Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Predicting malignant nodules by fusing deep features with classical radiomics features

Rahul Paul
Samuel H. Hawkins
Matthew B. Schabath
Robert J. Gillies
Lawrence O. Hall
Dmitry B. Goldgof

SPIE.

Rahul Paul, Samuel H. Hawkins, Matthew B. Schabath, Robert J. Gillies, Lawrence O. Hall, Dmitry B. Goldgof, "Predicting malignant nodules by fusing deep features with classical radiomics features," *J. Med. Imag.* **5**(1), 011021 (2018), doi: 10.1117/1.JMI.5.1.011021.

Predicting malignant nodules by fusing deep features with classical radiomics features

Rahul Paul,[a] Samuel H. Hawkins,[a] Matthew B. Schabath,[b] Robert J. Gillies,[c] Lawrence O. Hall,[a] and Dmitry B. Goldgof[a,*]
[a]University of South Florida, Department of Computer Science and Engineering, Tampa, Florida, United States
[b]H. Lee Moffitt Cancer Center & Research Institute, Department of Cancer Epidemiology, Tampa, Florida, United States
[c]H. Lee Moffitt Cancer Center & Research Institute, Department of Cancer Physiology, Tampa, Florida, United States

Abstract. Lung cancer has a high incidence and mortality rate. Early detection and diagnosis of lung cancers is best achieved with low-dose computed tomography (CT). Classical radiomics features extracted from lung CT images have been shown as able to predict cancer incidence and prognosis. With the advancement of deep learning and convolutional neural networks (CNNs), deep features can be identified to analyze lung CTs for prognosis prediction and diagnosis. Due to a limited number of available images in the medical field, the transfer learning concept can be helpful. Using subsets of participants from the National Lung Screening Trial (NLST), we utilized a transfer learning approach to differentiate lung cancer nodules versus positive controls. We experimented with three different pretrained CNNs for extracting deep features and used five different classifiers. Experiments were also conducted with deep features from different color channels of a pretrained CNN. Selected deep features were combined with radiomics features. A CNN was designed and trained. Combinations of features from pretrained, CNNs trained on NLST data, and classical radiomics were used to build classifiers. The best accuracy (76.79%) was obtained using feature combinations. An area under the receiver operating characteristic curve of 0.87 was obtained using a CNN trained on an augmented NLST data cohort. [DOI: 10.1117/1.JMI.5.1.011021]

Keywords: nonsmall cell lung cancer; National Lung Screening Trial; convolutional neural network; transfer learning; deep features; radiomics.

Paper 17204SSRRRR received Jul. 6, 2017; accepted for publication Feb. 19, 2018; published online Mar. 21, 2018.

1 Introduction

Worldwide, lung cancer[1] is the most frequently diagnosed cancer, consisting of 13.3% of all cancers diagnosed. Overall, lung cancer has a 5-year survival rate of 17.7%. In the US, more than 200,000 are diagnosed every year, and nearly 150,000 people die from lung cancer. In the US, lung cancers are primarily categorized into two types: small cell lung cancer (SCLC) and nonsmall cell lung cancer (NSCLC). NSCLC, which generally spreads and grows slower than SCLC, is the most common type of lung cancer (80% to 85% of all lung cancers). Lung cancer is often asymptomatic until the disease has reached an advanced stage. Recently, an early detection method was shown to decrease mortality rates for NSCLC. The National Lung Screening Trial (NLST)[2] demonstrated that lung cancer mortality was reduced significantly, by 20%, among high-risk individuals that were screened using low-dose computed tomography (LDCT) versus a standard chest radiograph. However, in the NLST, 96.4% of the intermediate pulmonary nodules (IPNs) identified by LDCT were false positives. Thus, an accurate noninvasive approach is necessary as a clinical decision tool to better identify nodules, especially IPNs, in the lung cancer screening setting.

Radiomics[3,4] is an approach to extract quantitative features from the standard of care medical images. The features can then be used in statistical analysis, machine learning, or other high dimensional analysis. Radiomics features are expected to provide an accurate noninvasive approach to better track nodules during lung cancer screening.

Deep learning,[5] an emerging area of research in the machine learning field, has recently become widely used for classification and categorization. Deep learning algorithms learn by enabling a multilevel representation of data via multiple hidden layers in a neural network. Though Fukushima[6] introduced the "neocognitron," one of the first artificial neural network models, convolutional neural network (CNN) of LeCun et al.[7] gained popularity for classification tasks. Krizhevsky et al.[8] with his "ALEXNET" architecture achieved a significant improvement for large-scale classification of images in the ILSVRC, 2012. Girshick et al.[9] showed that, when data are scarce, it is beneficial to use a pretrained network and then fine-tune it on new data. Erhan et al.[10] assessed how unsupervised learning could be used to most effectively initialize a deep neural network. Donahue et al.[11] examined whether the features extracted from a CNN can be applied to some other object recognition tasks with few training cases. This leads to the idea of transfer learning, applying previously learned knowledge in a different domain. Raina et al.[12] proposed an approach for supervised classification using unlabeled data.

This work is focused on utilizing deep learning to increase the accuracy and area under the receiver operating characteristic curve (AUC)[13–15] of classifier predictions of LDCT image lung nodules becoming malignant within 2 years. As labeled data are

*Address all correspondence to: Dmitry B. Goldgof, E-mail: goldgof@mail.usf.edu

2329-4302/2018/$25.00 © 2018 SPIE

limited, a major focus is on the transfer learning approach, whereas existing CNNs trained on the ImageNet dataset[16] are used to extract features for classifying lung nodules. The images that were used by others to train the CNNs and the lung nodule images are different in two aspects: first, the lung images are grayscale, whereas ImageNet contains color RGB images of objects and natural scenes, and the lung nodule images are much smaller in size, i.e., fewer pixels per object of interest. In previous work[17,18] on categorizing lung tumors by expected survival time, we normalized grayscale CT images using just the red channel and demonstrated that deep features extracted from a pretrained CNN can be used effectively and classification accuracy could be enhanced by merging classical radiomics features with deep features.

Here, the data consist of LDCT screening images of lung nodules, which may or may not become malignant. There are several ways to use CNNs. In transfer learning, the features from a CNN are used with a conventional machine learning algorithm to build a classification model. Experiments were performed with several learning algorithms, as the best for this type of work is currently unknown. A CNN with augmented training data may be directly learned. With limited data, the CNN must be small to enable learning and minimize overfitting. Once trained, a CNN can be used for predictions or features can be extracted from the network. Experiments were performed with both approaches to see how they compared.

Our data consisted of two cohorts from NLST data: cohort 1 used for training and the independent cohort 2 data used for evaluating trained classifiers. As deep neural networks are notoriously affected by their many parameters, we designed several small CNN architectures trained on cohort 1 of NLST data. For the trained CNN architectures, two approaches were experimented with: classification by using the sigmoid output layer of the CNN, extracting features from the last fully connected hidden layer of the CNN architecture, and then applying feature selection and using the selected features to build a classifier.

