This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2934577, IEEE Transactions on Medical Imaging

Multi-Task Deep Model with Margin Ranking Loss for Lung Nodule Analysis

Lihao Liu, Qi Dou *Member, IEEE,*, Hao Chen* *Member, IEEE,*,
Jing Qin *Member, IEEE,* and Pheng-Ann Heng *Senior Member, IEEE*

***Abstract*—Lung cancer is the leading cause of cancer deaths worldwide and early diagnosis of lung nodule is of great importance for therapeutic treatment and saving lives. Automated lung nodule analysis requires both accurate lung nodule benign-malignant classification and attribute score regression. However, this is quite challenging due to the considerable difficulty of lung nodule heterogeneity modeling and the limited discrimination capability on ambiguous cases. To solve these challenges, we propose a Multi-Task deep model with Margin Ranking loss (referred as MTMR-Net) for automated lung nodule analysis. Compared to existing methods which consider these two tasks separately, the relatedness between lung nodule classification and attribute score regression is explicitly explored in a cause-and-effect manner within our multi-task deep model, which can contribute to the performance gains of both tasks. The results of different tasks can be yielded simultaneously for assisting the radiologists in diagnosis interpretation. Furthermore, a Siamese network with a margin ranking loss is elaborately designed to enhance the discrimination capability on ambiguous nodule cases. To further explore the internal relationship between two tasks and validate the effectiveness of the proposed model, we use the recursive feature elimination method to iteratively rank the most malignancy-related features. We validate the efficacy of our method MTMR-Net on the public benchmark LIDC-IDRI dataset. Extensive experiments show that the diagnosis results with internal relationship explicitly explored in our model has met some similar patterns in clinical usage and also demonstrate that our approach can achieve competitive classification performance and more accurate scoring on attributes over the state-of-the-arts. Codes are publicly available at: https://github.com/CaptainWilliam/MTMR-NET**

***Index Terms*—Lung Nodule, Benign-Malignant Diagnosis, Attribute Score Regression, Deep Learning, Multi-Task**

I. Introduction

Lung cancer, which has a high morbidity and a low survival rate, is among the leading cause of cancer deaths worldwide [1]. For the year of 2019, an estimate of 228,150 new cases of lung cancer will be discovered, with over hundreds of thousands patients expected to die, accounting for approximately 26% of all cancer deaths in the United States [2]. Early lung cancer can be detected and screened by analyzing the lung nodules in chest computed tomography [3] images (known as CT) [4]–[7]. In clinical practice, the radiologists diagnose the benign and malignancy of nodules by observing the characteristics of lesion morphology. For example, as shown in Fig. 1, the lung nodules in the bottom row are labeled as malignant because their lobulation and spiculation score are higher than the benign lung nodules in the top row. However, this task is quite difficult even for well-trained radiologists, given the set of complicated attributes which are considered to be malignancy-related [8].

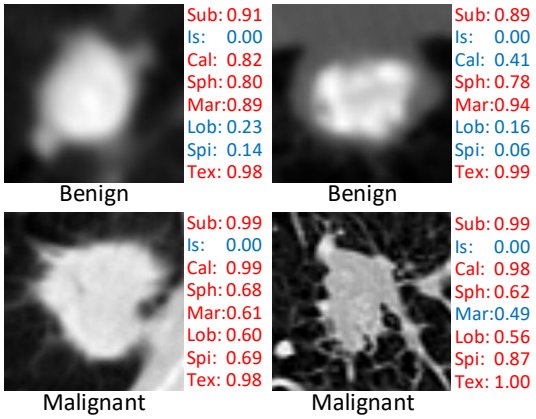

Fig. 1. Lung nodule diagnosis results of our proposed model. The top row shows two benign lung nodule diagnosis results; the bottom row shows two malignant lung nodule diagnosis results. The scores on the right side of each lung nodule are its corresponding eight attribute scores. Sub, Is, Cal, Sph, Mar, Lob, Spi, Tex denotes subtlety, internal structure, calcification, sphericity, margin, spiculation, lobulation and texture, respectively.

