HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

Comput Biol Med. 2019 June ; 109: 85–90. doi:10.1016/j.compbiomed.2019.04.018.

Deep Learning for identifying radiogenomic associations in breast cancer

Zhe Zhu[a,*], **Ehab Albadawy**[a], **Ashirbani Saha**[a], **Jun Zhang**[a], **Michael R. Harowicz**[a], **Maciej A. Mazurowski**[b]

[a]Department of Radiology, Duke University, USA

[b]Department of Radiology and Department of Electrical and Computer Engineering, Duke University, USA

Abstract

Rationale and Objectives: To determine whether deep learning models can distinguish between breast cancer molecular subtypes based on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI).

Materials and methods: In this institutional review board–approved single-center study, we analyzed DCE-MR images of 270 patients at our institution. Lesions of interest were identified by radiologists. The task was to automatically determine whether the tumor is of the Luminal A subtype or of another subtype based on the MR image patches representing the tumor. Three different deep learning approaches were used to classify the tumor according to their molecular subtypes: learning from scratch where only tumor patches were used for training, transfer learning where networks pre-trained on natural images were fine-tuned using tumor patches, and off-the-shelf deep features where the features extracted by neural networks trained on natural images were used for classification with a support vector machine. Network architectures utilized in our experiments were GoogleNet, VGG, and CIFAR. We used 10-fold crossvalidation method for validation and area under the receiver operating characteristic (AUC) as the measure of performance.

Results: The best AUC performance for distinguishing molecular subtypes was 0.65 (95% CI: [0.57,0.71]) and was achieved by the off-the-shelf deep features approach. The highest AUC performance for training from scratch was 0.58 (95% CI:[0.51,0.64]) and the best AUC performance for transfer learning was 0.60 (95% CI:[0.52,0.65]) respectively. For the off-the-shelf approach, the features extracted from the fully connected layer performed the best.

*Corresponding author. 2424 Erwin Rd. (Hock Plaza), Suite 302, Durham, NC, 27705, USA. zhe.zhu@duke.edu (Z. Zhu).

Conflict of interest

The authors have no conflict of interest to disclose.

Informed consent

Informed consent was obtained for experimentation with human subjects.

Conclusion: Deep learning may play a role in discovering radiogenomic associations in breast cancer.

Keywords

Deep learning; radiogenomic; breast cancer subtype

Introduction

Molecular classification into intrinsic subtypes has led to significant advances in the field of breast cancer. Each distinct molecular subtype is associated with a tendency of disease progression, which is why treatment recommendations from physicians hinge on the genomic analysis of each patient's tumor. Recently, the field of radiogenomics or imaging genomics has emerged, which aims at finding correlations between the imaging characteristics of cancer and its genomic composition. A specific area that has garnered significant attention is the prediction of genomics in breast cancer using MRI[1-16]. Previous work on this topic utilized either imaging features manually extracted by radiologists, which is a very time consuming and subjective process, or features automatically extracted by computer algorithms. Such features include tumor texture (e.g. Haralick features), tumor shape, or enhancement dynamics[6]. While this has shown promising results, hand-crafted features are limited because they require the researcher to anticipate which characteristics might be of use for a given tumor.

Deep learning approaches have shown superiority to handcrafted approaches in automatic feature extraction[17], image classification[18] and object detection[19]. In this study we propose using deep learning to conduct radiogenomic analysis of breast cancer. Specifically, we studied three different deep learning approaches: training from scratch, transfer learning and off-the-shelf deep features. Training from scratch is the most straightforward way of training deep neural networks when there are enough training data. In many medical imaging tasks, however, the number of training samples is insufficient for this method. In these situations using the training from scratch approach can cause overfitting. One way to alleviate the issue of limited data is to use transfer learning. The transfer learning approach initializes the network using a model pre-trained using different data (e.g., natural images) and then additionally trains the network using the specific data for the task at hand. Another option when faced with limited data is to use the deep features approach, which utilizes the pre-trained network as the feature extractor. Afterwards a traditional classifier such as a support vector machine is trained on the extracted features. Each of these approaches have been proven to work well on a specific subset of medical imaging tasks[20-22] .

