Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images

Sihong Chen†, Jing Qin, Xing Ji, Baiying Lei, *Member, IEEE*, Tianfu Wang, Dong Ni*, and Jie-Zhi Cheng*†, *Member, IEEE*

***Abstract*— The gap between the computational and semantic features is the one of major factors that bottlenecks the computer-aided diagnosis (CAD) performance from clinical usage. To bridge this gap, we exploit three multi-task learning (MTL) schemes to leverage heterogeneous computational features derived from deep learning models of stacked denoising autoencoder (SDAE) and convolutional neural network (CNN), as well as hand-crafted Haar-like and HoG features, for the description of 9 semantic features for lung nodules in CT images. We regard that there may exist relations among the semantic features of "spiculation", "texture", "margin", etc., that can be explored with the MTL. The Lung Image Database Consortium (LIDC) data is adopted in this study for the rich annotation resources. The LIDC nodules were quantitatively scored w.r.t. 9 semantic features from 12 radiologists of several institutes in U.S.A. By treating each semantic feature as an individual task, the MTL schemes select and map the heterogeneous computational features toward the radiologists' ratings with cross validation evaluation schemes on the randomly selected 2400 nodules from the LIDC dataset. The experimental results suggest that the predicted semantic scores from the three MTL schemes are closer to the radiologists' ratings than the scores from single-task LASSO and elastic net regression methods. The proposed semantic attribute scoring scheme may provide richer quantitative assessments of nodules for better support of diagnostic decision and management. Meanwhile, the capability of the automatic association of medical image contents with the clinical semantic terms by our method may also assist the development of medical search engine.**

***Index Terms*— Computer-aided diagnosis (CAD), lung nodule, computed tomography (CT), multi-task learning, deep learning, feature learning.**

I. Introduction

COMPUTED tomography (CT) is a widely used imaging modality for the assessment of pulmonary nodules. The phenotype features of the pulmonary nodule in CT images are important cues for the malignancy prediction [1]–[2], diagnosis and the further management [3]–[5]. For instance, as illustrated in the diagnostic guidelines from several societies [3]–[6], the high-level texture feature of nodule solidity and the semantic morphology feature of spiculation are crucial for the differentiation of pulmonary nodules. The nodule solidity is also an important cue for the diagnosis of adenocarcinomas and other subtypes [7]–[9]. Meanwhile, other semantic features like the calcification pattern, roundness, margin clearness, etc., are shown to be helpful for the evaluation of nodule malignancy [10] as well. In many previous studies and clinical image diagnosis, these semantic characteristics of pulmonary nodules in CT images are usually defined and rated in a subjective manner. In such cases, the determination of these clinical characteristics may differ from person to person and also depend on the observer's experience [11]–[14]. For example, several studies had suggested there exists perceivable intra- and inter-observer variation for the categorization of nodule solidity in the CT images [11]–[12], [15].

Computer-aided diagnosis (CAD) is an assistive software package to provide computational diagnostic references for the clinical image reading and decision support [16]–[20]. It has been shown to be effective to lower down the inter-observer variation [15], [21]–[22]. Most previous CAD studies focused on the direct differentiation of nodule malignancy and benignancy [18], [20]–[25]. To provide more quantitative diagnostic reference, a recent study [15] explored to categorize nodule solidity with the histogram features for the high-level texture analysis. To characterize the spiculation feature, Ciompi *et al.* [26] developed the bag-of-frequencies descriptor that can successfully distinguished 51 spiculated nodules from the other 204 non-spiculated nodules. In general, the profiling of pulmonary nodules in CT images with attributes of the commonly-used clinical semantic features may provide richer

Manuscript received October 12, 2016; revised November 7, 2016; accepted November 9, 2016. Date of publication November 16, 2016; date of current version March 2, 2017. This work was supported partly by National Natural Science Foundation of China (Nos. 61402296, 61571304, 81571758, 61501305 and 61427806), Shenzhen Key Basic Research Project (Nos. JCYJ20150525092940986, JCYJ20150525092940982, JCYJ20130329105033277 and JCYJ20140509172 609164), the (Key) Project of Department of Education of Guangdong Province (No. 2014GKXM052), Natural Science Foundation of SZU (No. 2016089) and Shenzhen-Hong Kong Innovation Circle Funding Program (No. JSE201109150013A). Sihong Chen and Jie-Zhi Cheng contributed equally. *Asterisks indicate corresponding authors.*

S. Chen, X. Ji, B. Lei, and T. Wang are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Department of Biomedical Engineering, School of Medicine, Shenzhen University, Shenzhen 518060, China.

J. Qin is with the Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.

*J.-Z. Cheng and D. Ni are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Department of Biomedical Engineering, School of Medicine, Shenzhen University, Shenzhen 518060, China (e-mail: nidong@szu.edu.cn; jzcheng@ntu.edu.tw).

Digital Object Identifier 10.1109/TMI.2016.2629462

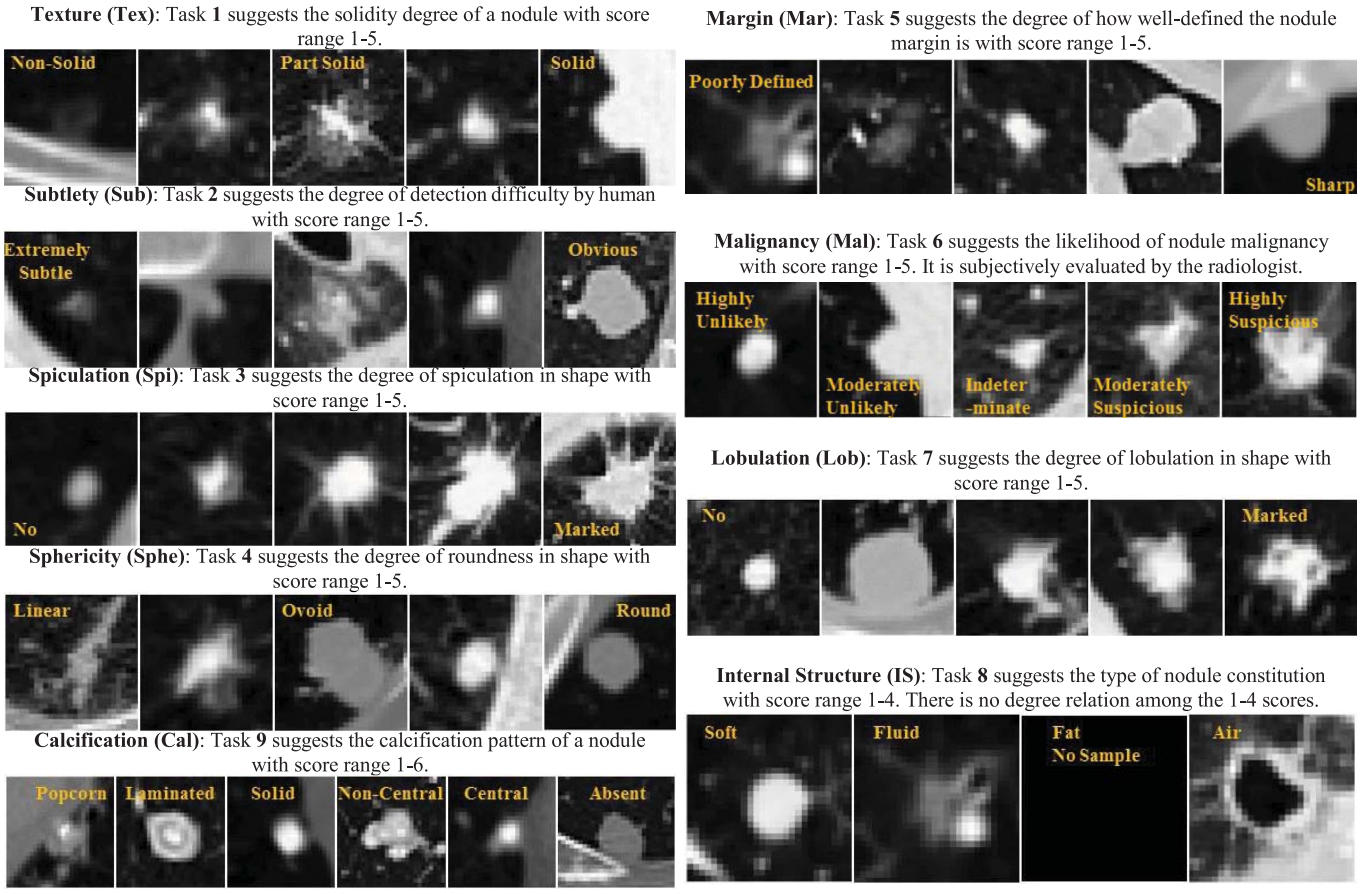

Fig. 1. Nodule patterns with respect to the annotated degrees of the 9 semantic features. The text abbreviations "Tex", "Sub", "Spi", "Sphe", "Mar", "Mal", "Lob", "IS", and "Cal" are "texture", "subtlety", "spiculation", "sphericity", "margin", "malignancy", "lobulation", "internal structure", and "calcification", respectively.

quantitative cues for deeper analysis of pulmonary nodules. However, the mapping from the low-level image features toward the high-level semantic features in the domain of clinical terms is neither straightforward nor a trivial task. It requires intensive elaboration on the design and selection of computational image features. Meanwhile, the designed computational image features for one specific semantic feature may not be generalized to the others easily.

