Explaining Deep Features Using Radiologist-Defined Semantic Features and Traditional Quantitative Features

Rahul Paul[1], Matthew Schabath[2], Yoganand Balagurunathan[3], Ying Liu[4], Qian Li[4], Robert Gillies[3], Lawrence O. Hall[1], and Dmitry B. Goldgof[1]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL; [2]Department of Cancer Epidemiology, H. L. Moffitt Cancer Center & Research Institute, Tampa, FL; [3]Department of Cancer Imaging and Metabolism, H. L. Moffitt Cancer Center & Research Institute, Tampa, FL; and [4]Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin

Corresponding Author:
Dmitry B. Goldgof, PhD
Department of Computer Science & Engineering, USF College of Engineering, Building II 4220 E. Fowler Avenue, Tampa, FL 33620, USA;
E-mail: goldgof@mail.usf.edu.

Key Words: deep features, radiomics, semantic features, interpretation of features, CNN, explainable AI, quantitative features
Abbreviations: Convolutional neural network (CNN), low-dose computed tomography (LDCT), screen-detected lung cancer (SDLC)

ABSTRACT

Quantitative features are generated from a tumor phenotype by various data characterization, feature-extraction approaches and have been used successfully as a biomarker. These features give us information about a nodule, for example, nodule size, pixel intensity, histogram-based information, and texture information from wavelets or a convolution kernel. Semantic features, on the other hand, can be generated by an experienced radiologist and consist of the common characteristics of a tumor, for example, location of a tumor, fissure, or pleural wall attachment, presence of fibrosis or emphysema, concave cut on nodule surface. These features have been derived for lung nodules by our group. Semantic features have also shown promise in predicting malignancy. Deep features from images are generally extracted from the last layers before the classification layer of a convolutional neural network (CNN). By training with the use of different types of images, the CNN learns to recognize various patterns and textures. But when we extract deep features, there is no specific naming approach for them, other than denoting them by the feature column number (position of a neuron in a hidden layer). In this study, we tried to relate and explain deep features with respect to traditional quantitative features and semantic features. We discovered that 26 deep features from the Vgg-S neural network and 12 deep features from our trained CNN could be explained by semantic or traditional quantitative features. From this, we concluded that those deep features can have a recognizable definition via semantic or quantitative features.

INTRODUCTION

Lung cancer is one of the most common causes of malignancy worldwide, with a 5-year survival rate of 18% (1). The American Cancer Society estimates 14% of new cancer cases will be lung cancer cases for 2018, making it the second most detected cancer in the United States. They also estimate 154,050 deaths from lung cancer, which is the most in the United States in 2018 (2). As lung cancer typically remains undetected during the initial stages, ~75% of patients with lung cancers are first diagnosed at the advanced stages (III/IV) (3). As a result, early detection and diagnosis is a high priority.

Low-dose computed tomography (LDCT) is a noninvasive and widely used imaging technique for detecting lung nodules. By analyzing CT scans, radiologists can generate specific features from one's lung nodule, which could provide guidance for detection and diagnosis. These distinctive features are named semantic features. They can be categorized into the following different groups: shape (eg, lobulation), location (eg, lobe location), margin (eg, spiculation), external (eg, peripheral emphysema). With CT scans, cavitation is discovered in 22% of primary lung cancers and often the cavities in benign nodules mimic the cavities of malignant nodules, which makes precise diagnosis difficult (4). In another study (5), it was found that the risk of lung cancer can be increased 3- to 4-fold owing to emphysema among heavy smokers. Nodule size also influences cancer diagnosis and treatment (6). Hence, semantic features can be used in creating a predictor of lung cancer.

Using CT scans, quantitative information from a lung nodule can be generated and analyzed using statistics, machine learning, or high-dimensional data analysis. This approach is termed radiomics (7). These quantitative features can be categorized into the following different groups: texture (eg, Law's

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00034

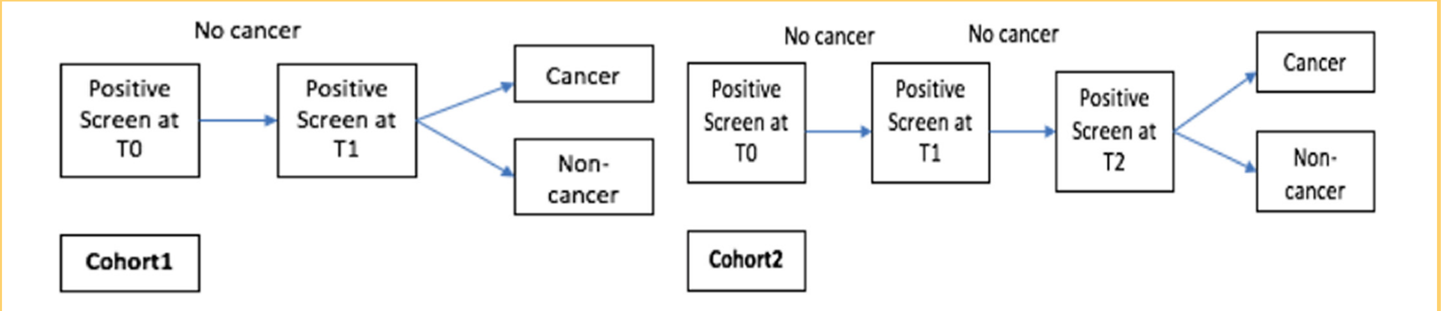

Figure 1. Selection of cohort 1 and cohort 2.

texture features, wavelet features), size (eg, longest diameter, volume), location (eg, attached to the pleural wall, distance from the boundary). These traditional quantitative features can be used to create a biomarker for tumor prognosis, analysis, and prediction (8-10).

