Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Digital mammographic tumor classification using transfer learning from deep convolutional neural networks

Benjamin Q. Huynh
Hui Li
Maryellen L. Giger

SPIE.

Benjamin Q. Huynh, Hui Li, Maryellen L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag.* **3**(3), 034501 (2016), doi: 10.1117/1.JMI.3.3.034501.

Digital mammographic tumor classification using transfer learning from deep convolutional neural networks

Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger*
University of Chicago, Department of Radiology, 5841 South Maryland Avenue, Chicago, Illinois 60637, United States

Abstract. Convolutional neural networks (CNNs) show potential for computer-aided diagnosis (CADx) by learning features directly from the image data instead of using analytically extracted features. However, CNNs are difficult to train from scratch for medical images due to small sample sizes and variations in tumor presentations. Instead, transfer learning can be used to extract tumor information from medical images via CNNs originally pretrained for nonmedical tasks, alleviating the need for large datasets. Our database includes 219 breast lesions (607 full-field digital mammographic images). We compared support vector machine classifiers based on the CNN-extracted image features and our prior computer-extracted tumor features in the task of distinguishing between benign and malignant breast lesions. Five-fold cross validation (by lesion) was conducted with the area under the receiver operating characteristic (ROC) curve as the performance metric. Results show that classifiers based on CNN-extracted features (with transfer learning) perform comparably to those using analytically extracted features [area under the ROC curve $(AUC) = 0.81$]. Further, the performance of ensemble classifiers based on both types was significantly better than that of either classifier type alone ($AUC = 0.86$ versus 0.81, $p = 0.022$). We conclude that transfer learning can improve current CADx methods while also providing standalone classifiers without large datasets, facilitating machine-learning methods in radiomics and precision medicine. [DOI: 10.1117/1.JMI.3.3.034501]

Keywords: transfer learning; mammography; deep learning; convolutional neural networks; computer-aided diagnosis; radiomics; precision medicine.

Paper 16029RR received Feb. 11, 2016; accepted for publication Aug. 1, 2016; published online Aug. 22, 2016.

1 Introduction

Computer-aided diagnosis (CADx) systems have been successfully used to support human decision-making in radiological image analysis and precision medicine in general.[1–3] In particular, many research efforts have been made to use CADx in the task of diagnosing breast cancer by using classification algorithms to determine whether a lesion is malignant or benign based on features extracted from the image data.

Traditional approaches to breast cancer CADx involve analytically extracting clinically specified tumor features (e.g., shape and density) and estimating malignancy probabilities based on those analytically extracted hand-crafted lesion features.[1–3] Alternative approaches involve learning features directly from the full images through methods such as convolutional neural networks (CNNs), in the hopes that learning features directly will yield unintuitive, hidden features that contain more information than analytically extracted features.[4–6]

Indeed, advances in recent years in deep learning and computer vision have been remarkable, with CNNs seeing great success in many benchmark image classification tasks.[7–9] However, CNNs are contingent on very large and properly labeled datasets, as well as substantial computational resources. As a result, training CNNs from scratch is often infeasible for CADx and medical image data.

Surprisingly, it has been shown that generic features can be transferred from pretrained CNNs to create powerful classifiers for a new target task different from the original task of the CNN—a process known as transfer learning.[10–13] In particular, success has been found in transferring knowledge from general object recognition tasks to classification tasks in which categories are visually similar. Examples include categorizing species of dogs or types of indoor scenes using CNNs trained on classifying everyday objects.[12,13] Since CADx includes similar subtle classification tasks, we hypothesized that structures within a CNN trained on everyday objects could be used to create a classifier for breast cancer CADx, thereby harnessing the predictive power of deep neural networks without the computational costs or large dataset requirements.

In this study, we present a breast imaging CADx system based on deep neural networks with transfer learning. We tested the optimal point at which to extract features from the pretrained CNN, identifying the specific utility of transfer learning in CADx. Further, we evaluated three different classifiers: one trained on our previously developed analytically extracted hand-crafted CADx features, one trained on pretrained CNN-extracted features, and an ensemble classifier trained on both types of features.