This paper examines four ways to increase the prediction power of classifiers built from screening CT images of lung nodules that are currently benign. First, we investigate whether classification using deep features extracted from a CNN trained on camera images can yield similar or improved results compared to using classical radiomics features like textures, shape, and size.[19] To do this, features were extracted from three slightly different trained deep neural networks, each trained on the same data. The reason for this is to see how parameters affect performance. Second, we hypothesized that the combination of classical radiomics features with deep features improves the classification result. We did not attempt to decide the right number of features on the training data. Instead, we used 5, 10, 15, and 20 of the best features from a feature selection method. Third, as the lung nodule images are grayscale, we examined the effects on classification results after sending nodule images through different color channels (red/green/blue) of a network trained on camera images. Fourth, even with small data, we evaluated the hypothesis that lung nodules can be effectively put into groups, which will remain benign or will become malignant by designing new CNN architectures and using features extracted from the new CNN. There were three architectures tried with an attempt to progressively minimize the number of weights. While more weights can provide more complex classifiers, with small data, they can overfit or simply train poorly, but small networks may not give good accuracy, hence, the spread of 3. Performance was measured by accuracy and AUC on unseen data. While we did not use results on the test data to modify experiments, the number of experiments could have yielded a result that would not otherwise have been found.

Comparisons with previous work,[19] which used classical radiomics features for classification, show that deep neural networks can be used to recreate the results with no explicit feature selection/extraction even with small training data. Furthermore, performance can be improved by combining classical and deep features.

2 Materials and Methods

2.1 *Study Participants, Data, and Feature Extraction*

In this study, we utilized LDCT images and data from the NLST study. Briefly, the NLST[2] was a randomized clinical trial comparing LDCT and chest X-rays conducted on 53,454 current or former smokers at 33 medical centers across the US. The NLST study had a baseline (T0) screening and two follow-up screenings ~1 year (T1) and 2 years (T2) after the baseline.

The pixel size varied across the image set as follows. Average image pixel size was 0.6642 mm, the standard deviation was 0.072 mm, the min was 0.4844 mm, and the max was 0.8594 mm. Normalizing the pixel size can introduce artifacts and was not done. The slice thickness in mm varied from 3.2 (maximum slice thickness) to 1 (minimum slice thickness) across the image set. The majority of cases of both cohorts had a 2.5-mm slice thickness with the rest mostly at 2 mm. The reconstruction field-of-view ranged between 248 and 460 mm. Tube potential varied from 140 to 120 kVp. No corrections were done based on the variations, and this is a potential limitation of this study.

Based on prior work from Schabath et al.,[20] we selected subsets of participants of screen-detected lung cancers (SDLC) and nodule positive controls from the LDCT-arm of the NLST. We chose two SDLC patient cohorts for this study, as depicted in Fig. 1. The lung cancer cases and the nodule positive controls were 1:2 frequency matched on age at randomization, sex, and smoking. At the baseline (T0) screen, both the lung cancer cases and the nodule positive controls had a positive screen that was not diagnosed as lung cancer. The lung cancer cases were diagnosed at either the first (T1) or second (T2) follow-up screen while the nodule positive controls had three consecutive positive screens (T0 to T2) not diagnosed as lung cancer. The lung cancer cases and nodule positive controls were split into a training cohort and a testing cohort. For both cohorts, we used cases from time T0, baseline screening, for training. Cohort 1 consisted of 261 cases of which 85 nodules became cancer and 176 did not become cancer. Cohort 2 consisted of a unique set of 237 cases of which 85 nodules became cancer and 152 did not become cancer. Cohort 1 was used as the training set and cohort 2 as the test set in our study.[21] Figure 1 shows a flowchart of the two cohorts. Nodule size was not used in selecting cohorts. Dataset selection is described in more detail in Refs. 19 and 20.

Table 1 presents the demographics of the lung cancer cases and nodule positive controls. There were no statistically significant differences between the lung cancer cases and nodule positive controls for age, sex, race, ethnicity, and smoking. Nodule segmentation was performed using the Definiens software suite (Cambridge, Massachusetts).[22]

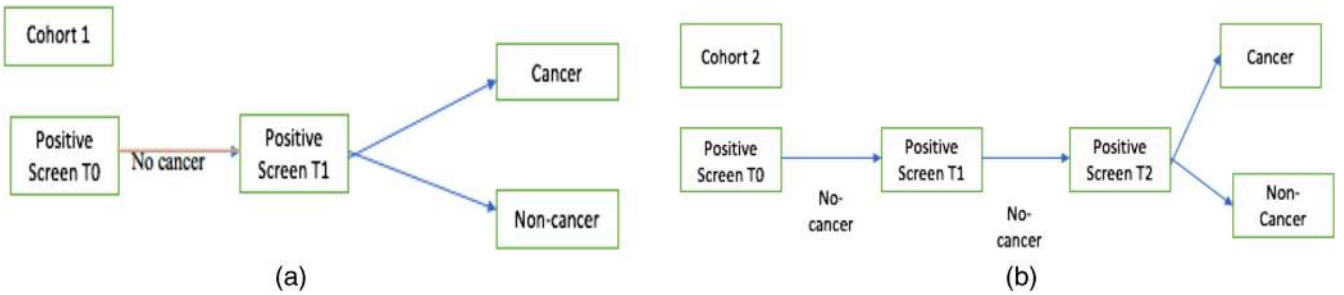

Fig. 1 Flowchart of (a) cohort 1 and (b) cohort2.

Table 1 Demographic and clinical characteristics of dataset.

Characteristic	Lung cancer cases ($n = 170$)	Nodule positive cases ($n = 328$)	p value
Age, mean ± SD, years	63.7 ± 5.11	63.5 ± 5.1	0.66
Sex, *n* (%)			
Male	94 (55.3)	192 (58.5)	0.28
Female	76 (44.7)	136 (41.5)	
Race, *n* (%)			
White	161 (94.7)	315 (96.0)	0.49
Black, Asian, other	9 (5.3)	13 (4.0)	
Ethnicity, *n* (%)			
Hispanic or Latino	0 (0.0)	2 (0.6)	0.55
Neither Hispanic nor Latino and unknown	170 (100.0)	326 (99.4)	
Smoking, *n* (%)			
Current	89 (53.4)	175 (53.4)	0.85
Former	81 (47.6)	153 (46.6)	
Pack-years smoked, mean ± SD			
Current smokers	63.2 (25.8)	62.0 (21.3)	0.69
Former smokers	64.5 (27.6)	63.7 (26.8)	0.83
Stage, *n* (%)			
I	117 (68.8)		
II	12 (7.1)		
III	21 (12.3)		
IV	18 (10.6)		
Carcinoid, unknown	2 (1.2)		
Histologic subtype, *n* (%)			
Adenocarcinoma	108 (63.5)		
Squamous cell carcinoma	38 (22.4)		
Other, NOS, unknown	24 (14.1)		