In this study, we aim to develop a new type of CAD scheme that can profile pulmonary nodules with a semantic attribute vector from CT images. Instead of singleton identification of malignancy, this new CAD scheme can yield several quantitative assessment scores w.r.t. the semantic features that may be more referential for clinical usage. Meanwhile, different from [15], [26] that describe the semantic features with binary or trinary values, our CAD scheme is able to rate each semantic feature with wider score range around 1-5 for better quantification analysis. The proposed new computer-aided diagnosis with attribute scoring is denoted as CADa for convenience. To support the training and testing of CADa scheme, we adopt the Lung Image Database Consortium (LIDC) dataset [13]–[14] for its rich annotation resource.

The LIDC dataset collects more than 1,000 thoracic CT scans from 1,010 patients at several medical centers in U.S.A. Each CT scan was reviewed for nodule annotation by four radiologists with two rigorous blinded and unblinded image reading sessions. Totally, there are 12 radiologists [13] from different institutes participating the annotation process. The annotated nodules with diameters larger than 3 mm are further rated by each radiologist w.r.t. the semantic features of "subtlety", "calcification", "sphericity", "margin", "spiculation", "texture", "lobulation", "internal structure", and "malignancy". Except the feature "calcification" that can be rated from score 1 to score 6 and the feature "internal structure" that was scored from 1 to 4, the score range of the remainder features is 1-5. Fig. 1 illustrates the cases w.r.t. each score of the 9 semantic features. Meanwhile, the semantic meanings of the scores are also shown in Fig. 1. The scoring of the first 7 features "Tex", "Sub", "Spi", "Sphe", "Mar", "Mal" and "Lob" hold the degree relation, where the 8^{th} scoring feature "IS" don't. There may exist degree relation for the scoring of the 9^{th} "Cal" feature but not as clear as the scoring for the first 7 features.

Referring to Fig. 1, it can be found the new CADa problem to be approached is quite challenging. The faced challenges can be summarized in threefold. First, our goal is to develop an automatic scoring model to quantify the degree of most semantic features. Therefore, the engineering of effective low-level image features for the scoring of each semantic feature may be more complicated than the feature extraction

problems for the discrete trichotomy/dichotomy of nodules in [15] and [26], respectively. Since useful computational features may vary from semantic features to semantic features, the whole feature engineering and selection process may turn out to be very arduous and tedious. The second challenge consists in the high inter-observer variation on the annotation ratings. Because the nodule annotation was done by many radiologists, there exists definition ambiguity over the degree of the 9 semantic features. In some cases, the definition variation may be significant as the rating process is subjective and may highly depend on rater's experience. The third challenge is the variation of slice thicknesses of the thoracic CT scans in the LIDC dataset. The slice thickness of the LIDC CT scans ranges from 1.25 mm to 3 mm. The different degree of anisotropic resolution between the z and x-y directions may thus impose more challenge on the computation of pure 3D image features as the cues along the z dimension are relatively less reliable.

To address the three challenges faced in the new automatic CADa scheme, we propose to utilize multi-task learning (MTL) framework that leverages heterogeneous computational features derived from the deep learning models of stacked denoising autoencoder (SDAE) [27] and convolutional neural network (CNN) [28], as well as general low-level Haar-like [29] and histogram of oriented gradients (HoG) [30] features, to approach the radiologists' ratings w.r.t. the 9 semantic features.

Since it is generally unknown what kinds of computational features can effectively describe the 9 semantic features, we aim to compute heterogeneous features as diverse as possible to reserve more selection flexibility for the latter MTL models. We employ the deep learning models for the advantage of automatic feature extraction from the training data. The SDAE features are general features because the pre-training of SDAE model doesn't require the labels of training data; the CNN features are more task-specific as the training of CNN needs the specification of data labels. The learnt neuron-crafted features are then pooled with the low-level Haar-like and HoG features.

Because a pulmonary nodule is profiled with the 9 semantic features here, there may exist some sorts of relation among the semantic features that can be further explored with the MTL. The MTL is a co-training framework to exploit the sharable knowledge across the involved tasks. We here specifically apply the MTL to select the effective features from the computational feature pool by treating each semantic feature as an individual task. The MTL is here cast as a regression framework to seek the best mapping between the selected computational features and the rated scores. The regression framework may accommodate the scoring ambiguity between the consecutive scores by relaxing the output scores from the annotated integer level to the estimated floating point level. We explore two MTL schemes: 1) multi-task linear regression [31] and 2) random forest regression [32] with auto-context boosting [33] across 9 tasks.

Except the "Cal" and "IS" tasks, the rating scores of the first 7 semantic tasks suggest the sequential degrees; see Fig. 1. Although some score levels of the "Cal" may have sequential relation, they are not as clear as the scoring defined in the first 7 tasks. Therefore, for the learning of the 2 exceptional "Cal" and "IS" tasks, it may be more suitable to be formulated as classification. Accordingly, we explore a third composite learning scheme to perform the regression for the first 7 tasks and classification for the last two 2 exceptional tasks in a joint-learning fashion.

To bypass the issue of slice thickness variation in the CT images, a slice-based scheme is implemented. Specifically, the training of the regression model is carried out with the unit of 2D slice ROIs. For the testing phase, the final score of a pulmonary nodule can be reached by averaging the scores derived from the member-slice ROIs.

The contributions of this work can be summarized in threefold. First, a new CADa scheme is proposed here to provide quantitative assessments over the 9 semantic features for a pulmonary nodule depicted in the CT images. Different from previous works that only focused on single semantic feature, broader studies on the 9 semantic features are performed in this paper. The yielded 9 semantic rating scores by our CADa scheme may support deeper analysis for a pulmonary nodule for either clinical diagnosis or educational purpose. Meanwhile, it may also help to thrust the CAD scheme toward the clinical usage closer with the quantitative implementation of diagnostic guidelines and recommendations [3]–[4], [6]–[7]. Second, a MTL scheme is proposed to effectively bridge the gap between the computational features and multiple clinical semantic features. The association between the medical image cues and clinical semantic terms may enable sophisticated retrievals of clinical reports and images [34]–[35] from medical databases for better diagnostic decision support. The semantic mapping may also help to correlate the medical image and clinical reports/documents as well as the histological and genomics data with the bridge of semantic features/terms. The correlation among multi-modal medical data may help to design more precise/personalized nodule treatment plan [36]. The third contribution lies the novelty of effective synergy between the heterogeneous computational features and the MTL schemes. The heterogeneous features include deep learning SDAE and CNN features as well as low-level Harr-like and HoG features. It will be shown that the semantic scores predicted from the synergetic schemes are more approaching to the radiologists' ratings than the scores from the single-task learning schemes with the extensive experimental evaluation. It thus corroborates the effectiveness of cross semantic task relation and usage of heterogeneous computational features. Such a synergetic scheme has been less exploited before. This paper is a significant extension of [37] with three major aspects. First, we exploit more MTL schemes and augment the feature pool with the HoG features for more thorough study. Second, larger number of 2400 nodules are involved in this study. Meanwhile, we also include the "Mal" semantic feature here for complete investigation. Third, more comprehensive experiments are performed in this study. Specifically, we elaborate the issues of different ROI settings, selection of rating instances from different radiologists of the same nodule, different implementations of

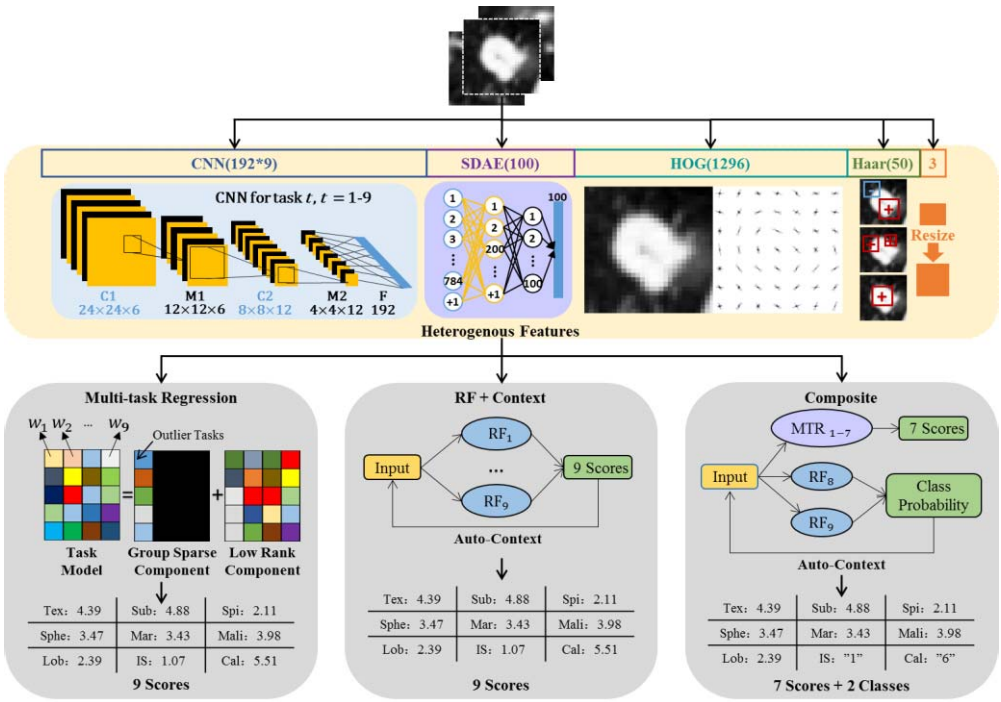

Fig. 2. Flowchart of the proposed computer-aided diagnostic attributing (CADa) scheme.