Materials and Methods

Patient Population

Design and execution of this study was approved by institutional review board. We collected consecutive preoperative dynamic contrast enhancement MRIs of 400 patients at Duke University Medical Center acquired from September 2007 to June 2009. Then, we excluded 114 patients for the following reasons: 19 had a previous history of breast cancer, 19 had a

history of benign elective breast surgery, 29 were undergoing breast cancer treatment at the time of MRI, 42 had missing pathology data, 3 had missing sequences, 1 had a discordant number of slices in pre-contrast and post-contrast sequences, 1 had no biopsy performed. The remaining 286 cases were split into 6 subsets without overlap and were annotated by six breast imagers. 11 cases were skipped by these readers (reasons include: tumor was not very clear in the MRIs and the reader was not confident for the cases). For the remaining 275 cases, 3 had errors in image processing, 1 had implants and 1 was missing the post-contrast sequences. The remaining 270 cases were used for our study. 90 cases belonged to luminal A subtype while the rest belonged to the other 3 subtypes.

Imaging and Pathology Data

All MRIs in this study were acquired using a 1.5 Tesla (Signa HDx, GE Healthcare, Little Chalfont, United Kingdom[44]; Signa HDxt, GE Healthcare[5]; MAGNETOM Avanto, Siemens, Munich, Germany[37]) or 3.0 Tesla scanner (Signa HDx, GE Healthcare[167]; MAGNETOM Trio, Siemens, Munich, Germany[25]) scanner using a breast coil (Invivo, Orlando, FL). More details were illustrated in Table 1. Each case had the following sequences: nonfat-saturated T1-weighted, fat-saturated T2-weighted sequence, and pre-contrast followed by three dynamic post-contrast T1-weighted gradient echo series with fat suppression after intravenous administration of gadopentetate dimeglumine (Magnevist, Bayer Health Care, Berlin, Germany) or gadobenate dimeglumine (MultiHance, Bracco, Milan, Italy). Contrast load was determined using a weight based dosing protocol (0.2 mL/kg). The estrogen receptor (ER), progesterone receptor (PR), and HER2-neu status were obtained from the initial breast biopsy pathology report.

Image Annotation

Six fellowship-trained breast-imaging radiologists with 6–20 years of experience finished the annotation of the dataset. For each case up to 5 bounding boxes were annotated by the reader. The bounding boxes represented a lesion(s). Finally, the annotated bounding boxes were converted to binary masks where 1 indicates inside the bounding box and 0 indicates outside of the bounding box.

MR Imaging Pre-processing and Tumor Patch Extraction

The MRIs in our dataset contain differences in pixel spacing as they were captured by different devices. Thus, the first step was to register all the MRIs to the same spatial resolution. This was achieved by accounting for the frequency of different spatial resolutions, and selecting the resolution that occurred most often as the target resolution. Next, MRIs with other spatial resolutions were scaled to the target resolution using bilinear interpolation. Three channel images were constructed by concatenating $T^{'} - T^{p}$, $T^{''} - T^{p}, T^{'''} - T^{p}$, where $T^{'}$, $T^{''}$ and $T^{'''}$ are three post-contrast sequences and T^{p} is the pre-contrast sequence. We assumed that the center of the bounding box is the center of the lesion, and we sampled square patches around the lesion center. The patch size ranged between 80 pixels and 120 pixels in our experiments. Here patch size relates to the size of patches cropped from original image. The motivation of altering the patch size is that larger regions lead to global features of the lesion while smaller regions focus more on local features, and choosing different patch sizes can be regarded as balancing between global/local features of

the lesion. For each lesion, four additional patches were generated by random translation and rotation. For each patient there are about 33 lesion regions on average, result in 44660(33.1 lesion regions × (1 original patch + 4 augmented patches) × 270 patients) patches generated in total for a given patch size.

Methods Overview

Three different deep learning approaches were used in our experiment: training from scratch, transfer learning and off-the-shelf deep features approach. Pipelines of these approaches are illustrated Figure 1. We used CAFFE[23] deep learning framework on a desktop with a NVIDIA GTX 1080 GPU. In our experiment we trained the models for 40 epochs.