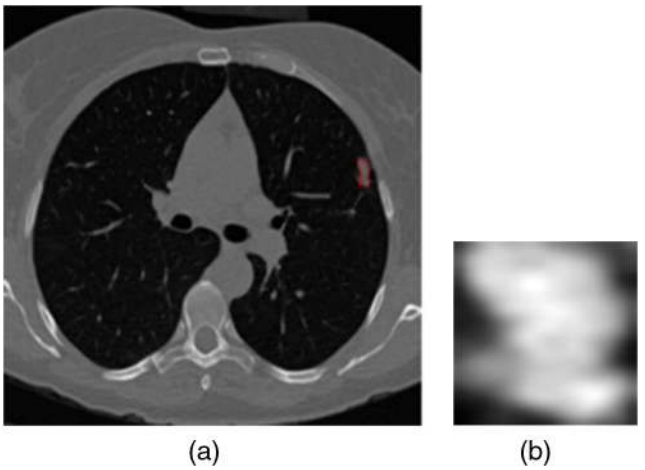

Fig. 2 (a) Lung image with nodule inside outlined by red (pixel size = 0.527 mm) and (b) nodule resized to 224 × 224.

In previous work, 219 radiomics features (see Ref. 19), consisting of size, shape, gray level co-occurrence matrix, wavelet, and laws features, were extracted from nodules. We merged radiomics features with deep features extracted from warped nodules.

In our study, for each cohort's cases, we chose the single slice that had the largest nodule area. We then extracted only the nodule region from each slice by taking a rectangular box of the image that completely covered the nodule region. Since it would be resized to match the input size of a CNN, as discussed in the proceeding, we called the resized nodule "warped." In Fig. 2, we show a resized nodule along with the actual LDCT scan slice. While each nodule was a different size, the input size required by pretrained CNN was 224 × 224 and, hence, a bicubic interpolation was used for resizing the images. In the Appendix, we show six additional lung images with nodules outlined and resized to 224 × 224. The range includes very small nodules, medium-sized and large-sized nodules.

2.2 *Convolutional Neural Networks and Transfer Learning*

CNNs,[23,24] a variant of multilayer feed forward networks, are recently used widely in image classification and object recognition tasks. A CNN architecture can be designed using a few convolutional layers, often followed by a max pooling layer, then fully connected layers and an activation function layer. As CNN consists of many layers, it needs to learn many connection weights, and for a big network, a lot of data are typically needed to avoid under- or overfitting. The dataset

Table 2 Pretrained CNN architectures [each architecture contains five convolutional layers (conv 1 to 5) and three fully connected layers (full 1 to 3)]. In the conv layers, the first number indicates how many convolutions and the next two numbers indicate the kernel size. St. is stride and pad is the number of 0s added to the right of the image. LRN indicates a type of regularization layer. The number in front of pool is the size of the max pool region.

Arch.	Conv 1	Conv 2	Conv 3	Conv 4	Conv 5	Full 1	Full 2	Full 3
Vgg-F	$64 \times 11 \times 11$	$256 \times 5 \times 5$	$256 \times 3 \times 3$	$256 \times 3 \times 3$	$256 \times 3 \times 3$	4096 dropout	4096 dropout	1000 softmax
	St. 4, pad 0	St. 1, pad 2	St. 1, pad 1	St. 1, pad 1	St. 1, pad 1			
	LRN, x2 pool	LRN, x2 pool			x2 pool			
Vgg-M	$96 \times 7 \times 7$	$256 \times 5 \times 5$	$512 \times 3 \times 3$	$512 \times 3 \times 3$	$512 \times 3 \times 3$	4096 dropout	4096 dropout	1000 softmax
	St. 2, pad 0	St. 2, pad 1	St. 1, pad 1	St. 1, pad 1	St. 1, pad 1			
	LRN, x2 pool	LRN, x2 pool			x2 pool			
Vgg-S	$64 \times 11 \times 11$	$256 \times 5 \times 5$	$512 \times 3 \times 3$	$512 \times 3 \times 3$	$512 \times 3 \times 3$	4096 dropout	4096 dropout	1000 softmax
	St. 4, pad 0	St. 1, pad 1	St. 1, pad 1	St. 1, pad 1	St. 1, pad 1			
	LRN, x3 pool	x2 pool			x3 pool			

we were using has just 276 cases for training, which is rather small for a CNN. So, a transfer learning approach was tried using a large network trained on the ImagNet set of camera images. Transfer learning[25,26] is a method where previously learned knowledge is applied to another task and the task domains may be different. In our case, the domain is very different. ImageNet consists of natural camera images and does not include any type of lung nodule or cancer image. Our image set consists of only lung nodules in CT images. We experimented with three different pretrained CNN's [vgg (visual geometry group)-m/vgg-f/vgg-s][27] in this study using a MATLAB toolbox named MATCONVNET.[28] The f, m, and s after vgg- stand for fast, medium, and slow and refer to training time (so partly the number of weights). We obtained deep features from the outputs of the last fully connected layer after applying the activation function using a rectified linear unit (post-ReLU), which changes all values <0 to be 0. The LDCT images were grayscale (no color component and we changed the voxel intensities of LDCT images to 0-255 or 8 bit), but the pretrained network was trained on RGB images, so we normalized the images by the average red, green, and blue channel images, and experimented by using each channel separately. The RGB images have three color channels (24-bit image), but the grayscale image had only a single grayscale image (8-bit image). In our previous experiment, we normalized the images of the pretrained network by each color channel separately. This approach lost the information provided by the other two channels. Here, we used the same grayscale LDCT image for each channel to make it somewhat analogous to an RGB image for the CNN. Doing so engages all the weights and exploits all the learned knowledge when extracting features from the pretrained network. Since the images experimented with were smaller than the required input size of the pretrained CNN (224×224), we used bicubic interpolation for resizing. The dimension of a deep feature vector extracted for each image was 4096. The features were the output of the last fully connected layer (the full 2 layer as shown in Table 2) before the output layer in an ImageNet pretrained CNN. The architectures and parameters for the pretrained CNNs used are described in Table 2.

We further experimented with three different CNN architectures by training from scratch (i.e., a random set of initial weights). We designed the architectures using Keras[29] with Tensorflow[30] as the CNN library. The architectures and parameters used are described in Tables 3–5. For each locally trained CNN architecture, the input image size was 100×100 pixels (used bicubic interpolation for resizing). Different size nodule images (their area varies from 16 to 10,354 pixels) were interpolated to 100×100. We performed the experiment with different input image sizes but obtained the best result from

Table 3 CNN Architecture 1.

Layers	Parameter	Total parameters
Input image	100×100	841,681
Conv 1	$64 \times 5 \times 5$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max pool 1	3×3, pad 0, stride3	
Conv 2	$64 \times 2 \times 2$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max pool 2	3×3, pad 0, stride3	
Dropout	0.1	
Fully connected 1 + ReLU	128	
Fully connected 2 + ReLU	8	
L2 regularizer	0.01	
Dropout	0.25	
Fully connected 3	1 sigmoid	

Table 4 CNN Architecture 2.