2 Materials and Methods

2.1 *Overview*

An overview of the classification methodology—in the task of distinguishing malignant and benign lesions in digital

*Address all correspondence to: Maryellen L. Giger, E-mail: m-giger@uchicago.edu

2329-4302/2016/$25.00 © 2016 SPIE

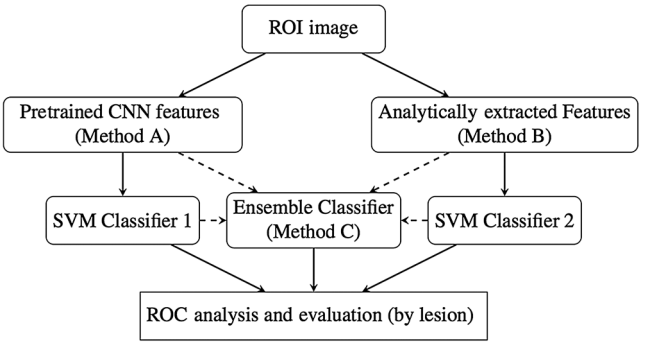

Fig. 1 An overview of the various methods used in the task of distinguishing between benign and malignant tumors.

mammograms—is illustrated in Fig. 1. First, features were obtained from the images by two different methods: extracting them automatically from pixel data via a pretrained CNN (Method A) and extracting them via segmented-tumor-based analytical methods (Method B). Support vector machine (SVM) classifiers were then trained on each of these feature sets, as well as an ensemble classifier averaged from both individual classifiers (Method C). From there, receiver operating characteristic (ROC) analysis and cross validation (by lesion) were used to evaluate and compare models.

2.2 *Imaging Data*

The data were retrospectively obtained under an Institutional Review Board-approved protocol from the University of Chicago Medical Center. The dataset consisted of 219 lesions on full-field digital mammography images from which within each image, a region of interest (ROI) about each lesion had been extracted, yielding 607 ROIs.[14] To extract the ROI from the full mammographic image, an expert mammographer marked the center of each lesion, and a 512×512 box was cropped around the center, with a pixel size of 0.1 mm. Note that initially, there were 287 lesions with 739 ROIs, but 132 ROIs were removed from the dataset prior to analysis due to visual artifacts obscuring the image data, including paddles within magnification views or location markers. The number of ROIs per lesion varied between 1 and 13, with most lesions having two to three ROI images. The ROIs had been labeled as either benign or malignant, based on pathologic reports. For our study, 261 of the ROIs were labeled as benign and 346 were labeled as malignant. The benign lesions ranged in diameter from 0.32 to 3.17 cm, with a mean diameter of 1.09 cm and a standard deviation of 0.52 cm. The malignant lesions ranged in diameter from 0.28 to 3.21 cm, with a mean diameter of 1.40 cm and a standard deviation of 0.56 cm.

2.3 *Method A: Pretrained Convolutional Neural Network Features*

AlexNet is a CNN model with three fully connected layers and five convolutional layers, three of which are followed by max-pooling layers.[7] A publicly available version of AlexNet is pretrained on the ImageNet[15] dataset, which consists of over 1 M images and 1 K possible classes. By using a pretrained version, the weights of the model are preinitialized, as opposed to being randomly initialized when training from scratch. For further explanation of AlexNet's architecture, see the definitive work by Krizhevsky et al.[7]

AlexNet was used to extract features directly from the ROIs, without the need for lesion segmentation. Since the outputs of each convolutional, pooling, and fully connected layer can be extracted as features, there are 11 layers from which features can be extracted and used as inputs for classification. Given the sparsity of the extracted features, all zero-variance columns were eliminated prior to input to classification. The illustration of extraction process is shown in Fig. 2.

Although it has been shown[11] that features from earlier layers in the architecture of a neural network are more generalizable to different tasks and that features from later layers tend to be more specific to their original task, it was unclear which layer of AlexNet in particular would be best suited for classification of breast tumor images. Thus, classifiers based on each layer were trained to determine the optimal layer. Features from each layer were extracted and used to train SVM classifiers, which were then evaluated via ROC analysis and five-fold cross validation, as detailed in Sec. 2.6. The best classifier was then selected based on predictive performance and computational costs, as discussed in Sec. 3.1.

