ribli et al for breast lesion classification and detection [6]. This method won second place in the Digital Mammography Dialogue for Reverse Engineering Assessments and Methods, or DREAM, Challenge in 2017. The *frcnn_cad* model used the entire mammography image as input and performs instance-based detection. The released model was trained with DDSM, MIAS, and a private data set. The model achieved 0.95 auROC on the validation database (INbreast).

Comparison Method

We compared the DDSM-trained deep CNN binary classifier (malignant versus benign) performance of the internal validation data set and the external testing data sets. More specifically, for each of the models, we generated four sets of prediction results: the validation result from the internal data set and three sets of testing results from the external data sets. Then, we compared the auROC of testing results with the validation result. Ideally, if a model has good generalization ability, the auROCs of the testing results should be similar to the validation results.

Training with a Mixed Data Set

In this experiment, we also trained one of our backbone models with a mixed data set. AlexNext was selected for this

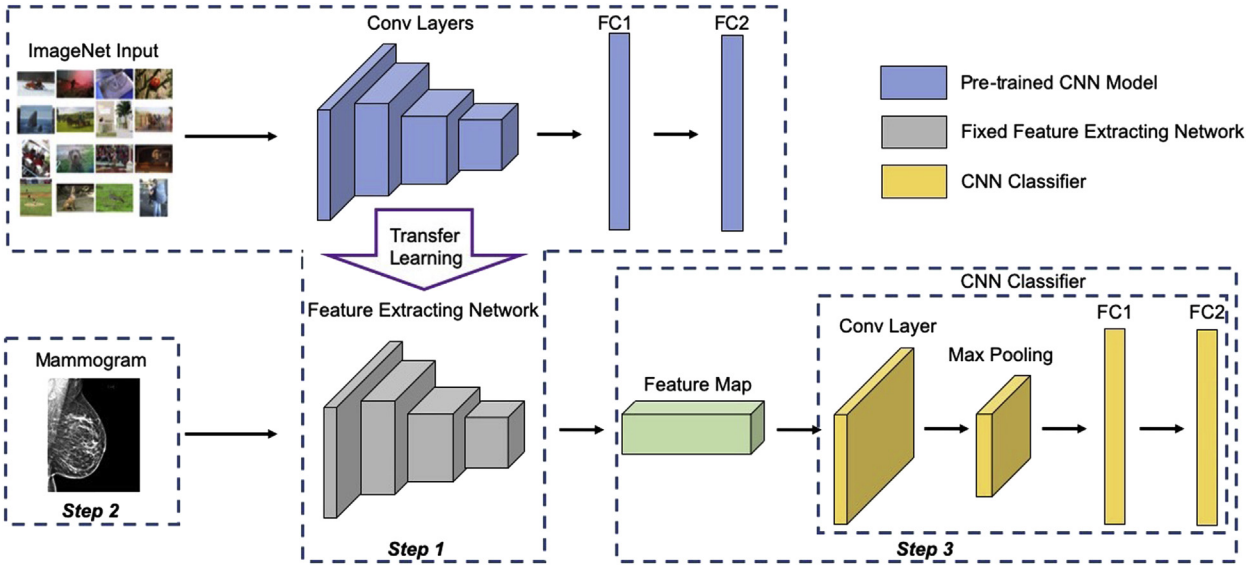

Fig 1. Stepwise illustration of the transfer learning deep convolutional neural network (CNN) models: (1) pretrain the CNN models with natural images from ImageNet and use them as the features extractors; (2) Feed the mammograms into the feature extractors; (3) Classify the feature maps using a CNN classifier. The blue component is the pretrained model. The gray component is the feature extractor. The yellow component is the CNN classifier. Note: FC1 = Fully Connected Layer 1 and FC2 = Fully Connected Layer 2.

test because it is one of the algorithms that has been most widely used for image classification. We first randomly split each of the external testing data sets (INbreast, MIAS, and UKy) into two subsets (denoted as $subset_1$ and $subset_2$) with a 4:1 ratio. Then, we combined the DDSM training set with the larger subset ($subset_1$) of each testing data set to form a new training data set, denoted *MixedData*, and used it to train a new model. The model was validated on the DDSM test set and tested on $subset_2$ of each external testing data set.