The automated analysis systems which can accurately estimate the malignancy as well as other eight attributes of lung nodules are highly demanded. Such reliable systems are very promising to reduce the mis-diagnosis rate and also improve clinical efficiency. This area has been frequently studied in recent several years and the proposed methods were mostly based on convolutional neural networks [9]. However, existing methods either separately consider these two tasks or implicitly explore the relationships between two tasks. Methods focus on benign-malignancy analysis often formulated this task as a binary classification problem, by classifying the nodule patches into benign or malignancy [10]–[12]. Chen et al. [10] proposed a multi-crop network, which incorporated size information for better benign-malignancy classification by using multiple max-pooling layers to extract feature maps at different scales. Xie et al. [11] utilized the

This work was supported by Hong Kong Innovation and Technology Commission under ITF ITSP Tier 2 Platform Scheme (Project No. ITS/426/17FP).

* indicates the corresponding author of this work.

L. Liu, H. Chen and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: lhliu@cse.cuhk.edu.hk; hchen@cse.cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

Q. Qou is with the Department of Computing, Imperial College London, UK (e-mail: qi.dou@imperial.ac.uk).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2934577, IEEE Transactions on Medical Imaging

overall appearance, nodule shape and gray-scale values of the nodules as inputs to three per-trained residual networks, and further integrated the output feature maps of the three networks to get the final classification result. Causey at el. [12] proposed a computational approach that systematically predicted lung nodule malignancy by combining quantitative image features extracted from pre-segmented CT image with a CNN model for lung nodule classification. Besides, researchers also try to explore the internal relationships between the significant morphologic attributes without the effect of malignancy. For example, Chen et al. [13] introduced a multi-task regression model to explore the internal relationship among the eight semantic features. Compared to standard methods such as lasso regression [14] and elastic net [15], they achieved lower absolute error between ground truth scores and model output scores.

As required in the radiological lung nodule diagnosis guideline [16], morphologic characteristics such as the smoothness of margin, spiculation contour and so on are to be carefully examined when performing the diagnosis. This domain knowledge indicates that predicting malignancy and analyzing other salient attributes are highly correlated and mutually supported. With this consideration, some other studies also work towards modeling the internal relationships between malignancy and other morphologic attributes of the nodules [17]–[20]. Instead of considering these two tasks independently, Hussein et al. [17] proposed a 3D CNN-based multi-task learning framework to explore the internal relationship between malignancy and attribute score by using graph regularized sparse least square optimization function. The relationship between the two tasks was well explored in this cause-and-effect manner as showed in Fig. 2(a). However, they only implicitly explored the internal relationship, thus, cannot output attribute scores. As mentioned before, Chen et al. [13] explored the internal relationship among the eight semantic features while excluding malignancy in their multi-task regression model. They further added malignancy back in their latest work [18], so they can explore the internal relationship of the malignancy and other eight attributes simultaneously. Similar to [18], Wang et al. [19] proposed to rank seven attributes (including malignancy) by using a WGAN-based over-sampling technique. Although their method jointly considered malignancy and other attributes simultaneously, the output attribute scores cannot be a definite assist-proof for the malignancy score, since it was not explored in a cause-and-effect logic, as shown in Fig. 2(b). To meet the high demands of interpretable computer-aided diagnosis (CAD) system, a method which can output benign-malignancy label and eight attribute scores in a cause-and-effect manner is indispensable.

To address these challenges, we first design a multi-task deep model [21] which explore the two tasks in a cause-and-effect manner. By applying this multi-task framework, we can not only train to get high accuracy results (benign-malignant classification) but also train to get other attribute scores (attribute score regression) which can further assist-proof the correction of the classification result. We build a two-branch multi-task architecture which not only predicts malignancy of the nodules but also outputs regressed scores of eight

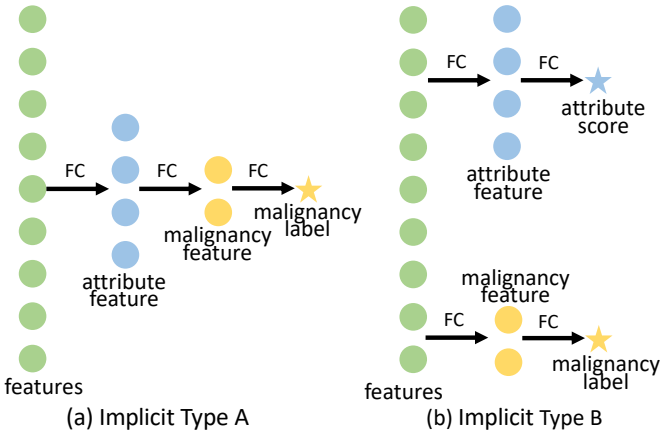

Fig. 2. Two types of multi-task framework. Note: the input feature vectors could be CNN / RNN extracted features, handcrafted features or any other pre-extracted features. (a). Extracting attribute features based on pre-extracted features, and further extract malignancy features from attribute features in a cause and effect relationship manner which only output malignancy label. (b). Extracting attribute features and malignancy features simultaneously based on pre-extracted features. This type can output malignancy label and attribute score at the same time but not in a cause and effect relationship manner.