2.4 *Method B: Analytically Extracted Computer-Aided Diagnosis Features*

To classify the breast lesion images based on analytically extracted features, the lesion was first segmented from the surrounding parenchymal background within the 512×512 ROI. The center of the lesion was manually indicated, then an automatic lesion segmentation was performed, based on a multiple-transition-point, gray-level, region-growing technique. Further details of the automatic lesion segmentation process can be found in the work of Li et al.[14] After the lesion was segmented from the rest of the image, image features (i.e., mathematical descriptors) were extracted from the lesion, including lesion size, shape, intensity (e.g., average gray level, contrast, and texture), and margin (e.g., spiculation and sharpness) of the mass. Extensive descriptions of the lesion features and how they were extracted can be found in various papers from our laboratory.[14,16,17] Finally, an SVM classifier was trained on the extracted features and evaluated as described in Sec. 2.6.

2.5 *Method C: Ensemble Classifier*

After individually performing classification with CNN features (Method A) and analytically extracted features (Method B), a simple ensemble technique known as soft voting[18] was used to combine the outputs from both individual classifiers. Through this technique, the output probabilities from the individual classifiers were averaged and then used as the final predicted probabilities. To be more precise, the soft-voting output was computed as $p_3 = [(p_1 + p_2)/2]$, where p_1 and p_2 are the output probability vectors from the individual classifiers.

2.6 *Classification and Evaluation Methods*

For all evaluation tasks, unless otherwise stated, the following methods were used. First, preprocessing was conducted by centering and scaling each feature to have zero mean and unit variance. For classification using extracted features, an SVM[19] with a polynomial kernel was employed. The kernel is defined as $K(x, y) = (ax^T y + c)^d$,[20] and optimized values of a, c, and d, were selected by cross validated grid-search with five folds. In the task of distinguishing between benign and malignant lesions,

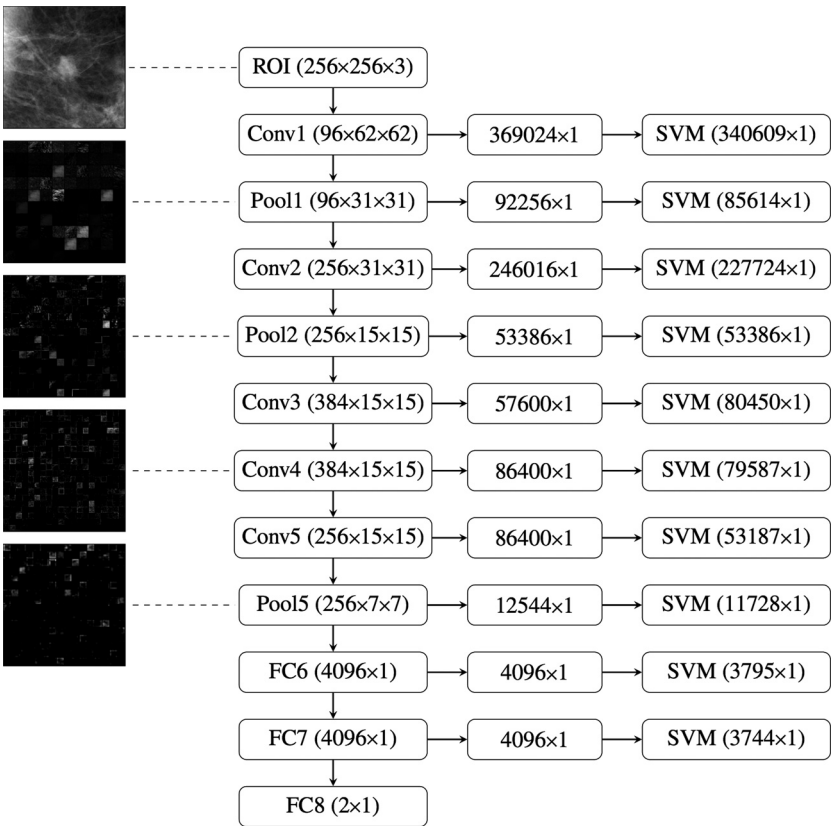

Fig. 2 A schematic of how features are extracted using a pretrained AlexNet. It should be noted that the ROI has three channels despite being grayscale to fit the original architecture that was designed for color images. Additionally, the ROIs are down-sampled to 256×256 from 512×512 to fit the architecture of the network. Each ROI is sent through the network and the outputs from each layer are preprocessed to be used as sets of features for an SVM. The filtered image outputs from some of the layers can be seen in the left column. The numbers in parentheses for the center column denote the dimensionality of the outputs from each layer. The numbers in parentheses for the right column denote the length of the feature vector per ROI used as an input for the SVM after zero-variance removal. For example, Pool1 outputs 96 31×31 images, which are all combined and flattened into a 92,256-length feature vector, which is then reduced to an 85,614-length vector to be used in an SVM. After a feature vector has been extracted from each ROI, the SVM is then trained and evaluated by cross validation.