RESULTS

Tables 2 and 3 show the validation and real testing results, respectively, of the six models. The validation and testing results of our models as well as the testing results of *end2end* and *frcnn_cad* were generated from our experiments. The validation results of *frcnn_cad* on INbreast and *end2end* on DDSM were taken from the corresponding publications [6,30].

Table 2. Validation results of deep learning models

Model	Mean auROC		
	AlexNet	VGG16	ResNet50
Ours	0.71 ±0.01	0.75 ±0.01	0.79 ±0.01
end2end		0.88	
Frcnn_cad	–	0.95	–

auROC = area under the receiver operating characteristic curve.

Table 2 reveals that the validation auROC of our transfer learning models with backbones of AlexNet, VGG16, and Resnet50 on DDSM are 0.71, 0.75, and 0.79, respectively. The reported validation results (auROC) of the *end2end* is 0.88, which surpasses human experts' performance by 28%. The *frcnn_cad* model has the highest internal validation result (0.95), which is 34% higher than expert radiologists.

However, the testing results of all models on all the external data sets significantly decreased from the validation results, regardless of the model design, transfer

Table 3. Testing results of models trained on the DDSM data set

Model	Testing Set	Mean auROC		
		AlexNet	VGG16	ResNet50
Ours	MIAS	0.46 ±0.01	0.58 ±0.01	0.54 ±0.01
	INbreast	0.51 ±0.01	0.48 ±0.01	0.51 ±0.01
	UKy	0.49 ±0.01	0.53 ±0.01	0.44 ±0.01
end2end	MIAS	–	0.63 ±0.01	0.65 ±0.01
	INbreast	–	0.53 ±0.01	0.60 ±0.01
	UKy	–	0.46 ±0.01	0.48 ±0.01
frcnn_cad	UKy	–	0.48 ±0.01	–

auROC = area under the receiver operating characteristic curve; DDSM = Digital Database for Screening Mammography; DSFM = digital screen film mammogram; FFDM = full field digital mammogram; MIAS = Mammographic Image Analysis Society.

FLA 5.6.0 DTD ■ JACR5065_proof ■ 7 March 2020 ■ 12:12 am ■ ce

Table 4. Testing result of model trained on mixed data set

Backbone Model	Mean auROC		
	Validation Result	Testing Set	Testing Result
AlexNet	0.69 ±0.01	MIAS	0.58 ±0.01
		INbreast	0.67 ±0.01
		UKy	0.68 ±0.01

auROC = area under the receiver operating characteristic curve; MIAS = Mammographic Image Analysis Society.

learning method, or image labeling technique. The testing results of the six models were between 0.44 (95% confidence interval [CI]: 0.43-0.45) and 0.65 (95% CI: 0.64-0.66), with a mean and median of 0.52 and 0.51, respectively, indicating essentially random performance (Table 3). None of the evaluated models performs at a near-human level on any of the external testing sets. The highest external testing auROC score (0.65; 95% CI: 0.64-0.66) was achieved with the ResNet50 version of the *end2end* model on the MIAS data set, which is still lower than the auROC of 0.71 achieved by human experts. Figure 2 shows the auROC curves of our three models on the external testing data sets. The dashed (diagonal) line represents 0.5 auROC, which indicates a random performance. For all three models, the validation results on DDSM (blue line) are significantly above the dashed line. However, all the other lines are close to the dashed line, which means none of the testing results from the external testing data sets are better than random performance or close to the validation results.

Table 4 demonstrates the testing result of our AlexNet backbone model trained with a mixed data set. The table reveals that by including the partial data from INbreast, MIAS, and UKy in the training set, the result of each testing data set is closer to the validation result and the model performance across those data sets is significantly improved.

DISCUSSION

Our results show that the high performance of deep CNN models achieved on the training mammographic data sets cannot be readily transferred or generalized to unseen external data sets, regardless of model architecture, training techniques, or data labeling methods. Our results are consistent with recent literature on AI applications in multiple medical imaging modalities by other researchers [20,21,26,27].