Deep learning is an emerging approach mainly applied in recognition-, prediction-, and classification-related tasks. Propagating data through multiple hidden layers will eventually help a neural network to learn and build a representation of data, which can be used further for prediction or classification. For image data, a convolutional neural network (CNN) typically uses several convolutional kernels to extract different textures and edges before propagating the extracted information through multiple hidden layers. For lung nodule analysis, CNNs have been used effectively in recent years (11). In the medical imaging field, data are currently scarce; so, as an alternative to building a new model, transfer learning has been used (12).

Convolution layers of CNNs, after learning, contain representations of edge gradients and textures, and when propagated through fully connected layers, various high-level features are posited to have been learned by the network. From fully connected layers, deep features (the outputs of units in the layer) are extracted and denoted by the number of the feature from the learning tool (the position of a neuron in a hidden layer row vector).

Two pretrained CNNs were used in the work described in this paper for extracting the following deep features: the Vgg-S network (13), which was trained on the ImageNet data set (14) of color camera images and our designed CNN (15), which was trained on lung nodule images. There were 23 traditional quantitative features [RIDER subset features (16)] used in this study along with 20 semantic features, which were generated by an experienced radiologist from Tianjin Medical University Cancer Institute and Hospital, China. This study is an extension of our previous study (17), which analyzes the similarity between deep features and semantic features. In this current study, we also focused on traditional quantitative features, that is, analyzed the similarity of deep feature(s) to traditional quantitative features. The analysis was conducted by replacing ≥1 deep features with traditional quantitative or semantic feature(s). The goal was to show that equivalent classification performance can be achieved. That means those deep features contained information similar to that of the semantic or traditional quantitative features. We can equate those deep features with the name of the corresponding semantic or traditional quantitative feature.

We found that location-based semantic features are difficult to replace, but size-, shape-, and texture-based semantic features can be replaced by deep feature(s). Therefore, shape and texture quantitative features can be used to explain deep feature(s). By "explain," we mean the features can replace deep features and a classifier will achieve the same accuracy. We successfully explained 26 deep features from the Vgg-S network out of 4096 features and 12 deep features from our trained CNN by semantic and traditional quantitative features. This provides a semantic meaning for the deep features.

METHODOLOGY

Data Set

A subset of cases from the LDCT-arm of the NLST (National Lung Screening Trial) data set was chosen for this study. The NLST study was conducted over 3 years: 1 baseline scan (T0) and 2 following scans (T1 and T2) in 2 subsequent years with an interval of ~1 year (18) between scans. For this study, a subset of nodule-positive and screen-detected lung cancer (SDLC) cases (years later) from the baseline (T0) scans were chosen, and the patient data were deidentified under an IRB-approved process. These subsets of cases were further divided into the following 2 categories: cohort 1 and cohort 2. Cohort 1 consisted of cases with a baseline scan (T0), which had a follow-up scan after 1 year (T1), wherein some of the nodules became cancerous. Whereas, cohort 2 consisted of nodules that became cancerous after 2 years (T2 scan) from the baseline scan (T0). Selection of cohorts is shown in Figure 1. Only Cohort 2 (SDLC, 85; positive control cases, 152) was chosen for our study. Between the SDLC and control-positive cases, there is no statistically significant difference with respect to sex, race age, ethnicity, and smoking (19). Nodule segmentation was performed using the Definiens software suite (20). From our initial set of cases, 52 cases were excluded owing to ≥1 of the following reasons: multiple malignant nodules, inability to identify the nodule, or unknown location of the tumor. So, finally, 185 cases (SDLC, 58; control-positive cases, 127) were selected for our study.

Semantic Features

Semantic features were described from the CT scan of a lung tumor, by an experienced radiologist. They can be used further