classifiers were evaluated with ROC analysis,[21,22] with the area under the ROC curve (AUC) as the performance metric. AUC was calculated by five-fold cross validation by lesion. Cross validation was conducted by lesion instead of by ROI since there were often multiple images, i.e., ROIs, of the same lesion. This method of cross validation is preferred in CADx and other image interpretation analyses due to correlations among ROIs of the same lesion, which would cause erroneous inflation of AUC values if one were to simultaneously train and test ROIs from the same lesion. Performance comparison t-tests as described by Hothorn et al.[23] were used to test for statistical significance between classifier AUC scores, with Bonferroni corrections for multiple comparisons.

3 Results

3.1 *Layer Comparisons*

Figure 3 shows the AUC performance of the SVM classifiers trained on CNN features extracted from each layer of the pretrained AlexNet. As shown, the performance increases with the number of layers until it peaks at the Conv4 layer, after which performance starts to drop slightly. There are sharp drops in performance after the Fc6 layer and the Fc7 layer.

We chose the Fc6 layer as the optimal layer due to its high predictive performance and relatively low dimensionality. The convolutional and pooling layers all had feature vector lengths one to two orders of magnitude higher than the fully connected layers, vastly increasing computational costs. Although the classifier based on features from Conv4 had the highest AUC, the difference between Conv4 and Fc6 was small (AUC $= 0.83$ versus 0.81).

3.2 *Model Evaluation and Comparison*

The ensemble classifier was shown to be significantly better than the classifier based on analytically extracted features after Bonferroni correction (AUC $= 0.86$ versus 0.81, $p = 0.022$). Table 1 and Fig. 4 show the AUC performance of each of the three classifiers. SVM1 is the SVM classifier based on the features extracted from the pretrained CNN. SVM2 is the SVM classifier based on the analytically extracted features. The ensemble classifier is the soft-voting classifier based on SVM1 and SVM2.

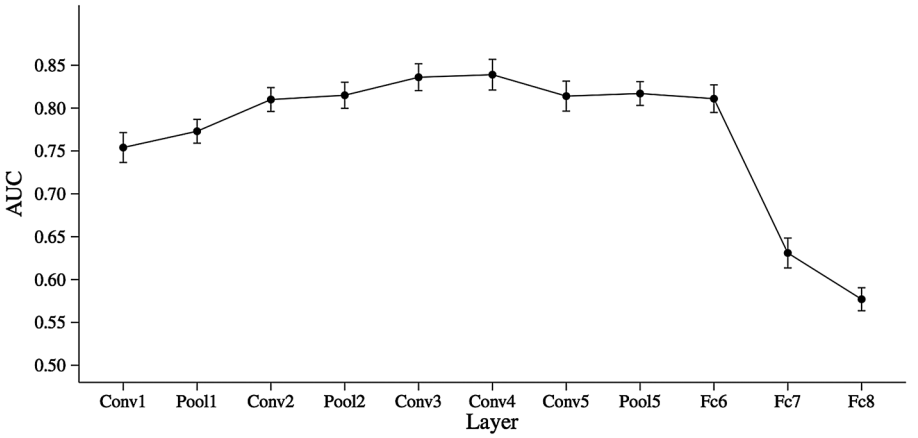

Fig. 3 Performance in terms of AUC for classifiers based on features from each layer of AlexNet in the task of distinguishing between malignant and benign tumors. The error bars represent the standard error of each AUC value, calculated by dividing the standard deviation by the square root of the number of cross validation folds.

Table 1 Model performance and evaluation.

Model	AUC	AUC.SD	Time
1 Pretrained CNN features (SVM1)	0.81	0.04	~7 min
2 Analytically extracted features (SVM2)	0.81	0.03	~5 min
3 Ensemble classifier (SVM1 and SVM2)	0.86	0.01	~10 min

Note: AUC is the area under the ROC curve, averaged for each fold of cross validation. AUC.SD is the standard deviation of the AUC, also averaged for each fold of cross validation. Time denotes the estimated computational time to extract features and train a classifier with five-fold cross validation over all 607 lesions.