Training from Scratch

Training from scratch is the most straightforward way to train convolutional neural networks. In this approach, the network is initialized with random weights and the training is performed with the available data. There are two factors that greatly affect the performance: training dataset and network architecture. Current neural networks contain tens or even hundreds of layers, resulting in millions of free parameters to train. Building a large dataset requires large amount of time and labor, and in some situations such as in medical imaging the amount of available data is often insufficient. Since improving the quality of training dataset is challenging, more attention has been paid to the improvement of network structure. Recently many neural network structures[24,25] have been proposed which perform well on a variety of different tasks. We choose three representative network architectures in our study: GoogleNet[24], VGGNet[25] and CIFAR[26]. For GoogleNet we used both original GoogleNet and reduced GoogleNet while for VGGNet we just used the reduced VGGNet, resulting in four specific architectures in total. In our problem the lesions on MRI have different sizes which requires the network to have multi-scale capability. We chose GoogleNet for this reason. The main building block of GoogleNet is the Inception module that consists of three convolution layers with different kernel sizes. These kernels can be used to capture the features of patterns of different sizes. VGGNet is much larger than GoogleNet, which means it has more weights and needs more data to train. Thus we only used reduced VGGNet due to the data insufficiency. For the original GoogleNet a much smaller learning rate (0.0001) than default (0.01) was chosen through experiment(details are in Results section), as image patches in our dataset had far less variation compared with those natural images. For convenience, we refer the reduced GoogleNet as GoogleNet_R and reduced VGGNet as VGG_R. To reduce the GoogleNet, in the first convolution layer we decreased the number of filters from 64 to 32, the pad size from 3 to 2, the kernel size from 7 to 5, the stride from 2 to 1 and the stride of the subsequent pooling layer from 2 to 1. The input size of the network was fixed to 64×64×3. To reduce VGGNet, we only retained the first 2 convolution layers and all the fully connected layers, and input size of the neural network was fixed to 80×80×3. The output dimension of the last layer of these networks were modified from 1000 to 2, to adapt to the binary classification setting. For CIFAR network, we set the input size to 80×80×3 and the output dimension of the last layer to 2. In training phase, the L2 regularization and dropout strategy were used to reduce overfitting. Details of these 4 networks were illustrated in Table 2.

Transfer Learning

Transfer learning is a two-step approach. The first step is to pre-train the network using a large dataset for a different task than the one at hand. The second step is fine-tuning the pre-trained network on the dataset representing the problem of interest (a.k.a., the target dataset). Usually the target dataset is much smaller than the dataset used for pre-training. In our scenario we chose natural image dataset for pre-training and fine-tuned the pre-trained network on our own dataset. The networks pre-trained on large natural image datasets are already publicly available. We chose GoogleNet and VGGNet pre-trained on ImageNet for transfer learning. Following modifications were made to adapt the networks to our specific problem: the output dimension of the last fully connected layer was changed from 1000 to 2, and the weights in this layer were re-initialized randomly. The first modification is intuitive as we are dealing with a binary classification problem. The second modification is based on the observation[27] that last fully connected layer is the most problem-specific layer, so it needs to be completely retrained. The learning rate of the last fully connected layer was set to 0.001, which is 10 times larger than the learning rate of all the other layers (0.0001). When pre-trained on ImageNet, the input resolution of GoogleNet and VGGNet were fixed to 224×224×3. To be consistent with this resolution, all the patches were resized to 256×256×3. Then a 224×224×3 sub-patch was randomly cropped from the 256×256×3 image in each epoch.

Off-the-shelf deep features

While training from scratch and transfer learning are both end-to-end approaches, off-the-shelf deep features approach explicitly separate feature extraction and classification. Since this approach only uses pre-trained networks as feature extractors and trains traditional classifiers using the extracted features, it can avoid the overfitting while taking advantage of the richness of deep features. Another benefit of using deep features approach is that the extracted features can be used for many different tasks. For example pre-trained network was originally used for image classification, but when used as feature extractor the extracted features were used for scene recognition[17]. We chose GoogleNet and VGGNet pre-trained on ImageNet as feature extractors and trained SVMs using the extracted features. While deep neural networks have multiple layers, it remains a question which layer's output should be used as features. The shallow layers may output low level but less specific features. Feature maps of deeper layers are higher level features but are also more problem-specific. We followed a previously described approach[17] and used the feature map of the last fully connected layer as default. Other layers' feature maps were also evaluated for comparison. We max pooled the feature map on the $x - y$ coordinates of the image plane, in order to limit the length of the feature vector. For our study there is a limited number of training samples, and SVM would not perform well when its feature length becomes larger than its training samples. SVMs with different kernel functions were then trained and compared.

Model Evaluation

Our patch-based approach treated each patch as a training sample in training phase. In test phase each patient was regarded as a test case. Five consecutive slices that contained the largest size of lesions were chosen and for each slice five patches (four corners and middle

region of the lesion) were sampled. The averaged score of the 25 patches was used as the final score.

We separated our dataset into 10 folds where each fold had approximately the same ratio between the number of positive samples and the number of negative samples. The performance of each fold was evaluated separately and the final AUC was obtained by averaging the AUC of the 10 folds. We ran 10 fold cross-validation 5 times and pick the medium result. The bootstrapping strategy was used to calculate the confidence interval.