attribute characteristics based on 50-layer ResNet [22]. The eight attribute characteristics are subtlety, internal structure, calcification, sphericity, margin, spiculation, lobulation and texture. The relatedness between two highly-correlated tasks is **explicitly** learned in our model, and both tasks can benefit from each other through the proposed multi-task learning scheme. We further explore the notable Siamese network architecture with a margin ranking loss to capture and harness the heterogeneity of lung nodules to increase sensitivity and accuracy of the proposed model on hard-classifying examples. The combination of pair-wise training strategy of Siamese network and margin ranking loss enables the model to be more accurate on those ambiguous nodules by referring to peer nodules. To further explore the internal relationship between malignancy and other attributes and validate the correctness of our model, we rank the relatedness between malignancy and eight attributes by applying the recursive feature elimination method. Moreover, based on the ranking results, we perform t-SNE visualizations using different sets of attributes.

We experimentally validate our proposed framework on the public LIDC-IDRI dataset and achieved competitive classification accuracy, sensitivity, specificity and area under the curve (AUC) results over the state-of-the-arts on benign-malignancy classification and lower absolute distance error on attribute scores regression. In addition, compared with previous approaches trying to either consider these two tasks independently or explore the relatedness of these two tasks implicitly which only output the final classification results, the proposed method can provide more clues and evidence for radiologists during the decision-making by yielding the scores of the eight attributes as assist-proofs for benign-malignancy label.

Our main contributions are summarized as follows:

1) We propose a multi-task deep learning model which explicitly explores the relatedness between lung nodule

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2934577, IEEE Transactions on Medical Imaging