Layers	Parameter	Total parameters
Input image	100×100	845,033
Conv 1	$64 \times 5 \times 5$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max Pool 1	3×3, pad 0, stride3	
Conv 2	$64 \times 2 \times 2$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max Pool 2	3×3, pad 0, stride3	
Dropout	0.1	
Fully connected 1 + ReLU	128	
LSTM 1 + ReLU	8	
L2 regularizer	0.01	
Dropout	0.25	
Fully connected 2	1 sigmoid	

Table 5 CNN Architecture 3.

Layers	Parameter	Total parameters
Left branch		
Input image	100×100	39,553
Max pool 1	10×10	
Dropout	0.1	
Right branch		
Input image	100×100	
Conv 1	$64 \times 5 \times 5$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max pool 2a	3×3, pad 0, stride3	
Conv 2	$64 \times 2 \times 2$, pad 0, stride1	
Leaky ReLU	alpha = 0.01	
Max pool 2b	3×3, pad 0, stride3	
Dropout	0.1	
Concatenate left branch + right branch		
Conv 3 + ReLU	$64 \times 2 \times 2$, pad 0, stride1	
Max pool 3	2×2, pad 0, stride2	
L2 regularizer	0.01	
Dropout	0.1	
Fully connected 1	1 sigmoid	

100×100. Here, we used the same grayscale CT image for each channel to make it somewhat analogous to an RGB image for the CNN. Doing so engages all the weights and exploits all the knowledge and information during feature extraction from the pretrained network.

The total number of epochs for training was 200. The learning rate for each architecture was kept constant at 0.0001 with the RMSprop[31] (root mean square propagation) algorithm, which was used for gradient descent optimization. Though we experimented with different batch sizes (8/16/24/32), a batch size of 16 was used for both training and validating the deep convolutional architecture, because it gave the best result. Leaky ReLU (alpha = 0.01), where negative values are occasionally allowed to propagate, was applied in convolutional layers 1 and 2. This provided nonlinearity on the output of the convolutional layers. As all our architectures were shallow, to prevent overfitting, both dropout[32] and L2 regularization[33] were applied before the classification layer.

In Architecture 1, we used two fully connected layers with 128 and 8 units, respectively, before the final classification layer. The total number of parameters was 841,681.

In Architecture 2, we incorporated one fully connected layer with 128 units, followed by one long short term memory (LSTM)[34] layer with 8 units before the final classification layer. LSTM is a type of recurrent neural network layer, consisting of a memory to remember information for a short or long time and various gates to control the flow of information going in or out of the memory. Using this architecture, we investigated whether the advantage of remembering information using LSTM was useful when it was the last layer before final classification instead of a fully connected layer. After using LSTM instead of a fully connected layer, the classification accuracy was further improved. The total number of parameters was a slightly larger 845,033.

Architecture 3 was a cascaded architecture,[35] where images were fed to both the "left" branch of the network, where there was a max pooling layer and more complex "right" branch. The right branch consisted of convolution and max pool layers. The cascading happened after getting the same size output (10×10 vector) from both the left and right branches. Another convolution and max pool layer were used after the cascade instead of using a fully connected layer or LSTM.

As a result, the number of total parameters was reduced by almost 100%[35] and classification accuracy was improved. The total number of parameters used was 39,553. In this architecture, we took image information more directly after applying max-pooling and merged it with information generated after convolutions. Features in the convolution layer are more generic (e.g., blobs, textures, edges, etc.). So, adding image information directly will create more specific information for each case. After merging, another convolution and max pooling layer before the final classification layer maintains the generic information about the image and can provide more features of about the image for getting a better classification result. Figure 3 shows a flowchart of CNN Architecture 3.

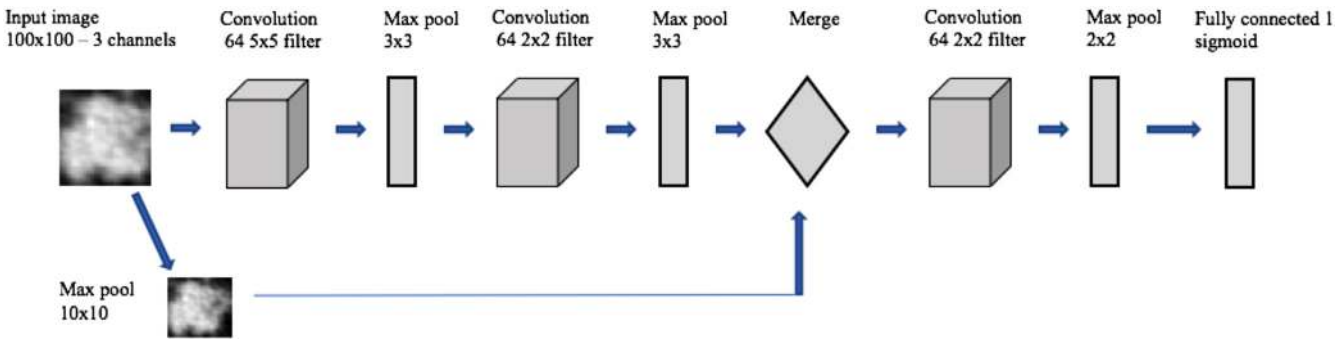

Fig. 3 CNN Architecture 3.

3 Experiments and Results

In the pretrained network, each image had to be normalized by an average image first, and as we know, the nodule images were grayscale images, so they were normalized by one channel (red, green, or blue) only. Experiments were performed with red, green, and blue channels separately, taking one channel at a time and ignoring the other two (i.e., removing the weights and connections of the other two from the CNN). We also used the grayscale image three times to simulate an image with three color channels and did normalization using the appropriate color channel image.

Deep features of dimension 4096 were extracted from the last fully connected layer after applying the ReLU activation function for each image.

We used the symmetric uncertainty[36] feature selector to select the top (5/10/15/20) features.[37]

For classification, we compared five classifiers: naïve Bayes,[38] SVM,[39] decision trees,[40] nearest neighbor,[41] and random forests [200 trees and $\log_2(n)+1$ features].[42]

We used the area under the receiver operating characteristic (ROC) curve[13–15] for performance measurement. For an ROC curve, at different cut-off points, the true-positive rate (Y-axis) is plotted against the false-positive rate (X-axis). For each decision threshold, the point on the ROC curve illustrates a true-positive/false-positive pair. The area under the ROC curve or AUC is a measurement of how well a model can differentiate the lung cancer cases from the control nodules. The maximum possible area under the ROC curve is 1. So, AUC values closer to 1 signify a better predictive model.