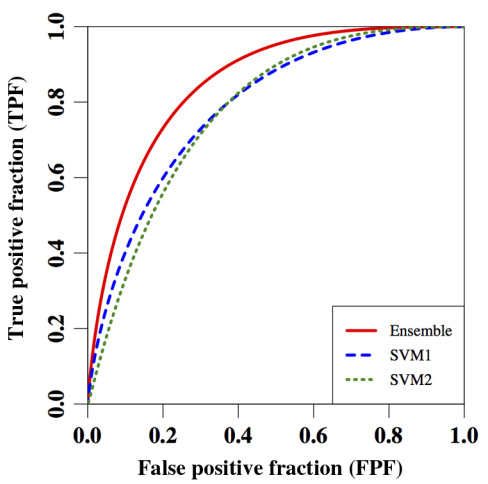

Fig. 4 Fitted binormal ROC curves depicting the performances of each classifier in the task of distinguishing between malignant and benign ROIs. Each curve corresponds to its respective classifier listed in Table 1.

4 Discussion

We have shown that transfer learning from nonmedical tasks can be used to significantly improve current methods of classification for CADx while also individually yielding competitive predictive performances. CNNs specifically trained for a given task are generally expected to outperform transfer learning methods, but only given sufficient data. As discussed in the introduction, modern advances in computer vision are often dependent on large, annotated datasets, which are not readily available for medical images. Thus, small-sample-size methods like the ones used in this study may continue to be vital in the future for CADx.

It should be noted that many of our methodological choices were not necessarily optimal in terms of predictive performance. We chose AlexNet as a pretrained CNN due to its simplicity and surrounding literature, but there are newer, more advanced networks, like Simonyan and Zisserman's "very deep" network,[8] that are known to be better for transfer learning.[24] Further, our ensembling method of soft voting was also chosen for simplicity. Some combination of advanced ensembling and feature selection methods could possibly yield higher predictive performance results. Finally, we chose to extract features from the Fc6 layer instead of the Conv4 layer despite their performance differences due to concerns of overfitting and computational costs. A 86,400-length feature vector seemed too unwieldy given a dataset of only 607 images, but Conv4 may be the better choice given a larger dataset.

It is also important to note that using a pretrained CNN restricts us to using its original architecture. Consequently, we had to fit our data to an architecture that is likely suboptimal for medical image data. While we could modify the input images and choose the layers for output, our ability to change the architecture of AlexNet was limited.

To our knowledge, this is the first study to use pretrained CNNs as fixed feature extractors in comparison and in combination with CADx-specific features for the task of diagnosing medical images. Bar et al.[25] and Anavi et al.[26] demonstrated the effectiveness of CNN-extracted features in characterizing chest x-rays, but did not provide comparison with traditional CADx methods as we report here. More specifically, Bar et al. compared CNN-extracted features with general image descriptors such as GIST instead of CADx-specific features. Wang et al.[27] used an ensembling method similar to ours to combine analytically extracted features and CNN features for the task of mitosis detection, but they did not use transfer learning and thus had to train a full CNN from scratch. Carneiro et al.[28]

implemented a similar transfer learning technique—fine-tuning, wherein a CNN is retrained on top of its original initialized weights—also to estimate malignancy probabilities in mammogram images. However, since it entails retraining a full network, fine-tuning thus requires the computational resources and training time that our method avoids. Further, we empirically found that a fine-tuned CNN performed only slightly better than an uninitialized one and worse than our methods. Yosinski et al.[11] suggest that fine-tuning should be used when the new dataset is large and similar to the original task and fixed feature extraction should be used when the new dataset is small and different from the original task, so our methods seem more appropriate given the task and size of our breast image datasets.

In this preliminary study, we demonstrated the potential usefulness of transfer learning for the task of CADx. Predictive performance is crucial to CADx and we believe that the performance gains from our transfer learning techniques could improve current applications of CADx in a variety of contexts. Comparing CNN-extracted features with human-designed features may allow for better interpretation and understanding of CADx output by clarifying the relationship between CNN-extracted features and calculated physical descriptors of lesions. We see these comparisons as important since CNN-extracted features are not intuitive or easily interpretable on their own. Other aspects, such as identifying classifier agreement rates or exploring how the different classifiers relate to specific physical qualities of lesions, may be investigated in the future. We believe that transfer learning for CADx can be further investigated in terms of types of training data, architectures used, and ensembling methods, further paving the way for improved CADx and precision medicine in general.