Results

We first made several experiments to choose a suitable learning rate. Learning rate value was picked in [0.00001-0.001], and the training-test AUC curve of 3 experiments using learning rate 0.0001,0.0005 and 0.00002 were illustrated in Figure 2, row 1. AUCs for those 3 learning rates were 0.54,0.53 and 0.54. We chose 0.0001 as the initial learning rate for all the 4 models in the following experiments. The results for the models trained from scratch are illustrated in Table 3. Regarding the performance on training set (third column in Table 3), the AUC of GoogleNet, reduced GoogleNet and reduced VGGNet were all above 0.90, indicating that those trained networks fitted the training data distribution well. Note that even the reduced version of GoogleNet and VGGNet were still large networks. For CIFAR which is a rather small network, the AUC on training set was much lower than those large networks, but it had the highest AUC on test set. This indicates that larger networks are more likely to overfit when trained on a small dataset. The 95% confidence intervals of the test AUCs of the 4 networks were [0.51,0.60], [0.51,0.62], [0.46,0.60], and [0.52,0.63] respectively. AUCs of reduced VGGNet of the training and test set during training phase are plotted in Figure 2a. While the performance of reduced VGGNet on the training set reached high level quickly, on test set it stayed around the 0.58 for the remainder of the training.

We did the same experiments to choose the suitable learning rate(fully-connected layer) for transfer learning. Results using learning rate 0.001, 0.005 and 0.0002 were illustrated in Figure 2, row 2. AUCs for those 3 learning rates were 0.60, 0.60 and 0.58. We chose 0.001 as the initial learning rate of the fully-connect layer for GoogleNet and VGGNet. The results of transfer learning are shown in Table 4. Regarding the test AUC, transfer learning always performed better than training from scratch. We also plot the AUCs of the training and test set of GoogleNet during training phase in Figure 2(b). Compared with the curves in Figure 2(a), the training AUC curve in Figure 2(b) reaches its asymptote more slowly than 2(a) while the test AUC curve in Figure 2(b) stays higher. This indicates that transfer learning can better avoid overfitting than training from scratch. It also indicates that pre-trained neural networks on large dataset can adapted to small dataset even when there exist large style differences between the two datasets. The 95% confidence intervals of the test AUCs of the 2 networks were [0.55,0.66] and [0.52,0.65].

For off-the-shelf deep features approach both patch size in the feature extraction and kernel function in the SVM affect the final performance. We tested several combinations and the results are shown in Table 5. We found that using GoogleNet with 120 patch size and polynomial kernel SVM achieved best performance. Results using other kernel functions(rbf

kernel and linear kernel) in SVM, other patch size(80) and other feature extraction network(VGGNet) are also given. These patches were finally resized to fit the input size of the pre-trained network. From the third, fourth and fifth row we can infer that polynomial kernel best suits our problem. Two results of VGGNet are given in sixth and seventh row. Compared with the corresponding results using GoogleNet but the same patch size and kernel function (row 2 and row 4), GoogleNet performed better than VGGNet on our dataset. The 95% confidence intervals of the test AUCs of the 6 results were [0.51,0.64], [0.51,0.65], [0.57,0.71], [0.50,0.60], [0.50,0.59] and [0.48,0.62] respectively.

We also tested the performance of features extracted from different layers, including intermediate convolution layers and fully connected layers. The performance of each layer's feature map is shown in Figure 3 and Table 6. Regarding the test AUC, feature map of the last fully connected layer has the best performance.

Discussion

We studied the molecular subtype classification of breast cancer using deep learning applied to dynamic contrast-enhanced magnetic resonance imaging. Three different deep learning approaches were investigated, and among them off-the-shelf deep features approach performed best. The highest AUC obtained was 0.65 using pre-trained GoogleNet as feature extractor and a polynomial kernel SVM trained on the extracted features. The highest AUC obtained for training from scratch was 0.58, which is only slightly higher than random guess. The performance of transfer learning is worse than off-the-shelf deep features but better than training from scratch. This implies that when the number of training samples is limited, overfitting is the main problem.

There are two potential ways to solve the overfitting problem. Instead of using pre-training networks on a natural image dataset, using pre-training networks on large medical image dataset seems more promising, as medical images share much more in common. Another way to deal with data insufficiency is to train a model that can do multiple tasks[28]. Tasks that have insufficient training data will benefit from tasks that have sufficient training data. We consider these potential solutions as future work.

The overall prognostic power of our results was low to intermediate with the highest AUC of 0.65. While direct comparison to results of other studies that used hand-crafted features is difficult since the specific goals of different studies, the datasets, and the evaluation metrics vary widely, the obtained level of performance is in the expected range[29]. A direct comparison of deep learning with hand crafted features will be a part of future work.