Cohort 1 had 261 cases, which is too small for training a new convolutional network from scratch. We performed image augmentation on cohort 1 by rotating each image 15 deg and then applying horizontal and vertical flipping. The total number of images generated from cohort 1 was thus 18,792. We randomly selected 70% of cohort 1 data for training and the remaining 30% was used for validation and each example is a particular subject at T0. The input size for the convolutional network was 100×100 pixels from "warped" images. A sigmoid layer was used for the classification in our CNN architectures. Architecture 3, which gave us the best classification accuracy from the sigmoid layer, was used to extract features of dimension 1024 from the last layer before the classification layer. These new deep feature vectors were used for further analysis by selecting 5/10/15/20 features using symmetric uncertainty and classifiers, as discussed previously.

We experimented with six different approaches to predict malignancy of cohort 2 subjects lung nodules imaged at T0: first, only deep features extracted from warped nodules using a pretrained CNN; second, merging deep features and classical radiomics features; third, constructing new CNN architectures and using a sigmoid layer for classification; fourth, deep features extracted from the newly trained CNN architecture; fifth, merging deep features from a lung nodule trained CNN and classical radiomics features; and finally, merging deep features from lung nodules, classical radiomics features, and deep features from a network pretrained on a camera image dataset (i.e., vgg architectures). Cohort 2 cases were not used in training and served as an unseen test set to validate the predictive power of our approaches.

With just deep features from a pretrained CNN, the best accuracy of 75.1% with a 0.74 AUC was obtained from warped nodules using the deep features obtained from vgg-s. The classifier used to obtain the result was a nearest neighbor (11 neighbors) classifier. In Table 6, we show only the best results obtained by using each color channel separately.

We then merged the 219 radiomics features with deep features extracted from warped nodules. We selected the top 5/10/15/20 features using the symmetric uncertainty feature selector separately on both deep features and classical radiomics features and merged them together to make a feature vector of

Table 6 Best results using each color channel using deep features only.

Color channel	Pretrained net	No. of features	Classifier	AUC (SE)	Accuracy
Red	Vgg-m	15	Nearest neighbor (7 neighbors)	0.65 (0.0359)	71.3
Blue	Vgg-s	15	Nearest neighbor (7 neighbors)	0.72 (0.0329)	73.4
Green	Vgg-s	20	Random forests (200 trees)	0.68 (0.0347)	68.3
RGB	Vgg-s	5	Nearest neighbor (11 neighbors)	0.74 (0.0318)	75.1

Table 7 Best results using each color channel and merging deep and classical radiomics features.

Color channel	Pretrained net	No. of features	Classifier	AUC (SE)	Accuracy
Red	Vgg-m	20	Random forests (200 trees)	0.793 (0.0285)	74.7
Blue	Vgg-f	5	Random forests (200 trees)	0.79 (0.0287)	74.2
Green	Vgg-m	15	Random forests (200 trees)	0.78 (0.0294)	73
RGB	Vgg-s	15	Random forests (200 trees)	0.78 (0.0294)	75.1

10/20/30/40 features. The best accuracy of 75.1% was obtained by merging the top 15 selected features from classical radiomics features and deep features from a vgg-s pretrained network using a random forests classifier. The best AUC of 0.793 was obtained by merging the top 20 selected features from classical radiomics features and deep features from a vgg-m pretrained network. The classifier used to obtain the result was a random forests classifier. In Table 7, we show only the best result obtained by merging deep features from each channel and classical radiomics features.

We designed three convolutional architectures and trained them on augmented images of cohort 1 and tested on cohort 2. Maximum classification accuracy of 76% with AUC 0.87 was obtained from Architecture 3.

We assessed the significance of the improvement of the AUC values between the current and previous best result[19] by calculating the standard error (SE) of the area for each AUC score using Eq. (1):

$$\mathrm{SE} = \sqrt{\frac{A(1-A) + (na-1)(Q-A^2) + (nn-1)(Q-A^2)}{na * nn}}, \tag{1}$$

Table 8 Results using different CNN architecture.

CNN architectures	AUC (SE)	Classification accuracy
CNN Architecture 1	0.82 (0.0266)	75.1%
CNN Architecture 2	0.86 (0.0242)	75.5%
CNN Architecture 3	0.87 (0.0224)	76%

where A is the calculated AUC, na and nn are the number of benign and malignant cases, respectively, and $Q1$ and $Q2$ are estimated by $Q1 = A/(2-A)$ and $Q2 = 2*(A*A)/(1+A)$.

Using the SE values from the two AUCs, AUC_1 and AUC_2, we calculated the z value using Eq. (2):

$$z = \frac{|\mathrm{AUC}_1 - \mathrm{AUC}_2|}{\sqrt{\mathrm{SE}^2_{\mathrm{AUC}_1} + \mathrm{SE}^2_{\mathrm{AUC}_2}}}. \tag{2}$$

From the z score, the p value was obtained and performance significance was evaluated at $p = 0.1$ and $p = 0.05$. Improvement of AUC was not statistically significant at $p = 0.05$, as it was 0.0656. In Table 8, we show the best result obtained from different CNN architectures.

CNN Architecture 3 was used as a pretrained network and 1024 features were extracted from the last layer (max pool $3-4\times4$ after pooling and 64 convolutions) before the classification layer. The symmetric uncertainty feature ranking algorithm was used to extract the top 5/10/15/20 deep features. Random forests, nearest neighbor, SVM, naïve Bayes, and decision tree classifiers were applied on the extracted new deep features for classification.

We further investigated merging the new deep features with the classical radiomics features to make a feature vector of size 10/20/30/40 and obtained 76.37% accuracy (AUC 0.75). We obtained further improved accuracy of 76.79% (AUC 0.78) by merging new deep features, classical radiomics features, and deep features obtained from the pretrained Vgg-s network (feature vector size 15/30/45/60). In Table 9, we show the best result obtained by extracting features from our CNN architectures and when combined with other features. Figure 4 shows a flowchart of the feature fusion process.

Table 9 Best results by extracting deep features from CNN Architecture 3.

Features	No. of features	Classifier	AUC (SE)	Accuracy
Deep features only	15	Random forests (200 trees)	0.62 (0.0369)	68.7
Deep features + classical radiomics features	20	Random forests (200 trees)	0.75 (0.0312)	76.37
Deep features + classical radiomics features+ vgg-f architecture's deep features	15	Random forests (200 trees)	0.77 (0.03)	75.1
Deep features + classical radiomics features + vgg-m architecture's deep features	15	Random forests (200 trees)	0.76 (0.0306)	72.5
Deep features + classical radiomics features + vgg-s architecture's deep features	10	Random forests (200 trees)	0.78 (0.0294)	76.79