Acknowledgments

This work was partially funded by NIH grant U01 CA195564 and the University of Chicago Metcalf program. MLG is a stockholder in R2/Hologic, co-founder and equity holder in Quantitative Insights, and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba. HL received royalties from Hologic. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

References

1. M. L. Giger, J. Boone, and H. Chan, "History and status of CAD and quantitative image analysis," *Med. Phys.* **35**(12), 5799–5820 (2008).
2. L. Hadjiiski, B. Sahiner, and H.-P. Chan, "Advances in CAD for diagnosis of breast cancer," *Curr. Opin. Obstet. Gynecol.* **18**(1), 64–70 (2006).
3. M. L. Giger, N. Karssemeijer, and J. A. Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annu. Rev. Biomed. Eng.* **15**, 327–357 (2013).
4. W. Zhang et al., "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Med. Phys.* **21**(4), 517–524 (1994).
5. S.-C. B. Lo et al., "Artificial convolution neural network for medical image pattern recognition," *Neural Networks* **8**(7), 1201–1214 (1995).
6. A. R. Jamieson, K. Drukker, and M. L. Giger, "Breast image feature learning with adaptive deconvolutional networks," *Proc. SPIE* **8315**, 831506 (2012).
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira et al., Eds., Vol. **25**, pp. 1097–1105, Curran Associates, Inc. (2012).
8. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).
9. C. Szegedy et al., Going deeper with convolutions, CoRR abs/1409.4842 (2014).
10. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.* **22** (10), 1345–1359 (2010).
11. J. Yosinski et al., How transferable are features in deep neural networks? CoRR abs/1411.1792 (2014).
12. A. S. Razavian et al., "CNN features off-the-shelf: anastounding baseline for recognition," in *2014 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 512–519, IEEE (2014).
13. J. Donahue et al., "Decaf: a deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531* (2013).
14. H. Li et al., "Evaluation of computer-aided diagnosis on a large clinical full-field digitalmammographic dataset," *Acad. Radiol.* **15**(11), 1437–1445 (2008).
15. O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision* **115**(3), 211–252 (2015).
16. Z. Huo et al., "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**(3), 155–168 (1998).
17. W. Chen et al., "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn. Reson. Med.* **58**(3), 562–571 (2007).
18. J. Kittler et al., "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 226–239 (1998).
19. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).
20. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, United Kingdom (2000).
21. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology* **143** (1), 29–36 (1982).
22. C. E. Metz, "Basic principles of roc analysis," *Semin. Nucl. Med.* **8**(4), 283–298 (1978).
23. T. Hothorn et al., "The design and analysis of benchmark experiments," *J. Comput. Graphical Stat.* **14** (3), 675–699 (2005).
24. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038* (2014).
25. Y. Bar et al., "Chest pathology detection using deep learning with non-medical training," in *2015 IEEE 12th Int. Symp. on Biomedical Imaging (ISBI)*, pp. 294–297, IEEE (2015).
26. Y. Anavi et al., "A comparative study for chest radiograph image retrieval using binary texture and deep learning classification," in *2015 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2940–2943, IEEE (2015).
27. H. Wang et al., "Mitosis detection in breast cancer pathology images bycombining handcrafted and convolutional neural network features," *J. Med. Imaging* **1**(3), 034003 (2014).
28. G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, pp. 652–660, Springer, New York City (2015).

Benjamin Q. Huynh currently works on applying deep learning methods to medical image analysis. His research interests include computational statistics, computer vision, and nonparametric Bayesian techniques with applications to biomedical tasks.

Hui Li has been working on quantitative imaging analysis on medical images for over a decade. His research interests include breast cancer risk assessment, diagnosis, prognosis, and response to therapy, understanding the relationship between radiomics and genomics, and their future roles in precision medicine.

Maryellen L. Giger has been working, for multiple decades, on computer-aided diagnosis and quantitative image analysis methods in cancer diagnosis and management. Her research interests include understanding the role of quantitative radiomics, computer vision, and deep learning in personalized medicine.