Our study had some limitations. The most significant one, discussed throughout the paper, is the relatively low number of cases in the context of deep learning. Our study explores this issue and suggests the deep learning solutions that are most likely to succeed in this imperfect situation. It is encouraging to see that even with such a small number of cases we were able to develop a deep learning model that showed some predictive value of molecular subtypes. Another limitation was considering only the distinction between the Luminal A subtype versus all other subtypes, as opposed to considering each molecular subtypes

individually. This was due to the small number of non-luminal A cases. Luminal A cases are typically less aggressive than some other subtypes such as triple negative cancers and can undergo different types of treatment.

In conclusion, in this study, we were able to demonstrate that deep learning can aid the investigation of the relationship between cancer imaging and tumor imaging in breast cancer. Future studies will include repeating this investigation with a larger dataset and comparison of deep learning to traditional machine learning models based on hand-crafted features.

Acknowledgements

This work was funded by the NIH (1R01EB021360) and North Carolina Biotechnology Center (2016-BIG-6520).

References

1. Uematsu T, Kasami M, Yuen S. Triple-Negative Breast Cancer: Correlation between MR Imaging and Pathologic Findings1. Radiology. 2009;250(3):638–647. doi:10.1148/radiol.2503081054. [PubMed: 19244039]
2. Costantini M, Belli P, Distefano D, et al. Magnetic resonance imaging features in triple-negative breast cancer: Comparison with luminal and HER2-overexpressing tumors. Clin Breast Cancer. 2012;12(5):331–339. doi:10.1016/j.clbc.2012.07.002. [PubMed: 23040001]
3. Yamamoto S, Maki DD, Korn RL, Kuo MD. Radiogenomic analysis of breast cancer using MRI: A preliminary study to define the landscape. Am J Roentgenol. 2012;199(3):654–663. doi:10.2214/AJR.11.7824. [PubMed: 22915408]
4. Sung JS, Jochelson MS, Brennan S, et al. MR imaging features of triple-negative breast cancers. Breast J. 2013;19(6):643–649. doi:10.1111/tbj.12182. [PubMed: 24015869]
5. Agner SC, Rosen MA, Englander S, et al. Computerized Image Analysis for Identifying Triple-Negative Breast Cancers and Differentiating Them from Other Molecular Subtypes of Breast Cancer on Dynamic Contrast-enhanced MR Images: A Feasibility Study. Radiology. 2014;272(1):91–99. doi:10.1148/radiol.14121031. [PubMed: 24620909]
6. Mazurowski MA, Zhang J, Grimm LJ, Yoon SC, Silber JI. Radiogenomic Analysis of Breast Cancer: Luminal B Molecular Subtype Is Associated with Enhancement Dynamics at MR Imaging. Radiology. 2014;273(2):365–372. doi:10.1148/radiol.14132641. [PubMed: 25028781]
7. Blaschke E, Abe H. MRI phenotype of breast cancer: Kinetic assessment for molecular subtypes. J Magn Reson Imaging. 2015;42(4):920–924. doi:10.1002/jmri.24884. [PubMed: 25758675]
8. Grimm LJ, Zhang J, Mazurowski MIA. Computational approach to radiogenomics of breast cancer: Luminal A and luminal B molecular subtypes are associated with imaging features on routine breast MRI extracted using computer vision algorithms. J Magn Reson Imaging. 2015;42(4):902–907. doi:10.1002/jmri.24879. [PubMed: 25777181]
9. Guo W, Li H, Zhu Y, et al. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. J Med Imaging. 2015;2(4):41007. doi:10.1117/1.JMI.2.4.041007.
10. Wang J, Kato F, Oyama-Manabe N, et al. Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: A pilot radiomics study. PLoS One. 2015;10(11):1–17. doi:10.1371/journal.pone.0143308.
11. Saha A, Grimm LJ, Harowicz M, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. Med Phys. 2016;43(8):4558–4564. doi:10.1118/1.4955435. [PubMed: 27487872]
12. Saha A, Yu X, Sahoo D, Mazurowski MA. Effects of MRI scanner parameters on breast cancer radiomics. Expert Syst Appl. 2017;87:384–391. doi:10.1016/j.eswa.2017.06.029. [PubMed: 30319179]