Table 1. Description of Semantic Features

Characteristic	Definition	Scoring
Location		
1. Lobe Location	Lobe location of the nodule	Left lower lobe (5), left upper lobe (4), right lower lobe (3), right middle lobe (2), right upper lobe (1)
Size		
2. Long-Axis Diameter	Longest diameter of the nodule	NA
3. Short-Axis Diameter	Longest perpendicular diameter of nodule in the same section	NA
Shape		
4. Contour	Roundness of the nodule	1, round; 2, oval; 3, irregular
5. Lobulation	Wavy nodule's surface	1, none; 2, yes
6. Concavity	Concave cut on nodule surface	1, none; 2, slight concavity; 3, deep concavity
Margin		
7. Border Definition	Edge appearance of the nodule	1, well defined; 2, slight poorly; 3, poorly defined
8. Spiculation	Lines radiating from the margins of tumor	1, none; 2 yes
Attenuation		
9. Texture	Solid, non-solid, part solid	1, non-solid; 2, part solid; 3, solid
10. Cavitation	Presence of air in the tumor at the time of diagnosis	0, no; 1, yes
External		
11. Fissure Attachment	Nodule attaches to the fissure	0, no; 1, yes
12. Pleural Attachment	Nodules attaches to the pleura	0, no; 1, yes
13. Vascular Convergence	Convergence of vessels to nodule	0, no significant convergence; 1, significant
14. Pleural Retraction	Retraction of the pleura towards nodule	0, absence of pleural retraction; 1, present
15. Peripheral Emphysema	Peripheral emphysema caused by nodule	1, absence of emphysema; 2, slight present; 3 severely present
16. Peripheral Fibrosis	Peripheral fibrosis caused by nodule	1, absence of fibrosis; 2, slight present; 3 severely present
17. Vessel Attachment	Nodule attachment to blood vessel	0, no; 1, yes
Associated Findings		
18. Nodules in Primary Lobe	Any nodules suspected to be malignant or intermediate	0, no; 1, yes
19. Nodules in Nonprimary Lobe	Any nodules suspected to be malignant or intermediate	0, no; 1, yes
20. Lymphadenopathy	Lymph nodes with short- axis diameter greater than 1 cm	0, no; 1, yes

for diagnosis. An experienced radiologist (Y.L.) with 7 years of experience from Tianjin Medical University Cancer Institute and Hospital, China, described 20 semantic features (21–24) on a subset of cases that intersected Cohort 2. Semantic features can be categorized into the following groups: shape, size, location, margin, external attenuation, and associated findings. These features have been derived with respect to lung nodules by our group. Table 1 shows a detailed description of our semantic features.

Traditional Quantitative Features

Definiens software (20), along with help from a radiologist, was used to segment lung nodules. Then 23 Rider stable features (16) were extracted using Definiens software. Table 2 shows a detailed description of the "traditional" quantitative features.

Deep Features from Vgg-S Network

Nowadays CNNs are used effectively for image classification and prediction (11, 13). A CNN has many layers of convolution kernels along with multiple hidden layers, which makes the network architecture deeper, and features extracted from such a network are called "deep features." In the medical imaging field, there is typically not enough original data available to train a CNN. As a result, transfer learning (12) is an alternative option. Applying previously learned knowledge from 1 domain to a new task domain is called transfer learning. To extract deep features from a CT scan, the 2-dimensional slice, which has the largest nodule area, was chosen for every case. We extracted only the nodule region by incorporating the largest rectangular box around the nodule. Bicubic interpolation was used to resize the nodule images to 224 × 224, which was the required input size of the Vgg-S network. Figure 2 shows a lung image with nodule

Table 2. Description of Rider Stable Traditional Quantitative Features

Characteristic	Features
Size	1. Long-axis diameter
	2. Short-axis diameter
	3. Long-axis diameter × short-axis diameter
	4. Volume (cm)
	5. Volume (pixel)
	6. Number of pixels
	7. Length/width
Pixel Intensity Histogram	8. Mean (HU)
	9. Stand deviation (HU)
Tumor Location	10. 8a_3D_ is attached to pleural wall
	11. 8b_3D Relative border to lung
	12. 8c_3D_Relative border to pleural wall
	13. 9e_3D_Standard deviation_ COG to border
	14. 9g_3D_max_Dist_COG to border
Tumor Shape (Roundness)	15. 9b-3D circularity
	16. 5a_3D- MacSpic
	17. Asymmetry
	18. Roundness
Run-length and Co-occurrence	19. Avg_RLN
Law's Texture Feature	20. E5 E5 L5 layer 1
	21. E5 E5 R5 layer 1
	22. E5 W5 L5 layer 1
	23. L5 W5 L5 layer 1

and the extracted nodule region. The Vgg-S network was trained using natural camera images, which were 3-channel (R, G, B), but the nodule images were grayscale (no color component and voxel intensities of the CT images were converted to 0-255). So, the same grayscale nodule image was used 3 times to mimic an image with 3 color channels and then normalization was performed using the appropriate color channel image. The deep features were generated from the last fully connected layer after applying the ReLU activation function. The size of the feature vector was 4096.

Deep Features from Our Trained CNN

We also experimented by extracting deep features from our designed CNN network (15). Augmented nodule images of Cohort 1 were used to train our CNN architecture. Each nodule image was augmented first by being flipped horizontally and vertically and then all images were rotated by 15°. Keras (25) with a Tensorflow (26) backend was used to train our CNN. We used the same 2-dimensional slice from a nodule for training the CNN and for transfer learning using the Vgg-S network. The input image size for the CNN architecture was 100 × 100 pixels.