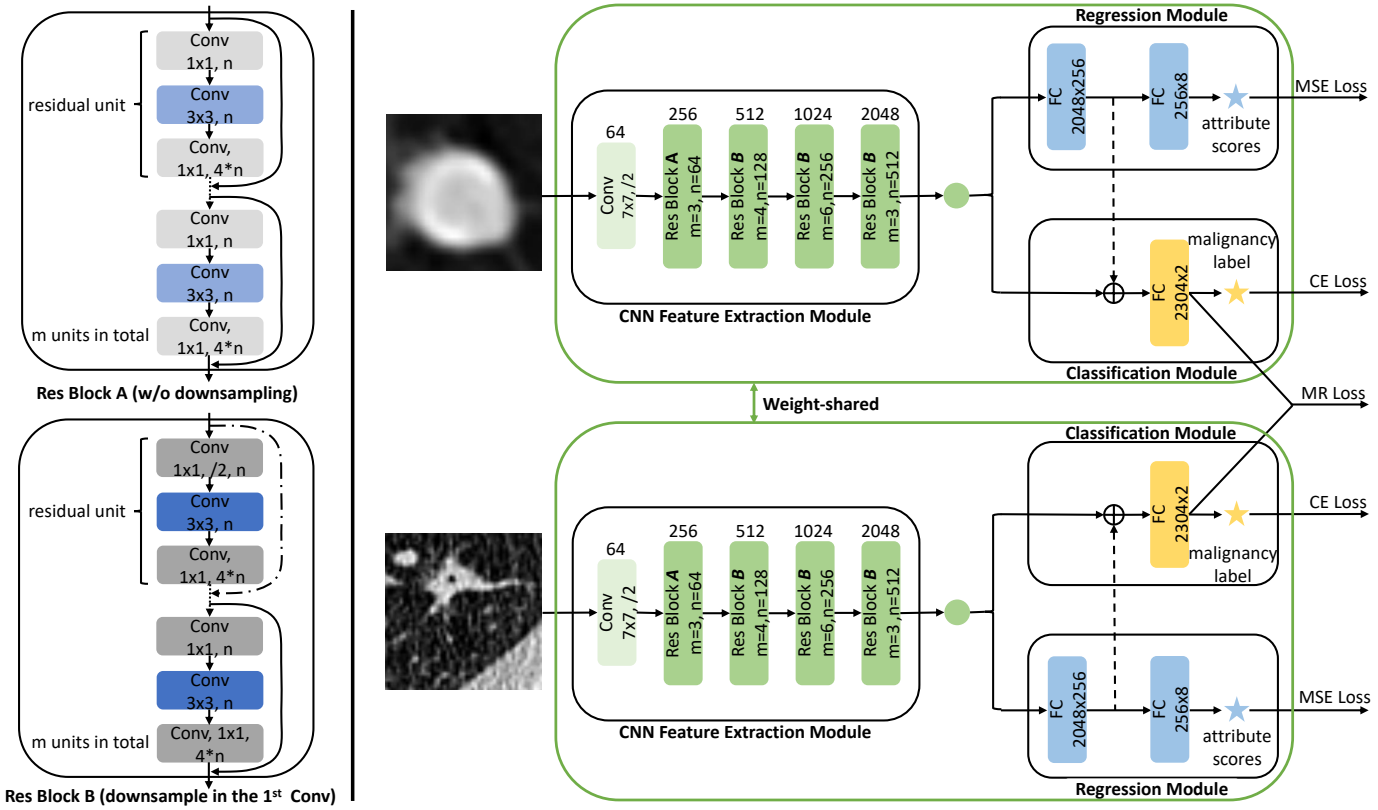

Fig. 3. The schematic illustration of the proposed network. Left part: Two residual blocks (Res Block A and Res Block B) described in residual network [23], where m is the total number of residual units and n is the channel number. Right part: The complete architecture of the proposed method. The green frame contains a multi-task learning framework in which there are three main modules: a feature extraction module, a classification module, a regression module. Feature extraction module consists of one convolutional layer, one Res Block A and 3 Res Block B. Classification Modules contains only one fully-connected layer followed by a cross entropy loss for the final benign-malignant label prediction. Regression module used 2 fully-connected layers followed by mean square error loss for the final attribute score regression. The extracted attribute feature map from the first fully-connected layer will be explicitly concatenated to the classification module as auxiliary deep supervisions. The "CE Loss", "MSE Loss" and"MR Loss"denotes cross entropy loss, mean square error loss and margin ranking loss, respectively. The entire network contains two weight-shared multi-task frameworks embedded with a margin ranking loss which was employed to increase discriminating capability of the model. Noted, weights for all three modules in the two frameworks are mutually shared.

benign-malignant classification and attribute regression in a cause-and-effect manner. The proposed model can yield benign-malignant diagnosis result and auxiliary attribute scores of lung nodules simultaneously, where the attribute scores work as assist-proofs for the diagnosis result.

2) We explore the renowned Siamese network architecture and its training strategy with a margin ranking loss in our model to better harness the heterogeneity of ambiguous lung nodules which further increase diagnostic accuracy.
3) We validate our proposed framework on the public benchmark LIDC-IDRI dataset using 5-fold cross validation. We achieved competitive benign-malignancy classification accuracy and superior performance on attribute scores regression over the state-of-the-arts.

This paper extends our preliminary work [24] by adding detailed architecture design of proposed method and redesigning extensive experiments to systematically evaluate the correctness of the model.

II. Method

Our proposed **M**ulti-**T**ask deep learning model with **M**argin **R**anking loss (MTMR-Net) consists of two components. Firstly, we proposed a multi-task deep learning framework for both benign-malignant classification task and attribute score regression task; see the green frame in Fig. 3. Noted, there are two multi-task frameworks (two green frames) which are weight-shared. The multi-task framework takes one 2d image as input, and outputs a benign-malignancy label and eight attribute scores simultaneously. Secondly, to further correctly classify these "marginal nodules" (lung nodules which is hard to classify due to close malignancy scores), we present a margin ranking loss for malignancy score ranking. Based on the Siamese network, we train the proposed model with this margin ranking loss in order to enhance the distinguishing capability among these "marginal nodules", as illustrated in the right part of Fig. 3.

A. Multi-Task Learning for Lung Nodule Analysis

There are two commonly-used methods for an automated lung nodule analysis system to provide evidence for a radiologist. The first method is malignant-benign classification, which simply classify a nodule into two categories [11], [12], [17], [25]–[27]. The binary label can give radiologists an intuitive computer-aided diagnosis result. The second method is to

provide attribute scores in the range of 0-1 for each nodule, which can be assist-proofs for the classification results [13], [17]. The score indicates the obviousness of each attribute of a lung nodule; the larger the score is, more obvious the attribute of the lung nodule is. In the proposed method, we integrate the two methods into a multi-task learning model to simultaneously produce a classification result and the attribute scores for a nodule.