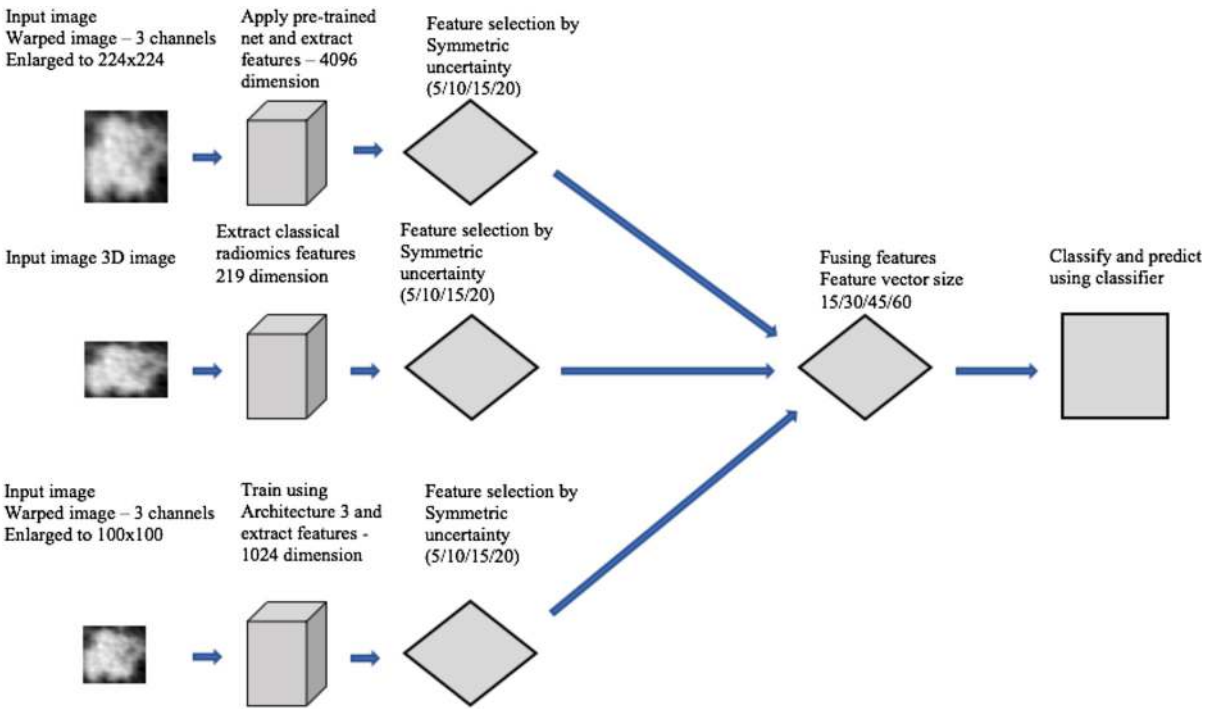

Fig. 4 Flowchart of feature fusion approach.

4 Discussion and Conclusions

Using transfer learning, features can be obtained from the last fully connected layer of a pretrained network using lung LDCT image data as input. LDCT screening data with labels will likely be a small set for training a new deep neural network, as it is in this research. Predictions were performed on baseline lung nodules from an unseen test set (cohort 2) to predict which patients would develop lung cancer by the second follow-up screen (T2). In this paper, we show transfer learning features, as well as training a new CNN will yield higher AUC when compared to our previous study[19] using only quantitative radiomics features. Quantitative radiomics features are mainly generated based on the tumor size, shape, histogram-based features, and texture. Deep features likely have a relation to texture-based features and, perhaps, shape, but are more opaque. So, we also fused deep features with quantitative radiomics features to evaluate whether fusing different types of features will provide better results.

In our study, three different convolutional networks pretrained on the ImageNet dataset were used and features were obtained from the last hidden layer after applying the ReLU activation function (post-ReLU). In this paper, we discussed two different approaches: using deep features only from warped nodules and merging deep features from warped nodules with classical radiomics features. In our previous study,[19] we used the red channel to normalize input images. In our current study, we performed further analyses, using red, green, and blue channels separately and using all three channels simultaneously. We generated a multichannel simulated RGB image, by using the same grayscale image three times. The best result obtained using only deep features extracted from the vgg-s pretrained network was 75.10% (AUC 0.74) with three channels using a nearest neighbor classifier. The accuracy is nearly as good as handcrafted features yet based on features from a network trained on color camera images. By merging the deep features with quantitative features, an improved AUC of 0.79 using a random forests classifier was enabled.

Tuning a pretrained ImageNet with our cohort 1 examples was attempted. However, the tuned networks features did not result in accuracy or AUC that was better than reported here.

We also trained our own CNN architectures using augmentation and obtained 76% accuracy (AUC 0.87) from Architecture 3. The AUC was significantly better compared to the previous best result[19] of 0.81 (SE = 0.273). We used our CNN Architecture 3 as a pretrained network to extract features of 1024 dimension from the max pool—layer 3. Using only the new deep features, the best accuracy obtained was 68.77% (AUC 0.62). We then merged the new deep features with the classical radiomics features and obtained an accuracy of 76.37% (AUC 0.75). By merging the new deep features, classical radiomics features, and the deep features from the pretrained vggs architecture, the best overall accuracy of 76.79% (AUC 0.78) was obtained. This matched the best accuracy previously attained with just radiomics features.

As the nodule images were grayscale and input images for pretrained CNN were color (RGB), how to choose a single-color channel from red, green, blue, or whether to use all of them is unclear. Here, we did a detailed analysis on the tradeoffs between color channels for extracting features from pretrained CNNs. We extracted deep features from an ImageNet trained CNN by sending nodule images through different color channels (red/blue/green) separately, as well as copied the nodule image to generate a RGB (24 bit image) and extract features. We noticed that features from the RGB outperformed the individual red, blue, and green channel. So, it shows that color channels can have a serious impact on performance while using pretrained CNNs to extract features from medical images. Merging ImageNet trained CNN features with radiomics features generates a newly designed feature vector for predictive analysis.

Our work also shows that it is possible to train a good performing deep neural network on small medical datasets using augmentation. The best-known AUC (0.87) on this data was obtained from a deep neural network trained on cohort 1. Using our designed CNN network trained on the NLST data as a pretrained CNN for feature extraction was also helpful, as it was solely trained on augmented nodule images. Merging these newly obtained deep features with classical radiomics features generates more powerful feature vectors, which eventually improved performance.

Generally, features that have at most small correlations are preferable for classification. We checked the correlation between deep features (transfer learning features/CNN extracted features) and traditional features and found the correlation between those features was low (in $[0.5, -0.5]$). So, constructing a new feature column by fusing quantitative features with deep features potentially added more information to the newly constructed feature column to enhance classification performance.

From this paper, three conclusions were obtained, which will be utilized in future research. First, we proposed a simple and effective CNN architecture with a small number of parameters useful for smaller (medical) datasets. Second, we show features obtained using transfer learning from all the channels of a CNN pretrained on a large corpus of camera images performed better than features extracted using any single channel. Third, we also constructed a new feature set by fusing quantitative features with deep features, which in turn enhanced classification performance.

One limitation of our study is not using a validation set for the transfer learning approach. So, we simply report all results on the test data. It might be that with a validation set, some combinations would not be tested. We will assess this limitation in our future work.