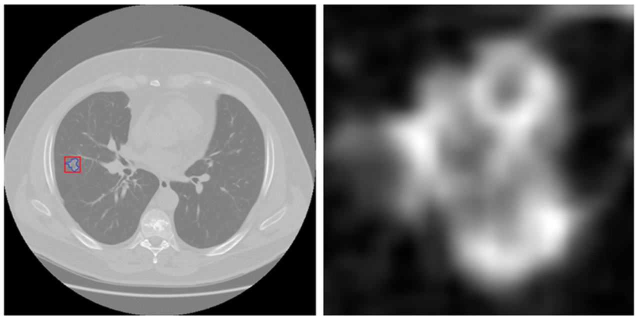

Figure 2. (Left) lung image with nodule inside outlined in blue (nodule pixel size =0.74 mm), with box used for extracted nodule in red, (Right) extracted nodule.

The augmented data set was divided into the following 2 parts: 70% of the data for training and the remaining 30% for validation. The CNN was trained for 100 epochs with 0.0001 learning rate with RMSprop (27) optimization and binary cross-entropy as loss function. A batch size of 16 was chosen for training and validation. L2 regularization (28) along with dropout (29) was used to reduce overfitting of our small and shallow CNN network. Our designed CNN is described in detail in Table 3. The deep features were extracted from the last layer before the

Table 3. Our Designed CNN architecture

Layers	Parameter	Total Parameters
Left branch		39,553
Input Image	100 × 100	
Max Pool 1	10 × 10	
Dropout	0.1	
Right branch		
Input Image	100 × 100	
Conv 1	64 × 5 × 5, pad 0, stride 1	
Leaky ReLU	alpha = 0.01	
Max Pool 2a	3 × 3, pad 0, stride 3	
Conv 2	64 × 2 × 2, pad 0, stride 1	
Leaky ReLU	alpha = 0.01	
Max Pool 2b	3 × 3, pad 0, stride 3	
Dropout	0.1	
Concatenate Left Branch + Right Branch		
Conv 3 + ReLU	64 × 2 × 2, pad 0, stride 1	
Max Pool 3	2 × 2, pad 0, stride 2	
L2 regularizer	0.01	
Dropout	0.1	
Fully Connected 1	1 sigmoid	

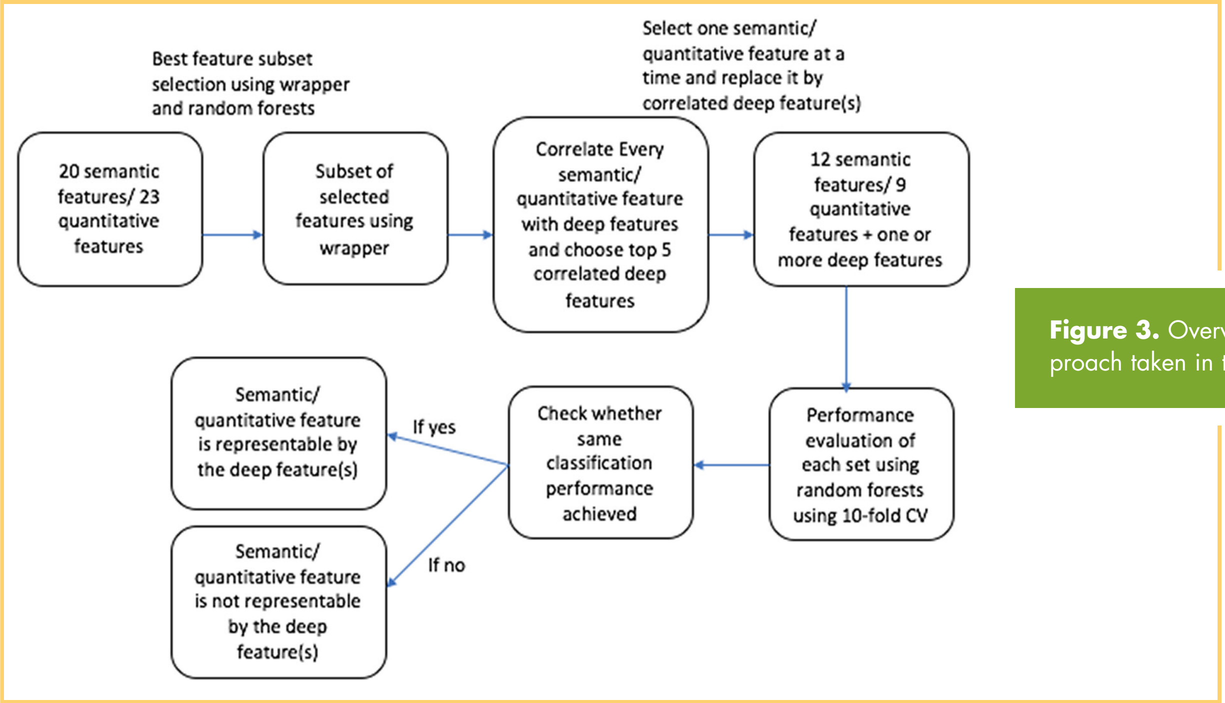
Best feature subset selection using wrapper and random forests
Select one semantic/ quantitative feature at a time and replace it by correlated deep feature(s)
20 semantic features/ 23 quantitative features
Subset of selected features using wrapper
Correlate Every semantic/ quantitative feature with deep features and choose top 5 correlated deep features
12 semantic features/ 9 quantitative features + one or more deep features
Performance evaluation of each set using random forests using 10-fold CV
Check whether same classification performance achieved
If yes
Semantic/ quantitative feature is representable by the deep feature(s)
If no
Semantic/ quantitative feature is not representable by the deep feature(s)

Figure 3. Overview of the approach taken in this study.

classification layer. The size of the feature vector was 1024. After applying the ReLU activation function, some features will be all zeros because ReLU truncates the negative feature values to zero. We removed such features, and as a result, the final number of feature vectors from Vgg-S pretrained CNN and our trained CNN became 3844 and 560, respectively.