CNN, and so on. With the three aspects of extension, the efficacy of the MTL schemes on the CADa problem can be better corroborated.

II. Related Works

To our best knowledge, there are less related studies that have conducted thorough exploration for the semantic mapping. An earlier work [38] proposed to use LASSO linear regression [39] to automatically profile liver lesions with semantic terms from the extension of RadLex [40], which is a standard ontology defined by RSNA. The LASSO regression was used to map the computational features to each semantic term in [38]. Depeursinge *et al.* [41] further improved the work [38] significantly. In [41], the Riesz-wavelet components were utilized as computational features with support vector machines (SVMs) to predict the semantic terms of liver lesions in CT scans. In [41], each semantic term was learned with an independent SVM without the exploration of MTL. 74 liver lesions were involved in [41]. In contrast, we focus on the analysis of lung nodules.

The works [15], [26] were formulated as nice 3- and 2-class differentiation schemes for the "Tex" and "Spi" tasks. Both works were realized with intermediate steps. [15] needed nodule segmentation to obtain histogram features. The computation of bag-of-frequencies [26] descriptor involves steps of nodule diameter estimation, profile sampling, codebook creation, etc. Particularly, diameter estimation seems to be crucial to exclude non-nodule profiles, which may affect the effectiveness of descriptor. The estimation scheme was based on nodule contrast. The segmentation of subtle/non-solid nodules as well as nodules with ill-defined margins and complex shapes is difficult. For nodules with complicated contexts, e.g., vessel, chest wall, etc., and low contrast to background, the diameter estimation may have issue. 150 nodules were involved in [15], and [26] used partial data (255/800 nodules), where 51 had spiculation, to identify spiculated and non-spiculated nodules.

Our MTL schemes approach the "Tex" and "Spi" tasks with regression framework for better modeling of the degree relation in the scoring. Meanwhile, complicated intermediate steps are not needed in our MTL schemes and hence potentially unreliable intermediate results from difficult nodules can be avoided. It will be shown that the MTL schemes can achieve reasonably good performance for the "Tex" and "Spi" tasks with the 2400 nodules from the LIDC database.

III. Methods

Referring to Fig. 1, the semantic features cover the high-level description about the shape and appearance of pulmonary nodules, and the effective computational features for each semantic feature (task) is generally unclear. Therefore, we aim to compute heterogeneous features as many as possible. This may enable the latter MTL schemes explore the inter-task (semantic feature) relation and retrieve sharable computational features across the 9 semantic features. We exploit two regression frameworks, i.e., multi-task linear regression and multi-task random forest regression, and one composite learning scheme to fulfill the CADa scheme for pulmonary nodules in CT images. The heterogeneous computational features include general SDAE, Haar-like and HoG features as well as the task-specific CNN features. The flowchart of the proposed method is illustrated in Fig. 2.

The feature computing and the training & testing of the MTL schemes is based on 2D nodule ROIs. We avoid the direct 3D feature computing from the lung CT data with

anisotropic resolution between x-y and z directions, because of the high slice thickness variation (1.25-3mm) in the LIDC dataset. At the testing of the three MTL schemes, the predicted score for a nodule is derived with the averaged scores over all its member slices. Each nodule ROI is defined as the bilaterally expanding bounding boxes of radiologists' outlines with given offset setting in the vertical and horizontal directions to include more anatomical contexts. For training and testing, all ROIs are resized as 28×28 for the convenience of SDAE and CNN models. Since the mean size over all original ROIs is around $(34.4 \pm 9.3) \times (34.3 \pm 9.4)$, the resized ROIs do not deviate significantly from their original image dimensionality. The scaled factors in x and y dimensions and aspect ratio are also included in the feature pool to preserve the original dimensionality features; see Fig. 2. The hyperparameters of the SDAE and CNN models used in this study can be found in the Tables A1 and A2 of Supplement, respectively.

A. Heterogeneous Feature Computation

1) Neuron-Crafted Features: The stacked autoencoder (SAE) is a deep learning model that is realized with two phases of unsupervised and supervised training. In the unsupervised training, the SAE architecture is built by greedily stacking up the encoder of a two-layer autoencoder, which is composed of one input layer and one hidden layer. An autoencoder is consisted of encoder and decoder and can be constructed by seeking the appropriate encoder and decoder for the minimization of reconstruction error. At the unsupervised training, the spatial features can be automatically discovered and encoded in the neurons of hidden layer from the given training data. The SDAE is an augmented version of SAE with the random corruption of input data in the training process of autoencoders for better generalization and noise tolerance.

Our goal is to use SDAE model for the feature computing. Only unsupervised training of SDAE is performed without the need of data labels. Therefore, the SDAE features shall be general to all tasks. We construct a three-layer SDAE architecture with 200 and 100 neurons at the first and second hidden layers, respectively; see Fig. 2. Accordingly, the dimensionality of SDAE features will be 100.

A typical CNN model can be constituted of convolutional (C), max-pooling (M), fully-connected (F), and soft-max layers. The C- and M-layers are generally coupled. The C-layers encode the local spatial features of the previous layer with learnable kernels. An M-layer is usually connected to C-layer for down sampling and can also equip the CNN model with the robustness against translation effect. The fully-connected layer is the traditional multi-layer perceptron network to explore the nonlinear interaction of the neurons from previous layers.

The training of a CNN model is usually realized with supervised fashion. During the training process of a CNN model, the learnable kernels of the C-layers at various levels can gradually capture the spatial features via adjusting the synaptic weightings of neurons. The CNN training is commonly optimized by the stochastic gradient descent approach with the fashion of mini-batch training. 9 CNN models are implemented here for the 9 semantic features to obtain the task-specific computational features with independent training. For each CNN, the training is cast as a classification problem by treating each descriptive an individual class. After training, the soft-max layer of each CNN is removed. The architecture configuration of the 9 CNN models without soft-max layer can be found in Fig. 2. The 192 neurons in the fully-connected layer are treated as the CNN features, and hence the overall number of CNN features in this study is $192 \times 9 = 1728$. The task-specific CNN features of one task may be helpful for other semantic tasks, and the sharable relation will be explored in the feature selection step.

2) Hand-Crafted Features: The hand-crafted Haar-like and HoG features are computed here to expand the diversity of computational features. The Haar-like features aim to characterize the low-level image appearance and context cues with the simple block-wise computation. It has been shown to be effective on many medical image analysis problems like landmark detection [29], etc. Specifically, the appearance and context cues of nodules are quantified with the single- and paired-block Haar-like features, respectively. The computation of the Haar-like features for a 2D nodule CT ROI I_R can be mathematically expressed as

$$f_{haar}\left(I_R|c_1, s_1, c_2, s_2, \epsilon\right) = \frac{1}{(2s_1+1)^2} \sum_{\|p-c_1\| \le s_1} I_R(p) + \frac{\epsilon}{(2s_2+1)^2} \sum_{\|q-c_2\| \le s_2} I_R(q), \tag{1}$$

where c_1, c_2 and s_1, s_2 are the centers and half-sizes of the square positive and negative blocks in the I_R, respectively, and ϵ can be -1, 0 or 1. The parameter ϵ can specify the Haar-like feature type of single-block (0) or paired-block (1 and -1). 50 Haar-like features are computed with the random specifications of parameter ϵ, the block centers, and the block half-sizes. The optional setting of the block half-size is 1, 2, and 3. Given that our ROIs are resized into 28×28, the 50 Haar-like features may be adequate to characterize the nodule ROI. The block sampling settings of the 50 features are randomly set once and then fixed for all ROIs in the training and testing.

HoG is a kind of low-level descriptor to characterize the shape of the object of interest with local gradient orientation histograms. The HoG descriptor may help to characterize several shape semantic features. For the computation of HoG descriptor, each nodule resized 28×28 ROI is decomposed into 6×6 overlapping blocks where each block are constituted of 2×2 cells with stride size of 4 pixels. A cell is defined as a 4×4 region and a gradient orientation histogram with 9 bins can be computed from each cell. There are totally 1296 HoG features computed as the low-level local shape features.

B. Joint Feature Learning Across Different Tasks

Some of the 9 semantic tasks may relate to each other as the perceptual rating process of a radiologist may involve cross referencing over several semantic tasks in his/her mind to give the rated scores for a nodule. Since some 9 semantic features are high-level shape and appearance features, relevance

within/between the high-level shape/appearance features may exist. In the viewpoint of image processing, the 9 semantic tasks may also affect each other, see Fig. 1, and hence impose more difficulty on the job of feature computing. The design of computational features for one semantic task may need to consider the effect of other tasks, and in the meantime maximize the within-task discriminative capability. However, the relation among the 9 tasks is generally unknown, and some semantic tasks may share some computational features with each other, whereas some other tasks may not.

To uncover the inter-task relation and find suitable features for each task, we explore two regression schemes that can jointly consider the interaction among tasks and select useful features w.r.t. each task. We firstly employ a multi-task linear regression scheme that encodes the relations of sharable features across different tasks with low rank structure and excludes the irrelevant tasks from feature sharing with the group-sparse structure. The second regression scheme is based on the random forest and auto-context techniques. Specifically, the random forest method is adopted as the regressor that performs nonlinearly mapping from the heterogeneous computational features to the scorings of each task, whereas the auto-context scheme attempts to fuse the regressed results across the 9 tasks as inter-task relation cues to boost the regression performance. The third composite scheme combines the multi-task learning regression, random forest and auto-context techniques to achieve the regression for 7 tasks and classification for 2 tasks. The two regression and the third composite schemes will be elaborated as follows.