Appendix

In this appendix, we show six representative lung images with nodules outlined in red along with the nodule resized by bicubic interpolation to 224×224. Figures 5 and 6 represent the larger nodules, Figs. 7 and 8 represent medium sized nodules and Figs. 9 and 10 represent very small nodules.

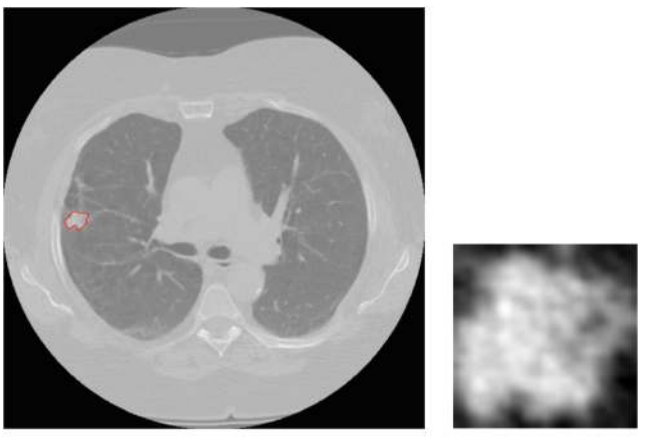

Fig. 5 (a) Lung image with nodule inside outlined by red (pixel size = 0.625 mm) and (b) nodule resized to 224×224.

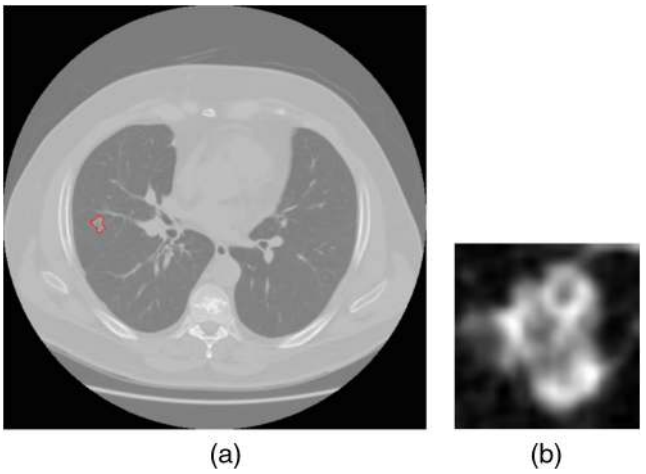

Fig. 6 Larger nodules: (a) lung image with nodule inside outlined by red (pixel size = 0.74 mm) and (b) nodule resized to 224×224.

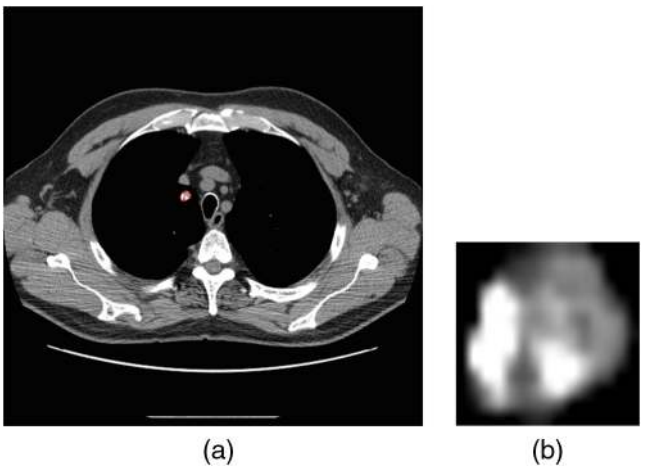

Fig. 7 Medium sized nodules: (a) lung image with nodule inside outlined by red (pixel size = 0.74 mm) and (b) nodule resized to 224×224.

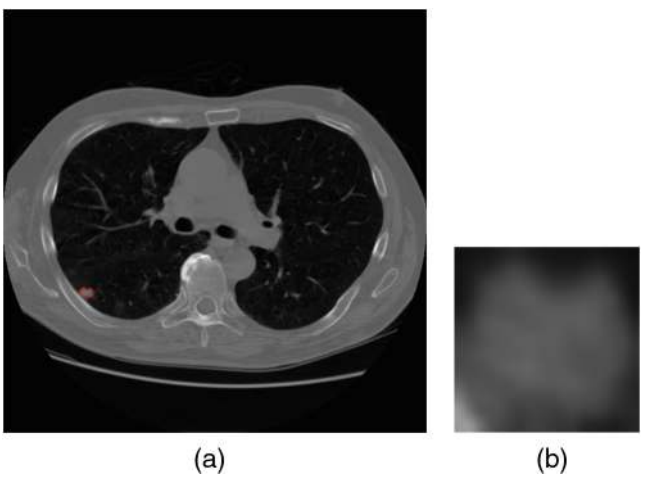

Fig. 8 Very small nodules: (a) lung image with nodule inside outlined by red (pixel size = 0.703 mm) and (b) nodule resized to 224×224.

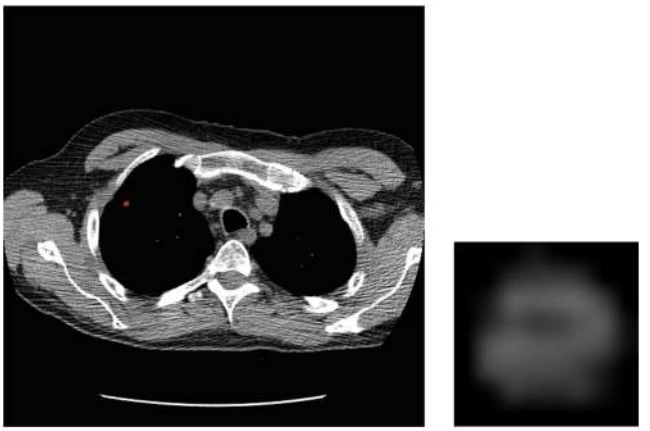

Fig. 9 (a) Lung image with nodule inside outlined by red (pixel size = 0.656 mm) and (b) nodule resized to 224×224.

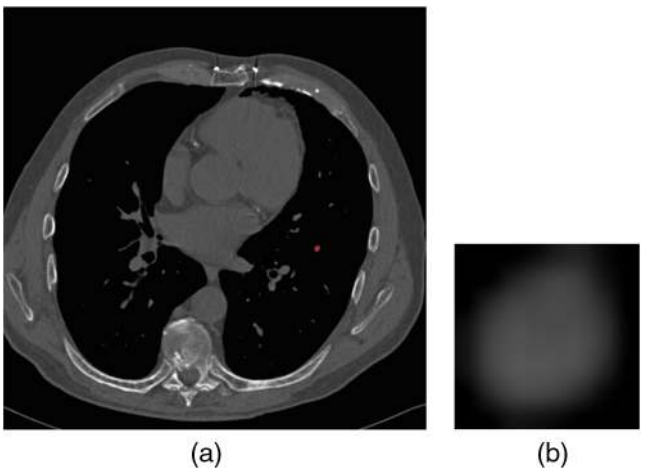

Fig. 10 (a) Lung image with nodule inside outlined by red (pixel size = 0.65 mm) and (b) nodule resized to 224×224.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

This research partially supported by the National Institutes of Health under grants (NIH U01 CA143062), (NIH U24 CA180927) and (NIH U01 CA200464), National Science Foundation under award number 1513126 and by the State of Florida Dept. of Health under grant (4KB17).