The consistently significant performance decrease is likely caused by a data distribution mismatch between the training and testing data. Generally speaking, deep learning algorithm models the training data distribution and assumes the distribution of the testing data is similar. The model will have relatively good performance only if the distribution is similar. Most available medical imaging data used for the model training are small, which neither cover the distribution of the external testing data nor reflect the complexity of real-world settings. In our study, the DDSM data set is quite different from the testing data in many aspects. First, the imaging acquisition and processing techniques for each data set are different. For instance, DDSM is a screen film mammography data set, but INbreast and UKy are digital mammography data sets, which have different optical density response to the x-ray exposure [34]. In addition, with machines from different vendors, the image acquisition parameter choices could also affect the image radiomic features [44]. Second, the imaging phenotype of different breast lesions on mammograms can also affect performance of the models. Breast cancer, like other cancers, is a heterogeneous disease with various intrinsic pathological types and imaging phenotypes. The four data sets used in this study include a variety of patient populations and demographics, which may have different disease prevalence. The images appearing in the training data set, DDSM, may not represent all the pathological types and phenotypes in the external data sets. Some rare breast disease may not be included in the training data, but it may be present in the testing data. In addition,

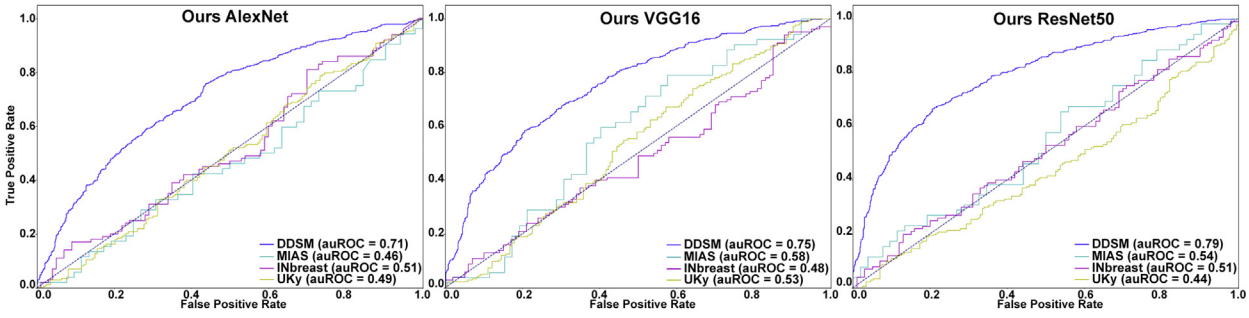

Fig 2. The auROC curves of our models. auROC =area under the receiver operating characteristic curve; DDSM = Digital Database for Screening Mammography; MIAS = Mammographic Image Analysis Society.

FLA 5.6.0 DTD ■ JACR5065_proof ■ 7 March 2020 ■ 12:12 am ■ ce

clinical scenarios in the real world can be more complex. In a study comparing the subtlety of malignant lesions in the DDSM and UKy images, radiologists determined that the appearance of lesions is more subtle on the UKy mammograms [34], which may indicate that the UKy data set, recently collected from a comprehensive breast care center, is more challenging to classify. Finally, because of the "black box" nature of the learning process of deep learning models, it is unclear how models "see" the image features and make decisions based on them. It is also possible that some imaging features in each data set are intrinsic to the computer models but are not yet recognized or explainable by humans.

Our results clearly indicate the need for external validation of model performance. However, it is known that clinical validation for the AI models is largely lacking [45]. A recent meta-analysis shows that only 6% of published AI algorithm studies performed external validation [46], and yet some high-performance models have received FDA approval despite the uncertainty of the model performance [47]. The rapid development of the AI models poses significant regulatory challenges. Guidelines are required for validating AI models across different imaging platforms and patient populations [48]. However, no clear pathways or standards have been established for clinical validation of the AI models even though numerous AI medical solution products are already in the markets.