*1) **Benign-Malignant Classification**:* Firstly, we extract a high-level semantic feature map (see the green circle in Fig. 3) from the input image by using the feature extraction module which is fine-tuned from the 152-layer residual network [22]. The feature extraction module consists of a single convolutional layer and two consecutive residual blocks (Residual block A and Residual block B) as shown in the left part of Fig. 3. Residual block A does not change the size of feature maps, while residual block B performs downsampling with a stride of 2 in the first convolutional layer within the first residual unit [23]. Repeated use of these two residual blocks allows the model to efficiently transfer spatial features to more complicated high-level semantic features. We name these features as classification features.

Secondly, based on those high-level features, we build a classification module. This module contains only one fully-connected layer; see the yellow layer in Fig. 3. In this module, we use the fully-connected layer to transfer the CNN-extracted high-level semantic features to the final class label. Compared to the original 152-layer residual network whose input channel of the last fully-connected layer is 2048, we change the input channel of this fully-connected layer to 2304. Because we concatenate the classification features with additional features before this fully-connected layer; see the dashed line connecting the classification and regression module in Fig. 3. These additional features are attribute features generated from the regression module with a size of 256. Hence, the final input size of fully-connected layer is 2304. This concatenation generates features consisting of both benign-malignancy and attribute information. In this regard, this concatenation allows the attribute features working as an auxiliary deep supervision for the classification module, which is similar to GoogleNet [28].

Thirdly, as it is a classification problem, we minimize the cross entropy loss (CE Loss) in the backward propagation process in the classification module to predict the classification label. The CE Loss is defined as:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i} log\ p_i^c\left(y_i^c|x_i; W_{cls}, W_s\right) \tag{1}$$

where x_i and p_i^c are the input image and output the correct probability from the classification module of network, while $y_i^c \in \{0, 1\}$ is the ground truth of lung nodule classification label, W_s and W_{cls} are the weights of shared feature extraction path and nodule classification task, respectively. N is the total number of training samples.

*2) **Nodule Attribute Score Regression**:* Motivated by the clinical observation that radiologists conduct diagnosis by analyzing the characteristics of attributes of nodules for malignancy assessments, we hypothesize that exploring the correlation between malignancy classification and attributes scoring would help to further improve the discrimination capability for lung nodule analysis. Therefore, besides the classification module, we add a regression module for attributes score prediction in the network to form a multi-task framework. The regression module is added based on the same feature extraction module and CNN-extracted high-level semantic features (classification features) as mentioned above. In the regression module, we firstly explicitly extract attributes features by using a single fully-connected layer from the high-level semantic features. Then we add another fully-connected layer to generate the final attribute scores; see the two blue layers in the regression module in Fig. 3.

In addition, rather than using these attributes features (extracted by the first fully-connected layer) solely for regression task, we concatenate the classification features in the classification module with these attributes features to get a new feature for the final benign-malignancy classification. This concatenating operation between classification features and attributes features enables more attribute information guidances in the nodule classification task. Hence, this multi-task architecture renders a mutual influence process between the classification and regression tasks. Moreover, in the backward propagation process [29], the classification loss (CE Loss) will also flow back to regression module through this concatenation operation and the dashed line shown in Fig. 3. To the end, this architecture also enables the interactions between the regression task and the classification task to boost the performance of both tasks.

For this attributes score regression task, we use mean square error loss (MSE Loss) to minimize the absolute distance error between ground truth values and output scores. The MSE Loss is defined as:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i} ||\hat{y}_i^r(x_i; W_s, W_{reg}) - y_i^r||_2^2 \tag{2}$$

where $y_i^r \in \mathbb{R}^{1\times n}$ is the output of regression task of network, while $\hat{y}_i^r \in \mathbb{R}^{1\times n}$ is the ground truth of attribute scores annotated by experts. N is the number of training samples. In our case, we use 8 semantic attributes, hence $n = 8$.