References

1. American Cancer Society, *Cancer Facts and Figures 2016*, American Cancer Society, Atlanta (2016).
2. National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *N. Engl. J. Med.* **365**, 395–409 (2011).
3. P. Lambin et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer* **48**(4), 441–446 (2012).
4. V. Kumar et al., "Radiomics: the process and the challenges," *Magn. Reson. Imaging* **30**(9), 1234–1248 (2012).
5. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
6. K. Fukushima, "Neocognitron: a hierarchical neural network capable of visual pattern recognition," *Neural Networks* **1**(2), 119–130 (1988).
7. Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.* **1**(4), 541–551 (1989).
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
9. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 580–587 (2014).
10. D. Erhan et al., "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.* **11**, 625–660 (2010).
11. J. Donahue et al., "DeCAF: a deep convolutional activation feature for generic visual recognition," in *Int. Conf. on Machine Learning*, Vol. 32, pp. 647–655 (2014).
12. R. Raina et al., "Self-taught learning: transfer learning from unlabeled data," in *Proc. of the 24th Int. Conf. on Machine Learning*, pp. 759–766 (2007).
13. K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian J. Intern. Med.* **4**(2), 627 (2013).
14. J. A. Hanley, "Receiver operating characteristic (ROC) methodology: the state of the art," *Crit. Rev. Diagn. Imaging* **29**(3), 307–335 (1989).
15. 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).
16. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 248–255 (2009).
17. R. Paul et al., "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT," in *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC '16)*, 002570 (2016).
18. R. Paul et al., "Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma," *Tomogr.: J. Imaging Res.* **2**(4), 388–395 (2016).
19. S. Hawkins et al., "Predicting malignant nodules from screening CT scans," *J. Thorac. Oncol.* **11**(12), 2120–2128 (2016).
20. M. B. Schabath et al., "Differences in patient outcomes of prevalence, interval, and screen-detected lung cancers in the CT arm of the national lung screening trial," *PLOS One* **11**(8), e0159880 (2016).
21. D. Cherezov et al., "Improving malignancy prediction through feature selection informed by nodule size ranges in NLST," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp. 001939–001944 (2016).
22. Definiens Developer XD, *2.0.4 User Guide*, Definiens AG, München, Germany (2009).
23. M. Oquab et al., "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1717–1724 (2014).
24. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conf. on Computer Vision*, pp. 818–833 (2014).
25. R. Caruana, "Multitask learning," in *Learning to Learn*, pp. 95–133, Springer, Boston, Massachusetts (1998).
26. S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *Advances in Neural Information Processing Systems*, pp. 640–646 (1996).
27. K. Chatfield et al., "Return of the devil in the details: delving deep into convolutional nets," arXiv preprint arXiv:1405.3531 (2014).
28. A. Vedaldi and K. Lenc, "Matconvnet: convolutional neural networks for MATLAB," in *Proc. of the 23rd ACM Int. Conf. on Multimedia*, pp. 689–692 (2015).
29. N. Ketkar, "Introduction to Keras," in *Deep Learning with Python*, pp. 95–109, Apress, Berkeley, California (2017).
30. M. Abadi et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467 (2016).
31. T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks Mach. Learn.* **4**(2), 26–31 (2012).

32. N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).
33. A. Y. Ng, "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," in *Proc. of the Twenty-First Int. Conf. on Machine Learning* (2004).
34. F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Comput.* **12**(10), 2451–2471 (2000).
35. H. Li et al., "A convolutional neural network cascade for face detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5325–5334 (2015).
36. L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," in *Proc. of the 20th Int. Conf. on Machine Learning (ICML '03)*, pp. 856–863 (2003).
37. M. Hall et al., "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009).
38. D. D. Lewis, "Naive (Bayes) at forty: the independence assumption in information retrieval," in *European Conf. on Machine Learning*, pp. 4–15 (1998).
39. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).
40. J. R. Quinlan, "Decision trees and decision-making," *IEEE Trans. Syst. Man Cybern.* **20**(2), 339–346 (1990).
41. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967).
42. T. K. Ho, "Random decision forests," in *Proc. of the IEEE Conf. on Document Analysis and Recognition*, pp. 278–282 (1995).

Rahul Paul is pursuing his PhD degree at the University of South Florida, Tampa. He received his MS degree in computer science from Indian Institute of Technology (ISM), Dhanbad, India, in 2015. He has worked to improve the prediction of malignancy of pulmonary nodules from CT screening by utilizing deep features and quantitative CT features from the nodule. His areas of interest include pattern recognition, image processing, deep learning, data mining, and lung cancer.

Samuel H. Hawkins is a postdoctoral fellow at H. Lee Moffitt Cancer Center, Tampa. He received his PhD degree from the University of South Florida and his MS degree from Emory University in 2007 in computer science. He has worked to improve the prediction of malignancy of pulmonary nodules from CT screening by utilizing quantitative CT features from the nodule. His areas of interest include pattern recognition, image processing, and lung cancer.

Matthew B. Schabath is an associate member of the Department of Cancer Epidemiology in H. Lee Moffitt Cancer Center & Research Institute, Florida. He received his PhD degree in epidemiology from the University of Texas in 2003 and has over 18 years of experience studying tobacco-related diseases, including lung cancer, bladder cancer, COPD, and heart disease.

Robert J. Gillies is the chair of Cancer Physiology Department, director of Center of Excellence in Cancer Imaging & Technology, and vice chair of Radiology Research in H. Lee Moffitt Cancer Center & Research Institute, Florida. His work is focused on defining and characterizing deregulated pathways with therapeutic relevance in subsets of human cancers. His colleagues have identified targetable cell surface receptors in pancreatic, melanoma, and breast cancers.

Lawrence O. Hall is a professor in the Computer Science and Engineering Department of University of South Florida, Tampa. He received his PhD degree in computer science from Florida State University. He is a fellow of IEEE and past president of NAFIPS. His research interests lie in distributed machine learning, extreme data mining, bioinformatics, pattern recognition and integrating AI into image processing.

Dmitry B. Goldgof is a professor in the Computer Science and Engineering Department of University of South Florida, Tampa. He received his PhD degree in electrical engineering from the University of Illinois at Urbana-Champaign. He is a fellow of IEEE, IAPR, AAAS, and AIMBE. His research interests are related to biomedical image analysis and machine learning with application in MR, CT, PET, and microscopy images, radiomics and bioinformatics, motion analysis with biometrics, face analysis, and surveillance applications.