13. Yamaguchi K, Abe H, Newstead GM, et al. Intratumoral heterogeneity of the distribution of kinetic parameters in breast cancer: comparison based on the molecular subtypes of invasive breast cancer. Breast Cancer. 2015;22(5):496–502. doi:10.1007/s12282-013-0512-0. [PubMed: 24402638]
14. Li H, Zhu Y, Burnside ES, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. npj Breast Cancer. 2016;2(1):16012. doi:10.1038/npjbcancer.2016.12. [PubMed: 27853751]
15. Fan M, Li H, Wang S, Zheng B, Zhang J, Li L. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer. PLoS One. 2017;12(2):1–15. doi:10.1371/journal.pone.0171683.
16. Wu J, Sun X, Wang J, et al. Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: Model discovery and external validation. J Magn Reson Imaging. 2017:1017–1027. doi:10.1002/jmri.25661. [PubMed: 28177554]
17. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning Deep Features for Scene Recognition using Places Database. Proc 27th Int Conf Neural Inf Process Syst 2014;1:487–495. doi:10.1162/153244303322533223.
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Adv Neural Inf Process Syst. 2012:1–9. doi:10.1145/3065386.
19. Zhu Z, Liang D, Zhang S, Huang X, Li B, Hu S. Traffic-Sign Detection and Classification in the Wild. 2016 IEEE Conf Comput Vis Pattern Recognit 2016:2110–2118. doi:10.1109/CVPR.2016.232.
20. Shin HC, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging. 2016;35(5):1285–1298. doi:10.1109/TMI.2016.2528162. [PubMed: 26886976]
21. Chen H, Qi X, Yu L, Heng P-A. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. 2016:2487–2496. doi:10.1109/CVPR.2016.273.
22. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. IEEE Trans Med Imaging. 2017;36(4):994–1004. doi:10.1109/TMI.2016.2642839. [PubMed: 28026754]
23. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. Proc ACM Int Conf Multimed - MM '14. 2014:675–678. doi:10.1145/2647868.2654889.
24. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2015;07-12-June:1–9. doi:10.1109/CVPR.2015.7298594.
25. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014:1–14. doi:10.1016/j.infsof.2008.09.005.
26. Krizhevsky A Learning Multiple Layers of Features from Tiny Images…. Sci Dep Univ Toronto, Tech…. 2009:1–60. doi:10.1.1.222.9220.
27. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013. Comput Vision–ECCV 2014. 2014;8689:818–833. doi:10.1007/978-3-319-10590-1_53.
28. Kaiser L, Gomez AN, Shazeer N, et al. One Model To Learn Them All. 2017. doi:10.1007/s11263-015-0816-y.
29. Saha A, Harowicz MR, Grimm LJ, et al. A machine learning approach to radiogenomics of breast cancer : a study of 922 subjects and 529 DCE-MRI features. Br J Cancer. 2018;(1). doi:10.1038/s41416-018-0185-8.

1. Deep learning can aid the investigation of the relationship between cancer imaging and tumor imaging in breast cancer.
2. The performance of transfer learning is worse than off-the-shelf deep features but better than training from scratch.
3. Deep learning may play a role in discovering radiogenomic associations in breast cancer.