Experiments and Results

This section describes the procedure of representing deep feature(s) using semantic or traditional quantitative features.

Wrapper feature selection (30) was applied on traditional quantitative or semantic features of Cohort 2 to select the best subset of features with maximum accuracy. Backward feature selection using the best first strategy and random forests classifier (31) with 200 trees was applied using the wrapper approach. Tenfold cross-validation was used for selecting the best subset of features. We analyzed quantitative features and semantic features separately. A subset of 9 quantitative features was chosen and it enabled a maximum accuracy of 84.32% (AUC 0.87), whereas a subset of 13 semantic features were selected, enabling a maximum accuracy of 83.78% (AUC 0.84). Here, we aim to use semantic features or traditional quantitative features to interpret/explain deep feature(s).

Explaining Deep Features With Respect to Semantic Features

The chosen semantic features (13) were location, long-axis diameter, short-axis diameter, lobulation, concavity, border definition, spiculation, texture, cavitation, vascular convergence, vessel attachment, perinodule fibrosis, and nodules in primary tumor lobe.

After selecting the best subset of semantic features, the correlation coefficient (Pearson correlation coefficient) was calculated for each semantic feature with the deep features, and the 5 most correlated features for each semantic feature were selected. We then replaced each semantic feature with the correlated deep feature(s) and checked whether the same classification accuracy of 83.78% could be achieved.

Our purpose for the study was to determine if semantic features could explain deep features. To do this, we replaced each semantic feature by ≥1 deep features to see if the same classification accuracy could be achieved. We replaced 1 semantic feature at a time from the subset of 13 features and substituted that semantic feature by, at first, the most correlated deep feature and, then 2 most correlated deep features and proceeded similarly to add features until the 5 most correlated deep features had been used as replacements. The accuracy was calculated using a random forests classifier with 200 trees using 10-fold cross-validation. Deep features from Vgg-S pretrained CNN and our trained CNN were examined separately. Figure 3 shows the approach taken for the analysis.

After replacing a feature with deep features extracted from the Vgg-S pretrained CNN, we secured the same original classification accuracy of 83.78% for the following 8 semantic features: long-axis diameter, lobulation, concavity, spiculation, texture, cavitation, vascular convergence, and peripheral fibrosis. Using the deep features acquired from our trained CNN, we achieved the same original classification accuracy of 83.78% for the following 4 semantic features: long-axis diameter, concavity, cavitation, nodules in primary tumor lobe. We found that 3 semantic features (long-axis diameter, concavity, cavitation) could be used to explain both deep features from Vgg-S and our trained CNN. Five semantic features could be used to explain only deep features from Vgg-S, and only 1 semantic feature could be used to explain deep features from our trained CNN. The Vgg-S network was trained on camera images from at least 1000 classes of objects, but not lung nodule images. The large training set helped the network to develop general features and which in turn were explained by texture, spiculation, lobulation, vascular convergence, and peripheral fibrosis. The replacement

Table 4. Classification performance After Features Removal

Features	Feature Names	Accuracy
Semantic Features	Long-axis diameter	82.70 (0.82)
	Lobulation	82.70 (0.83)
	Concavity	83.24 (0.83)
	Spiculation	83.24 (0.83)
	Texture	82.70 (0.83)
	Cavitation	82.70 (0.83)
	Vascular convergence	83.24 (0.84)
	Peripheral fibrosis	82.70 (0.83)
	Nodules in primary lobe	81.62 (0.83)
Traditional Quantitative Features	9b-3D circularity	82.16 (0.86)
	Roundness	82.70 (0.87)
	L5W5L5 layer 1	82.70 (0.87)

These features were from our chosen subset of features, leaving 12 features for training/testing.

of the first 3 and the last feature appear to result from training on lots of images of different types.

Table 4 shows the performance of each semantic feature after removing 1 semantic feature at a time from the subset of 13 features. So, we only calculated classification performance of 12 features at a time using random forests classifier using 10-fold cross-validation, to check whether by removing each feature, there was a change in classification accuracy. In Table 4, we show only the semantic features out of the chosen 13 feature subsets that could be used to explain deep feature(s). Table 5 shows the explainable deep features and their equivalent semantic feature(s). We also show the correlation value of each deep feature with a semantic feature in Table 5.

After replacing semantic features with deep feature(s), similar classification performance was obtained for 9 semantic features. For example, 2 deep features (3353 and 526) from the Vgg-S network could achieve the same classification performance of 83.78% if used in place of cavitation. The deep features 3353 and 526 had the correlation of 0.388 and 0.3551, respectively, with the semantic feature cavitation. Whereas, the deep feature 395 from our trained CNN, which had a correlation coefficient of 0.2748, was explained by cavitation. Similarly, 2 deep features (3353 and 2135) from the Vgg-S network and 1 deep feature (230) using the features from our trained CNN were explained long-axis diameter by providing equivalent performance.