Our study also reveals that by including partial data of INbreast, MIAS, and UKy in the training set, the model performance on those data sets can be significantly improved and the result of each testing data set is closer to the validation result. This performance improvement indicates that if the training data distribution covers the distribution of the testing data, the performance of the CNN model will improve. This may provide a pathway to optimize the AI models before clinical adoption. These models will likely perform well in clinic if the high performance persists after training with a small of real-world clinical data. As consumers, it is likely that radiologists will play a leading role in medical applications of AI [49].

Our study is limited by the common challenge of deep learning—small data size with likely limited data distribution in all the data set. Breast density and tumor size discrepancies between training and testing data sets may also affect model performance; however, this information is not reviewed in this study.

Clinical Relevance

Despite excitement regarding AI development in radiology, AI models may not perform well in real-world practice. When reviewing AI research in radiology, the future users of AI—radiologists—should understand that model performance strongly depends on the training data. Consumers of available AI products should be aware of generalizability issues and ensure that algorithms perform as expected at their own institutions before purchasing AI tools—even if they are FDA cleared.

In conclusion, our results demonstrate that deep learning models trained on a limited data set do not perform well on data sets that have different data distributions in patient population, disease characteristics, and imaging systems. This high variability in performance across mammography data sets and models indicates that the proclaimed high performance of deep learning models on one data set may not be readily transferred or generalized to external data sets or modern clinical data that have not been "seen" by the models. Fortunately, our study shows that training with mixed data may improve model performance. Guidelines and regulations are needed to catch up with the AI advancement to ensure that models with claimed high performance on limited training data undergo further assessment and validation before being applied to real-world practice.

TAKE-HOME POINTS

- Numerous deep learning models for automatic breast lesion classification on mammograms have reported exciting performance surpassing that of expert radiologists.
- Our results demonstrate high variability in performance across the mammography data sets and models, which indicates that the high performance of deep learning models on one limited data set cannot be readily transferred to unseen external data sets with different data distribution.
- Radiologists, as consumers of available AI products, should be aware of generalizability issues and ensure algorithms performance are validated at their own institutions before purchasing AI tools—even if FDA cleared.

ACKNOWLEDGMENTS

This work was supported by Grant No. IRG 16-182-28 from the American Cancer Society (principal investigator: Xiaoqin Wang) and Grant No. IIS-1553116 from the National Science Foundation (principal investigator: Jacob Nathan). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the American Cancer Society or National Science Foundation.

FLA 5.6.0 DTD ■ JACR5065_proof ■ 7 March 2020 ■ 12:12 am ■ ce

The Markey Cancer Center's Research Communications Office assisted with preparation of this manuscript.

ADDITIONAL RESOURCES

Additional resources can be found online at: https://doi.org/10.1016/j.jacr.2020.01.006.