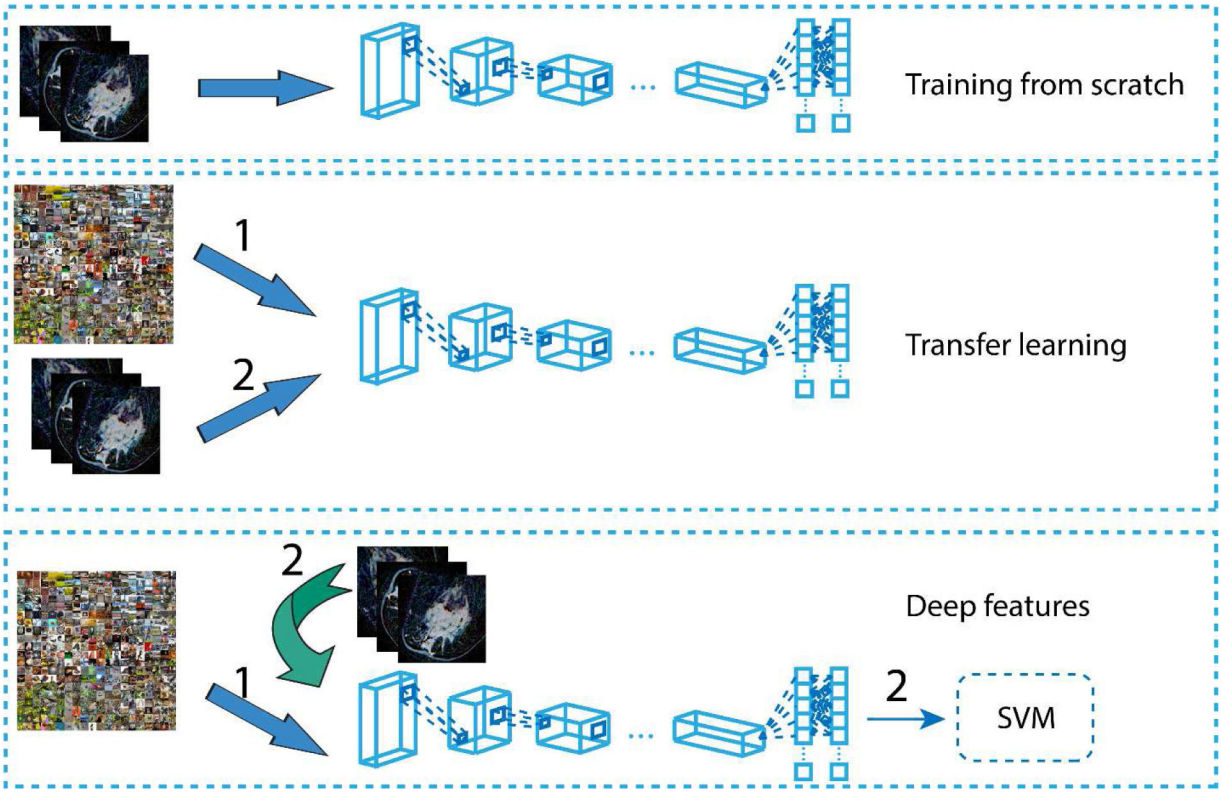

Figure 1.

Pipeline of the three approaches. Training from scratch: directly training a neural network using our dataset. Transfer learning: pre-training a model on natural image dataset, then fine-tuning it using our dataset. Off-the-shelf deep features: using pre-trained neural network as feature extractor, and then training a traditional classifier(SVM) using the extracted features.

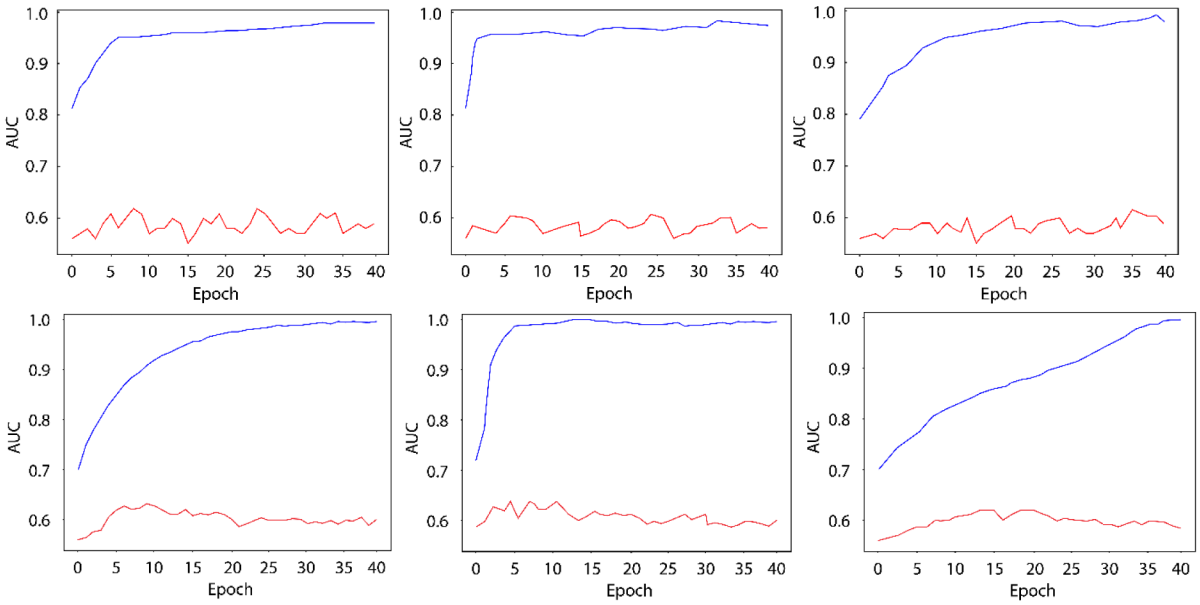

Figure 2.

Row 1, from left to right: AUCs of training from scratch for reduced VGGNet, using learning rate of 0.0001,0.0005 and 0.00002. Row 2, from left to right: AUCs of transfer learning using pre-trained GoogleNet, using learning rate 0.001,0.005 and 0.0002 for the last fully connected layer. Red line: test AUC; blue line: training AUC.