1) Multi-Task Linear Regression: The multi-task regression is to jointly seek the regressors that can correctly predict the semantic scores of the 9 tasks with proper coefficients of computational features. We denote the set of SDAE, CNN, Haar-like, and HoG features of a training sample i w.r.t. the task t as x_i, and its annotated score as y_i^t, for any $x_i \in R^{d\times 1}$ and $y_i^t \in \{1, \cdots, N^t\}$, where N^t is the maximal score of the task t and d is the dimensionality of computational features. The multi-task regression can be sought with the minimization of the cost function

$$\sum_i^n \sum_{t=1}^9 \frac{1}{9\times n}((l_t + s_t)^T x_i - y_i^t)^2 + \lambda_L \|L\|_* + \lambda_S \|S\|_{1,2}, \quad (2)$$

where $L = [l_1, \cdots, l_9]$, $S = [s_1, \cdots, s_9]$, $l_t \in R^{d\times 1}$, $s_t \in R^{d\times 1}$, n is the total sample number, and the terms $\|\cdot\|_*$ and $\|\cdot\|_{1,2}$ are the trace norm and $l_{1,2}$-norm, respectively. The columns of the matrices of L and S specify the coefficients of computational features of each task. The trace norm regularization of the second term, $\|L\|_*$ in the equation (3) prompts to find of low rank structure in the matrix L [31], whereas the $l_{1,2}$-norm of the third regularization term, $\|S\|_{1,2}$, encourages the group-sparse structure of the matrix S. The $l_{1,2}$-norm of S can be sought as $\sum_t^9 \|s_t\|_2$, where $\|s_t\|_2$ is the l_2-norm of column vector. Accordingly, the group-sparse structure shall suggest many zero columns in a matrix. λ_L and λ_S are the importance weightings for the trace norm and $l_{1,2}$-norm, respectively.

In the multi-task regression context, the sought low rank matrix L encodes sharable computational features across the related tasks, while the group-sparse matrix S identifies those non-zero columns as irrelevant tasks. Since the second term, $\|L\|_*$ in the equation (3) constrains the number of selected computational features, the low rank structure of L may select a few effective features across tasks and adjust the coefficients of these features w.r.t. each task. The minimization of the equation (3) is realized by seeking proper matrices of L and S with the accelerated proximal gradient descent method [42].

2) Forest Regression With Auto-Context Fusion Across Tasks: Random forest is a machine learning technique that ensembles the classification/regression results from a multitude of decision trees, where each tree is constructed with the bagging, i.e., bootstrap aggregating, of training data and features [32]. For the regression purpose with random forest, a final regressed value of a training sample i w.r.t. the semantic feature/task t, denoted as $\Upsilon^t(y^t|x_i)$, can be simply achieved by averaging over the predicted values from all decision trees as:

$$\Upsilon^t\left(y_i^t|x_i\right) = \frac{1}{M_t}\sum\nolimits_{k=1}^{M_t} \gamma_k^t\left(y_i^t|x_i\right), \quad (3)$$

where γ_k^t is the regressed value of a decision tree k and M_t is the total number of decision trees of task t, respectively. The variables x_i and y_i^t are the set of computational features and the task t annotated score of the sample i, respectively. The random forest regression is expected to approximate the regressed value $\Upsilon^t(y_i^t|x_i)$ to the annotated score y_i^t as close as possible for all training samples.

The random forest regression for individual task can be attained with the equation (4). To investigate if the inter-task relation can further boost the regression performance, we implement the auto-context scheme to iteratively fuse the regression results from other tasks into the random forest regression of each task. The auto-context is a recursive boosting method that concatenates the classification/regression results of previous iteration with the input data as augmented input data for the training of the classifier/regressor at current iteration. The optimal performance boosting can be usually achieved by taking 2-3 recursive iterations [33]. More recursive iteration may not be helpful. The concept of auto-context is quite simple but has been shown to be effective on performance improvement in many image segmentation applications.

In this study, the auto-context scheme is incorporated to augment the feature space with the regressed values from previous recursive iteration in the training of random forest, to see if better regression performance can be attained. Specifically, given the random forest regression value Υ_{j-1}^t for the task t at the recursive iteration $j-1$, the augmented features for the jth iteration of random forest training can be expressed as $\{x_i, \{\Upsilon_{j-1}^m\}|\exists m, 1 \le m \le 9, m \ne t\}$. The fusing of regression results over previous iteration of current and other tasks with the auto-context will be shown to be helpful in boosting the performance of random forest w.r.t. each task.

3) Composite of Regression and Classification: To address the issue of non-sequential relation on the scores

TABLE I

ABSOLUTE DISTANCE PERFORMANCE. THE TASK ABBREVIATIONS ARE THE SAME AS DEFINED IN FIG. 1. METHOD ABBREVIATIONS "IB", "R", "C", "RF", "LS", "EN", "COT_CNN" AND "STL_CNN" INDICATE THE INTER-OBSERVER VARIATION, MULTI-TASK LINEAR REGRESSION, COMPOSITE LEARNING, RANDOM FOREST, LASSO, ELASTIC NET, CO-TRAINING AND SINGLE-TASK CNN SCHEMES, RESPECTIVELY

		Task / M	Tex	Sub	Spi	Sphe	Mar	Mal	Lob	IS	Cal
		IB	**0.52±0.89**	**0.93±0.88**	**0.64±0.87**	**0.86±0.77**	**0.92±0.91**	**0.87±0.84**	**0.80±0.92**	**0.02±0.23**	**0.18±0.67**
Offset Setting (10-fold)	ROI1	R	0.53±0.62	0.70±0.62	0.85±0.73	0.80±0.57	0.81±0.64	0.83±0.62	0.88±0.72	0.05±0.24	0.31±0.57
		C(L2)	0.58±0.71	0.75±0.67	0.80±0.70	0.80±0.54	0.82±0.66	0.84±0.60	0.84±0.66	0.01	0.09
		RF(L2)	0.70±0.68	0.82±0.58	0.91±0.69	0.82±0.50	0.92±0.64	0.98±0.64	0.93±0.65	0.03±0.20	0.53±0.66
		LS	1.07±0.62	1.23±0.89	0.87±0.96	0.92±0.70	1.33±0.74	1.13±0.73	0.94±0.90	0.02±0.19	0.47±0.51
		EN	1.31±0.73	0.88±0.70	0.88±1.03	0.84±0.64	0.99±0.72	1.30±0.92	1.01±0.95	0.04±0.26	0.59±0.62
	ROI2	R	**0.55±0.65**	**0.72±0.66**	**0.84±0.71**	**0.79±0.54**	**0.82±0.65**	**0.84±0.60**	**0.89±0.70**	**0.04±0.19**	**0.35±0.58**
		C(L2)	0.62±0.72	0.75±0.65	0.79±0.67	0.81±0.53	0.83±0.66	0.88±0.60	0.83±0.65	0.01	0.08
		RF(L2)	0.74±0.69	0.82±0.58	0.91±0.70	0.83±0.51	0.94±0.66	0.99±0.65	0.93±0.65	0.02±0.20	0.53±0.67
		LS	1.13±0.63	1.19±0.83	0.84±0.96	0.93±0.71	1.14±0.70	1.13±0.72	0.92±0.90	0.02±0.19	0.47±0.52
		EN	1.13±0.68	0.89±0.67	0.88±0.65	0.84±0.60	0.92±0.68	1.21±0.88	0.94±0.60	0.02±0.19	0.56±0.68
	ROI3	R	0.56±0.66	0.74±0.65	0.84±0.73	0.83±0.57	0.85±0.68	0.89±0.66	0.87±0.70	0.05±0.19	0.38±0.64
		C(L2)	0.64±0.68	0.75±0.64	0.78±0.68	0.80±0.51	0.84±0.65	0.88±0.61	0.83±0.67	0.01	0.10
		RF(L2)	0.77±0.71	0.83±0.59	0.92±0.69	0.84±0.50	0.96±0.67	1.00±0.67	0.94±0.66	0.03±0.20	0.55±0.70
		LS	1.06±0.60	0.95±0.77	0.84±0.95	0.95±0.71	1.05±0.65	1.10±0.70	0.93±0.90	0.02±0.19	0.55±0.53
		EN	1.10±0.73	0.84±0.69	0.89±0.65	0.89±0.63	0.93±0.72	1.08±0.77	0.92±0.61	0.02±0.19	0.53±0.67
Average (10-fold)		R	0.53±0.65	0.65±0.62	0.70±0.63	0.67±0.51	0.72±0.60	0.70±0.55	0.76±0.64	0.06±0.18	0.33±0.52
		C(L2)	0.60±0.70	0.67±0.60	0.64±0.59	0.65±0.46	0.73±0.59	0.69±0.54	0.68±0.57	0.01	0.09
		RF(L2)	0.74±0.71	0.74±0.56	0.79±0.61	0.69±0.46	0.87±0.62	0.83±0.60	0.80±0.58	0.02±0.17	0.51±0.62
		LS	1.12±0.67	1.02±0.78	0.75±0.87	0.86±0.60	1.02±0.65	0.99±0.66	0.80±0.82	0.02±0.17	0.47±0.47
		EN	1.18±0.75	0.86±0.65	0.77±0.58	0.68±0.52	0.81±0.60	1.07±0.77	0.78±0.54	0.02±0.17	0.50±0.56
Leave-one-out (LOO)		R	0.61±0.63	0.81±0.63	0.69±0.62	0.75±0.50	0.80±0.61	0.74±0.53	0.75±0.61	0.03±0.18	0.38±0.52
		C(L2)	0.64±0.65	0.81±0.63	0.69±0.62	0.77±0.49	0.81±0.61	0.75±0.53	0.75±0.59	0.01	0.08
		RF(L2)	0.80±0.73	0.87±0.66	0.83±0.60	0.81±0.50	0.96±0.65	0.88±0.66	0.86±0.58	0.03±0.18	0.58±0.70
		COT_cnn	0.80±0.71	0.88±0.68	0.75±0.58	0.82±0.51	0.93±0.64	0.80±0.57	0.78±0.56	0.06±0.18	0.54±0.60
		LS	1.22±0.69	1.26±0.97	0.82±0.62	1.02±0.76	1.21±0.71	1.01±0.69	0.89±0.58	0.04±0.18	0.53±0.50
		EN	1.32±0.72	0.93±0.64	0.78±0.59	0.81±0.55	0.92±0.64	1.20±0.89	0.87±0.54	0.01±0.18	0.66±0.56
		STL_cnn	1.22±0.51	0.96±0.69	0.77±0.63	0.87±0.56	1.01±0.64	1.20±0.53	0.80±0.59	0.03±0.20	0.64±0.57
		Regression	1.23±0.50	1.21±0.69	0.87±0.85	0.91±0.63	1.18±0.66	1.08±0.74	0.84±0.86	0.04±0.18	0.60±0.44
Independent Data (100 nodules)		R	0.57±0.68	0.90±0.68	0.71±0.64	0.77±0.53	0.83±0.62	0.86±0.69	0.82±0.68	0.05±0.22	0.39±0.68
		C(L2)	0.55±0.68	0.88±0.67	0.74±0.63	0.77±0.48	0.83±0.59	0.80±0.63	0.86±0.67	0.01	0.2
		RF(L2)	0.75±0.66	0.89±0.64	0.84±0.59	0.83±0.50	0.99±0.63	0.92±0.70	0.94±0.64	0.05±0.22	0.57±0.83
		COT_cnn	0.73±0.59	0.89±0.57	0.75±0.62	0.86±0.52	0.90±0.62	0.78±0.58	0.85±0.64	0.10±0.23	0.51±0.55
		LS	1.16±0.56	1.22±0.77	0.91±0.68	1.04±0.67	1.14±0.73	1.59±0.87	1.10±0.72	0.26±0.29	0.59±0.76
		EN	1.23±0.64	0.99±0.62	0.80±0.60	0.84±0.53	1.02±0.58	1.01±0.82	0.94±0.60	0.02±0.23	0.64±0.54
		STL_cnn	1.15±1.01	0.94±0.63	0.75±0.67	0.93±0.62	0.98±0.62	0.93±0.66	0.86±0.72	0.04±0.22	0.70±0.61
		Regression	1.18±0.63	1.51±0.96	1.07±0.93	0.98±0.64	1.38±0.82	1.56±1.13	1.21±1.07	0.23±0.22	0.67±0.65