REFERENCES

1. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 2019;292:331-42.
2. Choe J, Lee SM, Do KH, et al. Deep learning-based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. Radiology 2019;292:365-73.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
4. Liang G, Fouladvand S, Zhang J, Brooks M, Jacobs N, Chen J. GANai: standardizing CT images using generative adversarial network with alternative improvement. 2019 IEEE International Conference on Health Informatics (ICHI). Xi'an, Shanxi, China, June, 10-13, 2019.
5. Mihail RP, Liang G, Jacobs N. Automatic hand skeletal shape estimation from radiographs. IEEE Trans Nanobioscience 2019;18:296-305.
6. Ribli D, Horvath A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep 2018;8:4165.
7. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 2016;35:1207-16.
8. Gupta H, Jin KH, Nguyen HQ, McCann MT, Unser M. CNN-based projected gradient descent for consistent CT image reconstruction. IEEE Trans Med Imaging 2018;37:1440-53.
9. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
10. Lee H, Mansouri M, Tajmir S, Lev MH, Do S. A deep-learning system for fully-automated peripherally inserted central catheter (PICC) tip detection. J Digit Imaging 2018;31:393-402.
11. Tseng K, Lin Y, Hsu W, Huang C. Joint sequence learning and cross-modality convolution for 3D biomedical segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, July, 10-13, 2017:3739-3746.
12. Yang Q, Yan P, Zhang Y, et al. Low-Dose CT Image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans Med Imaging 2018;37:1348-57.
13. Dhungel N, Carneiro G, Bradley A. Automated mass detection in mammograms using cascaded deep learning and random forests. 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA). Adelaide, Australia. November 23-25, 2015.
14. Jadoon MM, Zhang Q, Haq IU, Butt S, Jadoon A. Three-class mammogram classification based on descriptive CNN features. Biomed Res Int 2017;2017:3640901.
15. Lévy D, Jain A. Breast mass classification from mammograms using deep convolutional neural networks. 2016. arxiv.org/abs/1612.00542.
16. Mendel K, Li H, Sheth D, Giger M. Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. Acad Radiol 2019;26:735-43.
17. Yi D, Sawyer RL, Cohn D, Dunnmon J, Lam C, Xiao X, Rubin D. Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. 2017. arXiv:170506362.
18. Dhungel N, Carneiro G, Bradley A. Deep learning and structured prediction for the segmentation of mass in mammograms. Med Image Comput Comput Assist Interv 2015;9349:605-12.
19. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. AJR Am J Roentgenol 2014;203:909-16.
20. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:e1-36.
21. Abdelhafiz D, Yang C, Ammar R, Nabavi S. Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinformatics 2019;20:281.
22. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
23. McGuire S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. Adv Nutr 2016;7:418-9.
24. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7-30.
25. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7-34.
26. Gardezi SJS, Elazab A, Lei B, Wang T. Breast cancer detection and diagnosis using mammographic data: systematic review. J Med Internet Res 2019;21:e14464.
27. Harvey H, Heindl A, Khara G, et al. Deep learning in breast cancer screening. In: Ranschaert ER, Morozov S, Algra PR, eds. Artificial intelligence in medical imaging: opportunities, applications and risks. Switzerland AG: Springer, Cham; 2019:187-215.
28. Tsochatzidis L, Costaridou L, Pratikakis I. Deep learning for breast cancer diagnosis from mammograms—a comparative study. J Imaging 2019;5.
29. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening [E-pub ahead of print]. IEEE Trans Med Imaging 2019. https://doi.org/10.1109/TMI.2019.2945514.
30. Shen L. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. Sci Rep 2017;9.
31. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. Proceedings of the 5th International Workshop on Digital Mammography. 212-218, Medical Physics Publishing, Madison, 2001.
32. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol 2012;19:236-48.
33. Suckling J, Parker J, Dance D, et al. (2015). Mammographic Image Analysis Society (MIAS) database v1.21 [Dataset].
34. Chen Q, Liu J, Luo K, Zhang X, Wang X. Transfer deep learning mammography diagnostic model from public datasets to clinical practice: a comparison of model performance and mammography datasets. The Fourteenth International Workshop on Breast Imaging (IWBI 2018):10718. International Society for Optics and Photonics. Bellingham, USA, 2018.
35. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data 2017;4:170177.
36. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, June 20-25, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.
37. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. MIT Press, 2016.
38. Olivas ES. Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, 2009.
39. Dai J, Li Y, He K, Sun J. R-FCN: Object detection via region-based fully convolutional networks. 2016. arxiv.org/abs/1605.06409.

40. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271. Honolulu, Hawaii, July 22-25, 2017.
41. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60: 84-90.
42. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arxiv.org/abs/1409.1556.
43. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016:770-8. Las Vegas, Noveda, June 26th-July 1st, 2016.
44. Liang G, Zhang J, Brooks MA, Howard J, Chen J. Radiomic features of lung cancer and their dependency on CT image acquisition parameters. Med Phys 2017;44:3024.
45. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. Radiology 2019;293:246-59.
46. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20:405-10.
47. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-10.
48. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, Brink J. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018;15: 504-8.
49. Rubin DL. Artificial intelligence in imaging: the radiologist's role. J Am Coll Radiol 2019;16:1309-17.