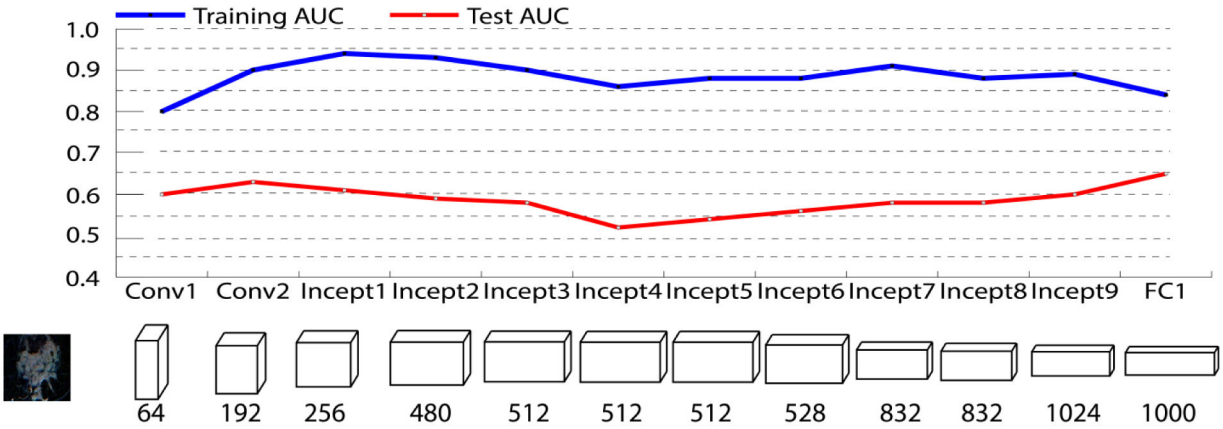

Figure 3.
AUCs of different layer's feature map of pre-trained GoogleNet.

Table 1.

Breast MRI Protocols by Scanner.

Scanner	Field strength(T)	TR	TE	FOV(cm)	Matrix size	Case number
Signa HDx/HDxt	1.5	5.3	2.4	38	350×350	44
MAGNETOM Avanto	1.5	4.0	1.3	36	448×448	36
Signa HDx	3.0	5.7	2.4	34	350×350	165
MAGNETOM Trio	3.0	4.1	1.4	36	448×448	25

Table 2.

Details about the 4 networks used in training from scratch. L2 regularization was used in convolutional layers while 0.5 dropout was used in fully-connected layers. The initial learning rate was set to 0.0001 and “step” learning rate policy was used(the learning rate was reduced 50% after each 5 epochs). MACC indicates the number of multiply-accumulate operations and Param indicates the number of parameters in the model.

	Regularization	Learning rate	Learning rate policy	MACC	Param
GoogleNet	L2,dropout	0.0001	Reduce 50% after each 5 epochs	1.59G	5.98M
GoogleNet_R	L2,dropout	0.0001	Reduce 50% after each 5 epochs	388.77M	5.98M
VGGNet_R	L2,dropout	0.0001	Reduce 50% after each 5 epochs	365.16M	317.38K
CIFAR	L2,dropout	0.0001	Reduce 50% after each 5 epochs	77M	92.13K

Table 3.

Results of training from scratch. We show the results of four different networks: original GoogleNet, reduced GoogleNet(GoogleNet_R), reduced VGGNet(VGG_R) and CIFAR. We list network input, training AUC and test AUC.

	Network Input	Training AUC	Test AUC
GoogleNet	224×224×3	0.93	0.56
GoogleNet_R	64×64×3	0.94	0.57
VGGNet_R	80×80×3	0.98	0.54
CIFAR	80×80×3	0.75	0.58

Table 4.

Results of transfer learning. We choose ImageNet pre-trained GoogleNet and VGGNet and perform transfer learning on our own dataset.

	Network Input	Training AUC	Test AUC
GoogleNet	224×224×3	0.93	0.60
VGGNet	224×224×3	0.89	0.59

Table 5.

AUCs with different patch sizes and kernels of SVM.

Network	Patch Size	Kernel Function	Training AUC	Test AUC
GoogleNet	80	poly	0.82	0.58
GoogleNet	120	rbf	0.85	0.59
GoogleNet	120	poly	0.84	0.65
GoogleNet	120	linear	0.74	0.56
VGGNet	80	poly	0.86	0.55
VGGNet	120	poly	0.83	0.56

Table 6.

Performance of features extracted from different layers.

	Conv1	**Conv1**	**Incep1**	**Incp2**	**Incp3**	**Incp4**	**Incp5**	**Incp6**	**Incp7**	**Incp8**	**Incp9**	**FC1**
Feature Length	64	192	256	480	512	512	512	528	832	832	1024	1000
Training AUC	0.80	0.90	0.94	0.93	0.90	0.86	0.88	0.88	0.91	0.88	0.89	0.84
Test AUC	0.60	0.63	0.61	0.59	0.58	0.52	0.54	0.56	0.58	0.58	0.60	0.65