of "Cal" and "IS" tasks, a composite scheme is developed. The first 7 tasks are firstly trained with the multi-task linear regression scheme, whereas the 2 exceptional tasks are independently trained with random forest classification. The inter-task relation among all 9 tasks are further exploited with the auto-context fusing. In each recursion of auto-context fusion, we remain perform regression for the 7 degree tasks and classification for the 2 exceptional tasks. The workflow of the composite scheme can be referred to Fig. 2.

IV. EXPERIMENTS

2400 pulmonary nodules are randomly selected from the LIDC dataset for the testing & training of the learning methods and feature computation. Both 10-fold and leave-one-out (LOO) cross validation (CV) schemes are implemented to

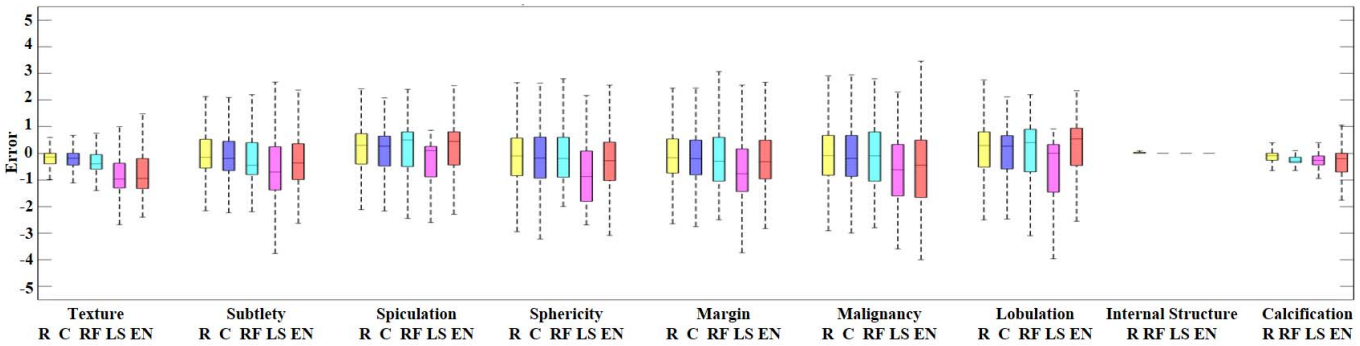

Fig. 3. Box-plots of signed distances from the predicted scores of learning methods using all features to the radiologists' scores. The method abbreviations are the same as the abbreviations defined in Table I.

TABLE II
CNN PERFORMANCE ANALYSIS ON DEPTH. THE TASK ABBREVIATIONS ARE THE SAME AS DEFINED IN FIG. 1. METHOD ABBREVIATIONS SHARE THE SAME DEFINITION IN TABLE I. THE COLUMN "D" SUGGESTS NUMBER OF WEIGHT LAYERS IN THE CNN

	D	Tex	Sub	Spi	Sphe	Mar	Mal	Lob	IS	Cal
R	4	0.64±0.67	0.74±0.59	0.86±0.66	0.84±0.49	0.88±0.64	0.90±0.59	0.88±0.64	0.04±0.19	0.52±0.63
	6	0.83±0.76	0.84±0.65	0.90±0.67	0.85±0.49	1.00±0.69	0.99±0.69	0.90±0.65	0.04±0.20	0.56±0.72
	10	0.84±0.76	0.84±0.65	0.90±0.67	0.85±0.49	1.01±0.69	0.99±0.69	0.90±0.65	0.04±0.20	0.57±0.72
C(L2)	4	0.64±0.66	0.73±0.58	0.86±0.67	0.84±0.50	0.86±0.63	0.91±0.60	0.88±0.64	0.01	0.08
	6	0.84±0.75	0.88±0.66	0.94±0.68	0.86±0.49	1.02±0.68	1.06±0.71	0.93±0.67	0.01	0.13
	10	0.85±0.75	0.89±0.67	0.95±0.68	0.86±0.49	1.02±0.68	1.06±0.71	0.93±0.67	0.01	0.12
RF(L2)	4	0.81±0.70	0.88±0.63	0.98±0.69	0.85±0.53	1.02±0.72	1.05±0.71	0.97±0.71	0.03±0.20	0.54±0.73
	6	0.85±0.76	0.91±0.66	0.96±0.69	0.86±0.51	1.04±0.70	1.10±0.74	0.95±0.67	0.04±0.21	0.60±0.72
	10	0.85±0.75	0.91±0.66	0.97±0.69	0.87±0.51	1.03±0.71	1.10±0.75	0.96±0.68	0.04±0.21	0.60±0.71
LS	4	1.23±0.56	1.02±0.64	0.98±0.99	1.09±0.77	1.25±0.74	1.23±0.80	0.95±0.84	0.03±0.19	1.12±0.39
	6	1.22±0.51	1.45±0.67	0.87±0.94	0.89±0.68	1.22±0.66	1.18±0.76	0.92±0.87	0.02±0.20	0.65±0.65
	10	1.15±0.49	1.07±0.58	0.88±0.94	0.89±0.68	1.28±0.67	1.18±0.77	0.93±0.87	0.02±0.20	0.65±0.65
EN	4	1.28±0.55	1.08±1.07	1.03±1.09	1.14±0.80	1.17±1.04	1.24±0.84	0.98±0.89	0.03±0.19	1.21±0.85
	6	1.88±0.56	1.18±0.58	0.83±1.17	1.00±0.74	2.10±1.02	1.73±1.12	0.91±1.15	0.33±0.16	1.52±0.20
	10	1.51±0.48	1.44±0.71	0.83±1.17	0.97±0.59	1.71±0.86	1.34±0.94	0.91±1.15	0.35±0.16	1.63±0.22

show the robustness of data dependence. In each fold of CV, the computation of SDAE, CNN, Haar-like and HoG features and the learning methods share the same data partition of training and testing. The data partition of the two CV schemes is based on the unit of nodule [15], [26], [43]. To further illustrate the robustness of regression performance to unseen data, we also prepare 100 extra independent LIDC nodules, which are not involved in the 10-fold and LOO CV schemes for evaluation.