Explaining Deep Features Using Traditional Quantitative Features

The 9 traditional quantitative features that enabled the best accuracy were: Mean (HU), 8a-3D_is_attached to pleural wall, 8c-3D_Relative border to pleural wall, 9b-3D circularity, Asymmetry, Roundness, Volume, E5W5L5, and L5W5L5. The Pearson correlation coefficient was calculated for each traditional quantitative feature with the deep features and the top 5 correlated deep features were selected to replace each traditional quantitative feature. We replaced each traditional quantitative feature by ≥ 1 deep features to try to achieve the same classification accuracy of 84.32%. After replacing deep features extracted from the Vgg-S pretrained CNN, we got the same original classification accuracy of 84.32% for the following 3 traditional quantitative features: 9b-3D circularity, roundness, and L5W5L5 layer 1. Hence, they can be used to explain what the deep features that replaced them have learned. Traditional quantitative features consist of tumor size, tumor shape, Law's texture features, tumor location, etc. As we have seen earlier for semantic features, deep features could be explained by shape-based quantitative features.

In Table 4, we only show the 3 quantitative features that can be replaced (used to explain) deep feature(s). Table 5 shows the quantitative features, their equivalent deep feature(s), and correlations.

DISCUSSION

We showed that some deep features can be explained by a semantic feature or traditional quantitative feature. From a lung nodule CT image, experienced radiologists generated semantic features of different types of information regarding a lung nodule, for example, size, shape, location of nodule, the boundary of the nodule, attachment to the vessel, fibrosis information, etc. These features were shown to provide useful information toward the prognosis and diagnosis of lung cancer. From a tumor phenotype, quantitative information can be extracted using various data characterization approaches, and these features are called traditional quantitative features.

Deep features are extracted from a CNN, generally from the last layer before the final classification layer. For this study, deep features were extracted from the last fully connected layer of the following 2 pretrained CNNs: the Vgg-S network, which was trained on the ImageNet data set, and our designed CNN, which was trained on LDCT lung nodule images. The Vgg-S architecture is a network with 5 convolution layers followed by 3 fully connected layers. Our designed CNN is a small and shallow network with 3 convolution layers and 1 fully connected layer. As the Vgg-S network was trained on a large set of classes of camera images, various textures and other features

Table 5. Semantic and Traditional Quantitative Features and Corresponding Deep Feature(s)

Features	Feature Names	Deep Features from Vgg-S With Correlation Value					Deep Features from Our Trained CNN With Correlation Value			
Semantic Features	Long-axis diameter	3353		2135			230			
		0.4334		0.42			0.3055			
	Lobulation	3534		1372	2975	2111	NA			
		0.5742		0.5614	0.5611	0.5520				
	Concavity	3534	2975	1372	2111	3246	547		440	
		0.5	0.4839	0.4837	0.475	0.4612	0.1776		0.1514	
	Spiculation	2811					NA			
		0.4111								
	Texture	1201		3350			NA			
		−0.3119		0.2936						
	Cavitation	3353		526			395			
		0.3888		0.3551			0.2748			
	Vascular convergence	1464		2115			NA			
		0.7052		0.701						
	Peripheral fibrosis	3305		3064			NA			
		0.2076		0.2043						
	Nodules in primary lobe	NA					425		57	
							0.1871		0.1836	
Traditional Quantitative Features	Roundness	1395		2510			160		20	
		0.3		0.27			0.16		0.13	
	9b-3d circularity	1395		1757	3401	2777	160		20	
		0.24		−0.234	−0.2069	−0.2069	0.14		0.13	
	L5W5L5 layer 1	51	66	163	476	928	547	169	265	309
		0.77	0.75	0.69	0.69	0.69	0.28	0.27	0.26	0.26

were extractable, which can be used effectively for tumor classification. Our trained CNN was trained with LDCT lung nodule images and gave us better performance than transfer learning in our previous study (15).

In this study, we attempted to explain deep features using semantic or traditional quantitative features. A subset of features was chosen from the semantic or traditional quantitative features using a wrapper with a random forests classifier. For the semantic features, the best subset had 13 features with an accuracy of 83.78% (AUC 0.84), whereas from traditional quantitative features, the size of the best subset was 9 features with an accuracy of 84.32% (AUC 0.87). The Pearson correlation coefficient was calculated with each of the chosen semantic features or traditional quantitative features and the deep features. For every semantic or traditional quantitative feature, the top 5 most correlated deep features were chosen. Now, from our chosen subset of semantic or traditional quantitative features, 1 feature was removed, and it was substituted by the most correlated deep feature and classification performance was calculated. With a single substituted deep feature, if we can achieve the classification performance then stop; otherwise, substitute that semantic feature or traditional quantitative feature by the 2 most correlated features and continue this process until the 5 most correlated deep features have been used. In total, 26 deep features from the Vgg-S network and 12 deep features from our trained CNN were explained by 9 semantic features and 3 traditional quantitative features. From this, we hypothesized that those deep features can have a recognizable definition from semantic or quantitative features. That is, those deep features can be given some meaningful definition.