B. Margin Ranking Loss for Discriminating Marginal Nodules

Although multiple correlated supervision information is employed in our multi-task deep learning model, we still observe there exists misclassification on marginal lung nodules. By saying marginal lung nodules, we refer to the lung nodules with a benign-malignancy score distributed near to the classification divide line (margin). These nodules are the key obstacles in improving classification accuracy, since they have different benign-malignancy label but very similar malignancy scores. Marginal nodules are very ambiguous and confusing even for experts. If a model can make a more precise and definite classification on these marginal nodules, the diagnosis performance of the model will greatly increase.

Siamese network adopts two networks with the same architecture to train the network in a pair-wise manner [30]–[32]. The parameters are shared in both of the networks.

Based on the outputs of the two weight-shared networks in a Siamese network, we can further explore some other significant attributes between the two outputs, e.g. similarity or difference, by employing additional losses. To tackle this misclassification problem among marginal nodules, we employ this Siamese network architecture with a margin ranking loss to model heterogeneity between marginal nodules. In details, in the forward propagation process, the two weight-shared networks take a pair of 2d images as inputs, and calculate the malignancy score and attribute scores from the two input image through the two weight-shared networks, respectively. Thus, there are 2 inputs and 18 outputs of the Siamese network (one output malignancy score and eight output attribute score for each network). The attribute scores will be used to calculate the MSE Loss directly. While the malignancy score will be used to calculate two losses: CE Loss and the margin ranking loss.

Inspired by support vector machine (SVM), which uses hinge loss to perform maximum-margin classification, we utilize the similar formulation as presented in [32], [33] to formulate our marginal ranking loss. The margin ranking loss (MR Loss) designed for capturing the ranking relationship between malignancy scores of training samples is defined as:

$$\mathcal{L}_{rank} = \frac{1}{2N} \sum_{i,j} max\left(0, \gamma - \delta\left(p_i^c, p_j^c\right) * \left(t_i^c - t_j^c\right)\right) \quad (3)$$

$$\delta\left(p_i^c, p_j^c\right) = \begin{cases} 1, & p_i^c \geq p_j^c \\ -1, & p_i^c < p_j^c \end{cases} \quad (4)$$

where $t_i^c \in [0,1], t_j^c \in [0,1]$ denotes the ground truth malignancy score for the ith, jth training sample, respectively. While $p_i^c \in [0,1], p_j^c \in [0,1]$ are the predicted malignancy score of ith, jth training sample, respectively. $\delta\left(p_i^c, p_j^c\right)$ is the indicator function. γ is the margin parameter.

We assume margin parameter is fixed (e.g., $\gamma = 0$). If the ranking between the two predicted scores is the same as the ranking between two ground truth scores (e.g., $t_i^c \geq t_j^c, p_i^c \geq p_j^c$), then the loss is 0. Otherwise, the loss is penalized during the training process (e.g., $t_i^c \geq t_j^c, p_i^c < p_j^c$). Applying this mechanism into a Siamese network can easily explore and model the difference among marginal lung nodules by adjusting the margin parameter γ.

Compared to similar margin-based losses [32] which perform nodules comparison based on different labels (e.g., $t_i^c \in \{0,1\}, t_j^c \in \{0,1\}$), we employ continuous values for the ground truth and output malignancy scores to perform the ranking, e.g. $t_i^c \in [0,1], t_j^c \in [0,1]$. Benefited from this difference, the margin ranking loss can not only consider the nodules difference from different groups, but also maximize the margin between similar nodules within the same group. Hence, during the training stage, instead of choosing a pair of nodules with different labels, we randomly choose a pair of nodules to maximum the margin between nodules with similar malignancy scores.

C. Joint Training of MTMR-Net

In summary, we adopt three independent yet complementary losses in our proposed MTMR-Net. We assign three weights ($1,\ \beta\ and\ \eta$) to the CE Loss, MSE Loss and MR Loss, respectively. Moreover, to train the model in an end-to-end manner, we sum them up with L2 regularization in order to get the final loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} + \beta \mathcal{L}_{rank} + \eta \mathcal{R} \quad (5)$$

$$\mathcal{R} = ||W_s||_2^2 + ||W_{cls}||_2^2 + ||W_{reg}||_2^2 \quad (6)$$

where $\mathcal{R}$ is the regularization term which can prevent overfitting. λ, β, η are three hyper-parameters balancing $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ and $\mathcal{R}$.