Since a CT scan was reviewed by 4 radiologists, one nodule can be annotated by at least one radiologist and has at most 4 rating instances. For the 2400 pulmonary nodules, there are totally 982, 554, 441, and 423 nodules with one, two, three and four annotation instances from different radiologists, respectively. As can be observed, not every nodule can be found by all radiologists. To consider the issue of multiple rating instances, we implement two strategies of rating instance selection for the training and testing. For the first strategy, only one annotation instance of the nodule is used in the training process, if the nodule is treated as training sample in one fold. On the other hand, all instances of the same testing nodule are involved to evaluate the predicted scores from the learning methods. The first strategy is denoted as "single" for convenience. With the second strategy, we average scores of all rating instances for every nodule in the training and testing. Therefore, the second strategy is abbreviated as "average".

The expanding offset to define nodule ROIs from the bounding boxes of nodule outlines is the factor to determine how much background shall be included in the training and testing. To illustrate the effect of offset setting on the results of our regression schemes, we explores three offset settings of 5, 10 and 15 pixels, where the corresponding ROI sets are denoted as ROI1, ROI2 and ROI3, respectively.

To illustrate the effectiveness of the multi-task schemes, three single-task regression schemes, i.e., LASSO [39], elastic net [44] and standard linear regression, are implemented for comparison. All single-task regression schemes use the same general SDAE, Harr-like and HoG features but only consider the CNN features learnt from the corresponding task. Meanwhile, since it is also unclear which type of computational

TABLE III
PERFORMANCE ANALYSIS ON HETEROGENEOUS FEATURES. THE TASK ABBREVIATIONS ARE THE SAME AS DEFINED IN FIG. 1. METHOD ABBREVIATIONS SHARE THE SAME DEFINITION IN TABLE I

Feature \ Task	M	Tex	Sub	Spi	Sphe	Mar	Mal	Lob	IS	Cal
Haar-like	R	1.51±1.16	1.5±1.13	0.91±1.01	1.56±1.06	1.44±1.06	1.58±1.03	0.95±1.00	0.07±0.27	2.62±1.50
	C(L2)	0.84±0.68	0.86±0.63	0.93±0.75	0.97±0.66	1.04±0.69	1.11±0.78	0.94±0.72	0.01	0.09
	RF(L2)	0.78±0.74	0.85±0.61	1.00±0.69	0.86±0.54	1.01±0.71	1.09±0.73	0.99±0.69	0.03±0.20	0.56±0.75
	LS	2.02±1.23	2.08±1.09	1.02±1.00	1.58±1.07	1.61±1.12	1.58±1.03	0.95±1.00	0.06±0.26	2.65±1.51
	EN	1.67±1.21	1.95±1.12	1.41±1.25	1.94±1.12	1.49±1.14	1.61±1.01	1.39±1.20	0.03±0.21	2.82±1.50
HoG	R	0.85±0.81	0.90±0.68	0.89±0.70	0.86±0.55	1.00±0.71	0.97±0.65	0.90±0.68	0.04±0.19	0.62±0.8
	C(L2)	0.84±0.81	0.89±0.68	0.89±0.70	0.86±0.55	0.99±0.70	0.97±0.64	0.90±0.68	0.01	0.11
	RF(L2)	0.93±0.70	0.91±0.63	1.01±0.70	0.86±0.52	1.00±0.69	1.08±0.73	0.97±0.71	0.02±0.19	0.57±0.73
	LS	1.44±0.65	1.35±0.73	0.99±1.05	1.07±0.74	1.16±0.67	1.59±1.02	0.92±1.07	0.04±0.19	1.18±0.48
	EN	1.15±0.63	1.44±0.75	1.08±1.16	1.02±0.82	1.10±1.02	1.65±1.07	1.00±0.98	0.06±0.25	1.44±0.43
SDAE	R	0.82±0.76	0.86±0.65	0.95±0.72	0.88±0.53	1.01±0.71	1.05±0.74	0.94±0.70	0.04±0.19	0.64±0.78
	C(L2)	0.77±0.77	0.85±0.66	0.92±0.74	0.88±0.54	0.98±0.72	1.05±0.69	0.93±0.71	0.01	0.19
	RF(L2)	0.83±0.69	0.95±0.61	0.99±0.71	0.86±0.54	1.01±0.68	1.11±0.77	0.98±0.70	0.03±0.20	0.62±0.68
	LS	1.24±0.52	1.20±0.65	1.04±0.90	1.43±0.85	1.21±0.63	1.26±0.82	1.00±0.97	0.03±0.19	1.10±0.34
	EN	1.26±0.83	1.31±0.82	1.08±0.91	1.44±0.84	1.29±0.84	1.41±0.94	0.95±1.02	0.04±0.19	1.50±0.83
CNN	R	0.64±0.67	0.74±0.59	0.86±0.66	0.84±0.49	0.88±0.64	0.90±0.59	0.88±0.64	0.04±0.19	0.52±0.63
	C(L2)	0.64±0.66	0.73±0.58	0.86±0.67	0.84±0.50	0.86±0.63	0.91±0.60	0.88±0.64	0.01	0.08
	RF(L2)	0.81±0.70	0.88±0.63	0.98±0.69	0.85±0.53	1.02±0.72	1.05±0.71	0.97±0.71	0.03±0.20	0.54±0.73
	LS	1.23±0.56	1.02±0.64	0.98±0.99	1.09±0.77	1.25±0.74	1.23±0.80	0.95±0.84	0.03±0.19	1.12±0.39
	EN	1.28±0.55	1.08±1.07	1.03±1.09	1.14±0.80	1.17±1.04	1.24±0.84	0.98±0.89	0.03±0.19	1.21±0.85

features is more helpful for the regression, we also compare the sole use of the SDAE, CNN, Haar-like and HoG features on the two regression and the composite learning schemes as well as the single-task regression baselines. All single-task and multi-task learning schemes with various combination of feature types are evaluated with the same data partition of the all CV schemes for fair comparison.

On the other hand, because the CNN is the most popular deep learning model, we further investigate several implementation issues of CNN. Firstly, the depth setting of the CNN architecture is studied. Specifically, we explore three settings of depth, i.e., number of layers, for the computation of CNN features in the multi-task regression. The three settings of the number of weight layers are 4, 6, and 10. The weight layers are convolutional and fully-connected layers, where synaptic weights are adjustable. The depth is set as 4 in all other experiments. To simply illustrate the factor of CNN depth, the multi-task regression in this experiment will only consider the CNN features. The network details for the three depth networks can be found in Table A3(I) and Table A3(II) of the Supplement. For CNN features, we only take the fully connected layer of 192 neurons. Secondly, we directly perform the regression with the CNN architecture under either co-training and single-task scheme. It aims to illustrate whether the regression capability with only CNN architecture can be better than the regression power with the synergy among the heterogeneous features and the adopted regression methods. The details for the used CNN networks in the direction regression experiments can be found in the Table A3(III) and Fig. A1 of Supplement.

V. RESULTS

Table I summarizes the mean ± standard deviation statistics of the absolute difference between the computer predicted values and all radiologists' ratings w.r.t. the experiments of offset setting, annotation instance selection strategy, 10-fold, LOO and the independent 100 nodules. Meanwhile, the performance of co-training and single-task regression with only CNN, as well as the performance of single-task standard linear regression, can also be found in the experiments of LOO and independent 100 nodules. It is worth noting that we use the ROI2 setting, i.e., the expanding offset of 10 pixels, for all experiments, except the offset setting experiment, in this study. Meanwhile, all experiments (not including the "average" in Table I) use the "single" strategy for rating instance selection.

The inter-observer variation among the radiologists' ratings is also reported in Table I. Specifically, the inter-observer variation is computed from all possible pairs of annotation instances w.r.t. each pulmonary nodule and is quantitatively illustrated with the mean ± standard deviation statistics in Table I. It is worth noting that the performances of the "IS" and "Cal" tasks of t**he composite learning scheme** in Table I are **classification** error rates in 10-fold CV. On the other hand, the classification error rates of the "Cal" task from the R, RF, LS

and EN are 0.17, 0.21, 0.29 and 0.43, respectively, whereas the error rates of the "IS" from all methods are 0.01. Accordingly, the composite scheme has smaller error rate in "Cal" task. The classification error rates from the regression results are computed by rounding the predicted values. The performance for the tasks "IS" and "Cal" are very good, because the most nodules are rated in one degrees in these two tasks. For "IS" task, 2385, 3, 0 and 12 nodules were rated in scores 1, 2, 3 and 4, respectively. The nodule numbers for "Cal" task are distributed as 2, 4, 249, 17, 20, and 2108 for scores 1-6, respectively.

Referring to Table I, the three multi-task schemes are relatively robust to the various offset settings, whereas the two single-task schemes are more sensitive to different ROIs. On the other hand, comparing to performance of the ROI2, it can be found the absolute differences with the average strategy are slightly smaller. It may be because the inter-observer variation of the rating scores is reduced by the average operation.

The performance of co-training CNN regression is relatively close to the performance of random forest regression with auto-context augmentation, as can be found in the experiments of LOO and independent 100 nodules. On the other hand, the multi-task regression remain achieve less absolute differences in all semantic tasks than the CNN and auto-context random forest regression methods do. It may therefore corroborate the efficacy of the synergy between the heterogeneous features and multi-task regression (R) method. For single-task regression, the elastic net and CNN regression methods are relatively better but still general yield larger absolute differences than the multi-task methods with perceivably margins.