We also trained our CNN on cohort 2 (all 237 cases) and then extracted deep features for only the subset of 185 cases for which semantic features were available. The deep feature vector size was 1024. We removed all zero features to get 699 features from cohort 2. We then used these deep features to represent semantic and quantitative features. We found that some additional semantic features could be used to explain deep features from our CNN trained on cohort 1 (shown in Table 5) in addition to the ones previously found useful. Lobulation, spiculation, vascular convergence, perinodule fibrosis and border definition could explain features from our new deep feature set (CNN trained on cohort 2 data only). Among these semantic features, "border definition" was found to explain 4 deep features (147, 160, 504, and 372) and it could not explain any deep features from Vgg-S or our CNN (trained on cohort 1).

For this study, we extracted only the nodule region from a CT slice. As the nodule region was extracted the information regarding pleural wall attachment, fissure attachment, relative border to the lung, or distance was lost. However, deep features from our trained CNN were explained by only 1 location-based semantic feature (nodules in primary lobe). For training the CNN, we performed data augmentation by rotation and flipping, which enabled the extracted deep features to achieve comparable accuracy. The deep features capture the boundary and shape information quite well because that information could be obtained from the extracted nodule region, and thus, 2 traditional quantitative features (9b-3D-circularity and roundness) and 3 semantic features (lobulation, concavity, and spiculation) were able to explain deep features. Deep features are known to grasp texture-based information as well. As a result, L5W5L5 Law's texture feature and cavitation were useful for explaining deep features. We also found out that deep features 3353, 3534, 1372, 2975, and 2111 from the Vgg-S network were correlated with and explained by >1 semantic features, and feature 1395 was correlated with and explained by 2 traditional quantitative features (roundness and 9b_3D_circularity). Deep features 160 and 20 from our trained CNN network were explained by 2 traditional quantitative features (roundness and 9b_3D_circularity).

In this work, the 5 most correlated features were used to replace a semantic or radiomics feature. Our requirement was some nonzero correlation. Now, with all the comparisons, there will potentially be some spurious correlations. Hence, the Bonferroni correction was used to look at the significance of correlations between deep features and every semantic (or radiomics) feature. As an example, cavitation could be replaced by 2 deep features from the Vgg-S network. Fea 1 (3353) had an original *P* value = 4.8651e-08 and fea 2 (526) had an original *P* value = 7.0822e-07. After the Bonferroni correction, the *P* value of fea 1 was 9.73e-08 and that of fea 2 was 1.4164e-06. Now both Bonferroni-corrected *P*-values were less than the more rigorous significance level. However, when combined, they added more information to our model and hence appear to be associated with cavitation.

After using the Bonferroni correction, we found some of the features with the 5 highest correlation values did not have a significant correlation with a semantic or radiomics feature. Nonetheless, the weakly correlated features were able to explain some CNN features. We interpret this to mean that insignificant, but nonzero, correlations taken together can provide insight into (some) deep features.

In total, 26 deep features from the Vgg-S network and 12 deep features from our trained CNN were explained by 9 semantic features and three traditional quantitative features.

CONCLUSIONS

The recent success of CNNs in various classification-type tasks leads to the question of what they have learned. Here, deep features are explained with respect to semantic features and traditional quantitative features.

In this study, we found explanations for 26 deep features from the Vgg-S network out of 4096 features and 12 deep features from our trained CNN by semantic and traditional quantitative features. One can also look at this as providing semantic information about deep features. Although there has been some research (32-39) regarding semantic understanding of natural scenes using deep CNN features, to our knowledge, this is the first work to explain deep features with respect to traditional quantitative features and semantic features extracted from a lung nodule. In the future, deep features with semantic meaning can be included in biomarkers for tumor prognosis and diagnosis of lung nodules from CT scans, along with semantic features and traditional quantitative features.

There were 2 limitations in our study, first, only 10-fold cross-validation was used to evaluate the performance as we had a limited set of expensive to obtain semantic information. The second limitation of our study was using a single slice for every patient to extract deep features, whereas semantic information was generated from multiple slices. In the future with more semantic annotated data, we will investigate deep features from a 3D CNN.