By and large, the prediction scores of the multi-task linear regression and the composite schemes are quite close to the radiologists' ratings. The scores from random forest with auto-context fusing as well as those form co-training CNN regression, deviate a little bit from the radiologists' scores but remain generally more accurate than the single-task regression schemes. The distributions of signed differences between the predicted scores from the 5 algorithms and the radiologists' ratings are also illustrated with box-plots in Fig. 3.

Table II compares the regression performance w.r.t. various depth settings of CNN architecture for the feature computing in the 5 regression schemes. As can be observed, deeper CNN architecture doesn't help in our problem. This may reflect the finding that deeper CNN may not always promise better performance in the work [45].

Table III summarizes the performance of sole use of each type of computational features w.r.t. the 5 methods. The experiment is done in 10-fold CV. It can be found that the deep learning features can help to achieve relatively good performance. By cross referencing the Table III and ROI2 in Table I, the combination of all heterogeneous features can effectively boost the regression results.

The performances of the random forest and the composite learning schemes are the results of the auto-context cross-task fusing with the second recursions, denoted as L2 in Tables I and II. The auto-context fusing doesn't help much for the composite learning scheme but does make some improvement for the random forest regression. The best performance can be attained at the second recursion iteration. The performance analysis w.r.t the number of recursions in the auto-context scheme can be found in the Tables A4 and A5 of Supplement.

TABLE IV
KENDALL'S COEFFICIENT OF CONCORDANCE (KCC) PERFORMANCE. THE TASK AND METHOD ABBREVIATIONS ARE THE SAME AS THOSE IN FIG.1 AND TABLE I. ROW GROUPS OF G1-G4 ARE THE KCC PERFORMANCES FOR THE GROUP 1-4. ALL ARE THE AGGREGATION FROM THE 4 GROUPS

	Tex	Sub	Spi	Sphe	Mar	Mal	Lob	IS	Cal
R-G1	0.77	0.72	0.72	0.64	0.73	0.74	0.71	0.99	0.89
C-G1	0.76	0.68	0.74	0.64	0.73	0.72	0.71	0.99	0.93
RF-G1	0.72	0.66	0.73	0.60	0.64	0.69	0.75	0.99	0.8
LS-G1	0.66	0.58	0.79	0.62	0.65	0.67	0.77	0.99	0.84
EN-G1	0.60	0.60	0.72	0.65	0.66	0.60	0.75	0.99	0.75
IB-G2	0.79	0.67	0.76	0.65	0.70	0.72	0.74	0.99	0.93
R-G2	0.70	0.60	0.62	0.55	0.63	0.69	0.61	0.99	0.87
C-G2	0.70	0.58	0.65	0.53	0.63	0.68	0.64	0.99	0.91
RF-G2	0.62	0.54	0.63	0.50	0.52	0.59	0.61	0.99	0.83
LS-G2	0.63	0.47	0.68	0.51	0.55	0.62	0.64	0.99	0.82
EN-G2	0.57	0.50	0.63	0.53	0.59	0.53	0.64	0.99	0.77
IB-G3	0.73	0.63	0.65	0.54	0.60	0.63	0.57	0.98	0.91
R-G3	0.69	0.59	0.58	0.50	0.58	0.62	0.52	0.98	0.86
C-G3	0.67	0.58	0.60	0.49	0.57	0.61	0.53	0.98	0.91
RF-G3	0.64	0.54	0.58	0.46	0.51	0.54	0.51	0.98	0.81
LS-G3	0.62	0.46	0.60	0.47	0.52	0.56	0.54	0.98	0.83
EN-G3	0.57	0.51	0.58	0.48	0.54	0.49	0.53	0.98	0.78
IB-G4	0.72	0.56	0.60	0.49	0.55	0.61	0.57	0.99	0.89
R-G4	0.69	0.55	0.57	0.46	0.53	0.61	0.53	0.99	0.88
C-G4	0.67	0.54	0.58	0.46	0.52	0.60	0.55	0.99	0.89
RF-G4	0.62	0.52	0.55	0.44	0.50	0.54	0.52	0.99	0.80
LS-G4	0.63	0.42	0.58	0.46	0.47	0.57	0.53	0.99	0.84
EN-G4	0.59	0.51	0.58	0.44	0.49	0.47	0.54	0.99	0.77
R-All	0.81	0.76	0.73	0.68	0.74	0.79	0.71	0.99	0.92
C-All	0.80	0.75	0.75	0.68	0.74	0.78	0.74	0.99	0.93
RF-All	0.75	0.73	0.72	0.65	0.69	0.71	0.72	0.99	0.81
LS-All	0.70	0.58	0.76	0.68	0.65	0.72	0.75	0.99	0.85
EN-All	0.64	0.67	0.73	0.66	0.69	0.60	0.74	0.99	0.77

Kendall's coefficient of concordance (KCC) [46] is also employed to evaluate not only the agreement between the computer predicted scores and radiologists' ratings but the inter-radiologist agreement (IB). The KCC analysis is applied on 4 nodule groups of the 2400 nodules, which contain 1, 2, 3 and 4 annotations, respectively, as well as the **All** group which aggregate all nodules from the 4 groups, and is reported in Table IV. The KCC ranges from 0 to 1, suggesting no agreement to complete agreement, respectively.

To further illustrate the significance of rating variation among radiologists, Fig. 4 lists several cases with significant annotation ambiguity from radiologists w.r.t. several semantic

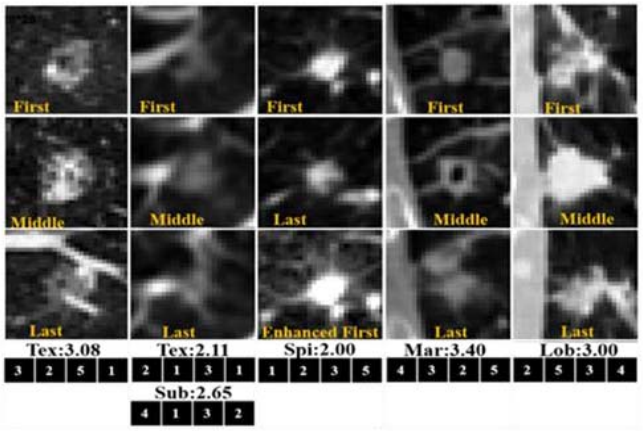

Fig. 4. Nodules with significant annotation ambiguity. The blue texted task scores are the predicted scores of multi-task linear regression. The task abbreviations are the same as those defined in Fig. 1.

tasks that are also shown under the ROIs of the five nodules. The left most nodule has high rating variation in "texture" and one radiologist gave the score 1 probably because the bright components was taken as vasculature or simply because of human error. The second nodule from the left has ambiguity in "texture" and "subtlety", as the rating process of these two semantic tasks may be subjective and depends on experience. One radiologist rated the third nodule as highly spiculated with score 5. It was probably because the spiculation can be salient with proper adjustment of HU window level as shown in the lowest ROI. The variation in the rating of semantic features of "margin" and "lobulation" for the fourth and fifth nodules may be caused by the bias from some particular slices.

VI. Discussion and Conclusion

Referring to Table I, the most promising performance can be achieved by the multi-task linear regression and the composite learning schemes, whereas the auto-context random forest can also achieve reasonably good performance. The three learning schemes are trained with annotation records from 12 of radiologists at different institutes in U.S.A. Therefore, the three learning schemes will consider inter-observer variation. Meanwhile, the relaxation from the integer annotated ratings to the floating regressed scores may also reserve the numeric flexibility to accommodate the rating ambiguity from different radiologists for less difference variation as shown in Table I.

The efficacy of heterogeneous computational SDAE, CNN, Haar-like and HoG features can also be observed in Table III. The sole usage of Haar-like features is less effective, whereas CNN features are the most promising type of features for either multi-task or single-task learning schemes. The effectiveness of the combination of all these heterogeneous features can be also be observed w.r.t. all learning schemes in Table I.

The performance tuning of the single-task LASSO and elastic net w.r.t. each task are carried out independently and hence the best parameter setting for each task is different. We try our best to achieve the best performance of the single-task regression for each task. The single-task regression schemes can also attain satisfactory prediction performance in the tasks like "spiculation" and "lobulation". However, the performances of "texture" and "subtlety" tasks are relatively more difficult to adjust and the predicted scores remain mostly deviate from the radiologists' ratings with more than 1. The performances of the three multi-task schemes can be tuned jointly across the 9 semantic tasks and achieve promising prediction results in all tasks. Therefore, the efficacy and efficiency of the MTL framework on our problem can be thus corroborated.

The inter-observer variation on the degree rating is discernible as it can be found in Table I and Fig. 4. For some nodules, the rating disagreement of some semantic features from different radiologists can be very significant. Meanwhile, it may also sometimes be quite difficult to give certain rating of semantic degree as the definition boundaries between the consecutive degrees/scores of the semantic features/tasks can be very vague and may vary from person to person. Even for the same radiologist, the semantic ratings at different time may also be inconsistent [11], as the semantic rating is a complicated perceptual process and also depend on the rater's condition.

Referring to Table I, the co-training CNN regression scheme can do better than single-task CNN regression, but still is not able to surpass the performance of multi-task regression. Since the multi-task regression method is originally designed for regression purpose and considers more heterogeneous features here, it may therefore perform better than the co-training CNN regression. On the other hand, because most well-known CNN models were formulated in the framework of classification, the loss function is usually set as soft-max loss. To fit in the context of regression, the loss function of the CNN regression schemes here is based on the Euclidean distance to measure the differences between the predicted and annotated scores. In such case, the error distance can be back propagated to derive useful neural parameters in the training of CNN regression schemes. The CNN training needs to optimize plenty parameters and usually require large training data to achieve good performance. Compared to many studies on the natural images, the number of 2400 nodules is relatively small. With more training samples, better CNN regression results may be possibly achieved.

The wider quantitative attributes of pulmonary with CADa may support medical doctors for more precise nodule analysis and management, e.g., correlation to histological images, radiology reports and genomics data, etc. Another potential application of the CADa scheme may consist in the support of medical search engine. As shown in an earlier study [47], 74% of the semantic feature terms defined in the LIDC dataset match to the terms in RadLex [40]. The automatic association between the lung CT contents and the clinical semantic terms may help to uphold sophisticated retrievals of clinical report and medical images from medical database [34-35] for better diagnostic decision support. Specifically, in the context of the content-based image retrieval (CBIR), the gap between the clinical semantic concepts and the low level image features is the major factor that thwarts the retrieval accuracy [35], [48]. Since simple computational image features may not be easily to address the issues of high intra-class variation and low inter-class difference, direct usage of low level image

features may not easily provide effective similarity measure. Therefore, the step of computational mapping between the clinical semantic terms and image contents to define the similarity measure at semantic level can possibly help to improve the CBIR performance [34]. With our automatic semantic scoring method, the computational visual similarity between pulmonary nodules may be simply leveraged to the semantic level with the differences of semantic scores. Accordingly, the retrieval can be based on overall similarity (considering all semantic features) or each specific semantic feature. The visual similarity can be easily combined with the ontological dissimilarity [34] with proper weighting.

To thrust the CBIR scheme toward clinical usage, we may further investigate the relation of semantic feature degrees and nodule subtypes as well as other clinical documents, e.g., treatment outcomes, to support more sophisticated retrieval functions. Specially, we may automatically establish the association between the medical image content and diagnosis/treatment level with the bridge of semantic scoring, to reach more precise retrieval. We will explore the image retrieval issue based on our multi-task regression framework in future study, as it may involves the correlation of other types of clinical data. In summary, the automatic scoring of semantic features for pulmonary nodules may support more precise CBIR and deeper research with more numerical references.

REFERENCES

[1] A. McWilliams *et al.*, "Probability of cancer in pulmonary nodules detected on first screening CT," *New Eng. J. Med.*, vol. 369, no. 10, pp. 910–919, 2013.

[2] V. K. Patel *et al.*, "A practical algorithmic approach to the diagnosis and management of solitary pulmonary nodules: Part 2: Pretest probability and algorithm," *Chest J.*, vol. 143, no. 3, pp. 840–846, 2013.

[3] D. P. Naidich *et al.*, "Recommendations for the management of subsolid pulmonary nodules detected at CT: A statement from the fleischner society," *Radiology*, vol. 266, no. 1, pp. 304–317, 2013.

[4] M. K. Gould *et al.*, "Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer: American college of chest physicians evidence-based clinical practice guidelines," *Chest J.*, vol. 143, no. 5, p. e93S, 2013.

[5] H. MacMahon *et al.*, "Guidelines for management of small pulmonary nodules detected on CT scans: A statement from the fleischner society," *Radiology*, vol. 237, no. 2, pp. 395–400, 2005.

[6] W. D. Travis *et al.*, "International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma," *J. Thoracic Oncol.*, vol. 6, no. 2, pp. 244–285, 2011.

[7] M. C. Godoy and D. P. Naidich, "Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: Recommended interim guidelines for assessment and management," *Radiology*, vol. 253, no. 3, pp. 606–622, 2009.

[8] J. H. M. Austin *et al.*, "Radiologic implications of the 2011 classification of adenocarcinoma of the lung," *Radiology*, vol. 266, no. 1, pp. 62–71, 2013.

[9] E. J. Hwang *et al.*, "Pulmonary adenocarcinomas appearing as part-solid ground-glass nodules: Is measuring solid component size a better prognostic indicator?" *Eur. Radiol.*, vol. 25, no. 2, pp. 558–567, Feb. 2015.

[10] J. J. Erasmus *et al.*, "Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions," *Radiographics*, vol. 20, no. 1, pp. 43–58, 2000.

[11] C. A. Ridge *et al.*, "Differentiating between subsolid and solid pulmonary nodules at CT: Inter–and intraobserver agreement between experienced thoracic radiologists," *Radiology*, vol. 278, no. 3, p. 150714, 2015.

[12] S. J. van Riel *et al.*, "Observer variability for classification of pulmonary nodules on low-dose CT images and its effect on nodule management," *Radiology*, vol. 277, no. 3, pp. 863–871, 2015.

[13] S. G. Armato, III, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[14] S. G. Armato, III, *et al.*, "Lung image database consortium: Developing a resource for the medical imaging research community," *Radiology*, vol. 232, no. 3, pp. 739–748, 2004.

[15] C. Jacobs *et al.*, "Solid, part-solid, or non-solid?: Classification of pulmonary nodules in low-dose chest computed tomography by a computer-aided diagnosis system," *Invest. Radiol.*, vol. 50, no. 3, pp. 168–173, 2015.

[16] J.-Z. Cheng *et al.*, "Computer-aided US diagnosis of breast lesions by using cell-based contour grouping," *Radiology*, vol. 255, no. 3, pp. 746–754, 2010.

[17] M. L. Giger *et al.*, "Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM," *Med. Phys.*, vol. 35, no. 12, pp. 5799–5820, 2008.

[18] A. El-Baz *et al.*, "3D shape analysis for early diagnosis of malignant lung nodules," in *Information Processing in Medical Imaging*. New York, NY, USA: Springer, 2011, pp. 772–783.

[19] T. W. Way *et al.*, "Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features," *Med. Phys.*, vol. 36, no. 7, pp. 3086–3098, Jun. 2009.

[20] T. W. Way *et al.*, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," *Med. Phys.*, vol. 33, no. 7, pp. 2323–2337, 2006.

[21] K. Awai *et al.*, "Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance," *Radiology*, vol. 230, pp. 347–352, 2004.

[22] K. Awai *et al.*, "Pulmonary nodules: Estimation of malignancy at thin-section helical CT-effect of computer-aided diagnosis on performance of radiologists," *Radiology*, vol. 239, pp. 276–284, 2006.

[23] H. Wang *et al.*, "Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image," *Eur. J. Radiol.*, vol. 74, no. 1, pp. 124–129, Apr. 2010.

[24] J.-Z. Cheng *et al.*, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans," *Sci. Rep.*, vol. 6, no. 1, p. 24454, 2016.

[25] J. Shi *et al.*, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, Jun. 2016.

[26] F. Ciompi *et al.*, "Bag-of-frequencies: A descriptor of pulmonary nodules in computed tomography images," *IEEE Trans. Med. Imag.*, vol. 34, no. 4, pp. 962–973, Apr. 2015.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[29] Y. Gao and D. Shen, "Collaborative regression-based anatomical landmark detection," *Phys. Med. Biol.*, vol. 60, no. 24, p. 9377, 2015.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, Jun. 2005, pp. 886–893.

[31] J. Chen *et al.*, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. SIGKDD ACM*, Aug. 2011, pp. 42–50.

[32] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[33] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, Oct. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20724753

[34] C. Kurtz *et al.*, "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications," *Med. Image Anal.*, vol. 18, no. 7, pp. 1082–1100, 2014.

[35] F. Zhang *et al.*, "Pairwise latent semantic association for similarity computation in medical imaging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 5, pp. 1058–1069, May 2016.

[36] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, Jun. 2014, Art. no. 4006.

[37] S. Chen *et al.*, "Bridging computational features toward multiple semantic features with multi-task regression: A study of CT pulmonary nodules," in *Medical Image Computing and Computer-Assisted Intervention*. New York, NY, USA: Springer, 2016.
[38] F. Gimenez *et al.*, "On the feasibility of predicting radiological observations from computational imaging features of liver lesions in CT scans," in *Proc. IEEE HISB*, Jul. 2011, pp. 346–350.
[39] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[40] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," *Radiographics*, vol. 26, pp. 1595–1597, 2006.
[41] A. Depeursinge *et al.*, "Predicting visual semantic descriptive terms from radiological image data: Preliminary results with liver lesions in CT," *IEEE Trans. Med. Imag.*, vol. 33, no. 8, pp. 1669–1676, Aug. 2014.
[42] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. ICML*, Jun. 2009, pp. 457–464.
[43] W. Shen *et al.*, "Multi-scale convolutional neural networks for lung nodule classification," in *Information Processing in Medical Imaging*. New York, NY, USA: Springer, 2015, pp. 588–599.
[44] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
[45] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE CVPR*, Jun. 2015, pp. 5353–5360.
[46] M. G. Kendall and B. B. Smith, "The problem of *m* rankings," *Ann. Math. Statist.*, vol. 10, no. 3, pp. 275–287, 1939.
[47] P. Opulencia *et al.*, "Mapping LIDC, RadLex, and lung nodule image features," *J. Digit. Imag.*, vol. 24, no. 2, pp. 256–270, 2011.
[48] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1276–1288, Jun. 2012.