ACKNOWLEDGMENTS

This research partially supported by the National Institute 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). DARPA's SocialSim program under award number FA8650-17-C-7725.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015 Jan;65:5–29.
2. American Cancer Society. Key Statistics for Lung Cancer. [cited 31 Aug 18] Available from: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
3. Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, Bergström S, Hanna L, Jakobsen E, Kölbeck K, Sundstrøm S. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. Thorax. 2013;68:551–564.
4. Gill RR, Matsusoka S, Hatabu H. Cavities in the lung in oncology patients: imaging overview and differential diagnoses. Appl Radiol. 2010;39:10.
5. Li Y, Swensen SJ, Karabekmez LG, Marks RS, Stoddard SM, Jiang R, Worra JB, Zhang F, Midthun DE, de Andrade M, Song Y. Effect of emphysema on lung cancer risk in smokers: a computed tomography-based assessment. Cancer Prev Res (Phila). 2010;4:43–50.
6. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung AN, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M, Rubin GD. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology. 2017;284:228–243.
7. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563–577.
8. Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based prognosis analysis for non-small cell lung cancer. Sci Rep. 2017;7:46349.
9. Chen CH, Chang CK, Tu CY, Liao WC, Wu BR, Chou KT, Chiou YR, Yang SN, Zhang G, Huang TC. Radiomic features analysis in computed tomography images of lung nodule classification. PloS One. 2018;13:e0192002.
10. Chaddad A, Desrosiers C, Toews M, Abdulkarim B. Predicting survival time of lung cancer patients using radiomic analysis. Oncotarget. 2017;8:104393.
11. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;1097–1105.

12. Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: transfer learning from unlabeled data. Proceedings of the 24th International Conference on Machine Learning. 2007;759–766.
13. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531. 2014 May 14.
14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 2009 Jun 20 (pp. 248–255). IEEE.
15. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging (Bellingham). 2018;5:011021.
16. Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, Goldgof DB, Hall LO, Korn R, Zhao B, Schwartz LH. Test–retest reproducibility analysis of lung CT image features. J Digit Imaging. 2014;27:805–823.
17. Paul R, Liu Y, Li Q, Hall L, Goldgof D, Balagurunathan Y, Schabath M, Gillies R. Representation of Deep Features using Radiologist defined Semantic Features. Proc Int Jt Conf Neural Netw. 2018;1–7.
18. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409.
19. Schabath MB, Massion PP, Thompson ZJ, Eschrich SA, Balagurunathan Y, Goldof D, Aberle DR, Gillies RJ. Differences in patient outcomes of prevalence, interval, and screen-detected lung cancers in the CT arm of the national lung screening trial. PloS One. 2016;11:e0159880.
20. Definiens Developer XD. 2.0. 4 User Guide. Definiens AG, Munich, Germany. 2009.
21. Liu Y, Kim J, Qu F, Liu S, Wang H, Balagurunathan Y, Ye Z, Gillies RJ. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. Radiology. 2016;280:271–280.
22. Li Q, Balagurunathan Y, Liu Y, Qi J, Schabath MB, Ye Z, Gillies RJ. Comparison Between Radiological Semantic Features and Lung-RADS in Predicting Malignancy of Screen-Detected Lung Nodules in the National Lung Screening Trial. Clin Lung Cancer. 2018;19:148–156.
23. Liu Y, Wang H, Li Q, McGettigan MJ, Balagurunathan Y, Garcia AL, Thompson ZJ, Heine JJ, Ye Z, Gillies RJ, Schabath MB. Radiologic features of small pulmonary nodules and lung cancer risk in the National Lung Screening Trial: a nested case-control study. Radiology. 2017;286:298–306.
24. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, Smith G, Massion PP, Schabath MB, Gillies RJ. Radiological image traits predictive of cancer status in pulmonary nodules. Clin Cancer Res. 2016;23:1442–1449.
25. Chollet F. Keras: The python deep learning library. Astrophysics Source Code Library. 2018.
26. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. Tensorflow: a system for large-scale machine learning. OSDI. 2016;16:265–283.
27. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. 2012;4:26–31.
28. Ng AY. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine learning. 2004;Jul 4:78. ACM.
29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014; 15:1929–1958.
30. Kohavi R, Sommerfield D. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. In KDD 1995 Aug 20 (pp. 192–197).
31. Ho TK. Random decision forests. In Document analysis and recognition, 1995, Proceedings of the Third International Conference on 1995 Aug 14 (Vol. 1, pp. 278–282). IEEE.
32. Gudi A. Recognizing semantic features in faces using deep learning. arXiv preprint arXiv:1512.00743. 2015 Dec 2.
33. Ufer N, Ommer B. Deep semantic feature matching. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on 2017 Jul 21 (pp. 5929–5938). IEEE.
34. Aubry M, Russell BC. Understanding deep features with computer-generated imagery. In Proceedings of the IEEE International Conference on Computer Vision 2015 (pp. 2875–2883).
35. Zhao RW, Wu Z, Li J, Jiang YG. Learning semantic feature map for visual content recognition. In Proceedings of the 2017 ACM on Multimedia Conference 2017 Oct 23 (pp. 1291–1299). ACM.
36. Chen C, Wu Z, Jiang YG. Emotion in Context: Deep Semantic Feature Fusion for Video Emotion Recognition. In Proceedings of the 2016 ACM on Multimedia Conference 2016 Oct 1 (pp. 127–131). ACM.
37. Li H, Peng J, Tao C, Chen J, Deng M. What do We Learn by Semantic Scene Understanding for Remote Sensing imagery in CNN framework? arXiv preprint arXiv:1705.07077. 2017 May 19.
38. Lynch C, Aryafar K, Attenberg J. Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016 Aug 13 (pp. 541–548). ACM.
39. Li S, Zhao Z, Liu T, Hu R, Du X. Initializing convolutional filters with semantic features for text classification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing 2017 (pp. 1884–1889).