III. Experiments

A. Dataset and Preprocessing

*1) **LIDC-IDRI Dataset**:* We validated the proposed MTMR-Net on the LIDC-IDRI dataset, which consists of 1018 CT scans [34] and 1422 lung nodules (972 benign lung nodules and 450 malignant lung nodules). We only keep the lung nodules with a diameter from 3 to 30 mm (which means each nodule contains 3 to 30 slices). In total, there are 9520 2D slices (6221 benign and 3299 malignant 2D slices).

Each nodule was rated from 1 to 5 by four experienced radiologists signifying the degree of malignancy in increasing order, e.g., a score of 1 means low malignancy while a score of 5 indicates high malignancy. Only those nodules with a diameter over 3 mm were considered in our experiments. To address the annotation disagreement among the radiologist, we follow [11], [13] and take the average of the benign-malignant ranking as the malignant score. For benign-malignant classification task, nodules with an average malignant score less than 3 and greater than 3 were labeled as benign and malignant, respectively. Nodules with an average malignant score of 3 were left out in our experiments similar to other works [11]–[13] since these nodules contain no benign-malignancy information. Besides malignancy, eight semantic attributes (i.e., subtlety, calcification, sphericity, margin, spiculation, texture, lobulation and internal structure) were also scored in the LIDC-IDRI dataset. Seven attributes were scored in an increasing order like malignancy. The higher the score is, the more obvious the characteristic is while internal structure's given score indicated 4 types of internal structure. Most features were rated in the range of 1-5, while the internal structure and calcification were given scores in the range of 1-4 and 1-6, respectively. We also take the same average strategy to obtain the ground truth attribute score. Furthermore, we take the ground truth label and scores for each nodule as the corresponding labels and scores for all slices of that nodule.

*2) **Five-fold Cross Validation**:* To demonstrate the effectiveness and robustness of the proposed methods, we conducted the experiments using 5-fold cross validation. We randomly split the data into 5 folds, where each fold contains 250 lung nodules. In the meantime, we keep the benign-malignant ratio of each fold similar to the ratio of the dataset. We trained on 4 folds and tested on the remained fold until each fold is used as testing data once. The final results are reported based on the average testing results of the 5 folds.

TABLE I
COMPLETE ARCHITECTURE OF MULTI-TASK DEEP LEARNING MODEL. NOTED, DOWNSAMPLING IS PERFORMED AT ALL RES BLOCK B'S 1^{st} CONV LAYER WITH STRIDE OF 2.

Module Name		Module Details	Output Size
Feature Extraction Module	Conv	$conv$ 7x7, stride 2	112x112x64
	Maxpooling	$maxpooling$ 3x3, stride 2	56x56x128
	Res Block A (m=3, n=64)	$\begin{bmatrix} conv\ 1x1, 64 \\ conv\ 3x3, 64 \\ conv\ 1x1, 256 \end{bmatrix}$ x3	56x56x256
	Res Block B (m=4, n=128)	$\begin{bmatrix} conv\ 1x1, 128 \\ conv\ 3x3, 128 \\ conv\ 1x1, 512 \end{bmatrix}$ x4	28x28x512
	Res Block B (m=6, n=256)	$\begin{bmatrix} conv\ 1x1, 256 \\ conv\ 3x3, 256 \\ conv\ 1x1, 1024 \end{bmatrix}$ x36	14x14x1024
	Res Block B (m=3, n=512)	$\begin{bmatrix} conv\ 1x1, 512 \\ conv\ 3x3, 512 \\ conv\ 1x1, 2018 \end{bmatrix}$ x3	7x7x2048
	Avgpooling	$avgpooling$ 7x7	2048
Regression Module		fc 2048x256	256
		fc 2048x8	8
Classification Module		fc (2048+256)x1	1

*3) **Preprocessing**:* To employ transfer learning from ImageNet models, we presented a nodule in multiple slices which can be feed directly into a 2D model. 3D CNN has more powerful ability to preserve spatial information than 2D CNN [25], [35]–[38]. However, we used 2D CNN to explore malignancy and semantic attribute scores of each slice, and then averaged the probability scores of slices enclosing nodule to get the final results as mentioned in [11]. This 2D method might lose some spatial information compared to 3D methods but the average operation can effectively prevent overfitting [11].

Moreover, we rescaled the average ground truth scores from 1-5, 1-6, 1-4 to 0-1 for normalization before training. We cropped an adaptive patch region according to the diameter and position of the nodule and resized the patch to 256×256 using bilinear interpolation. The diameter and position information can be directly obtained from the original dataset. In this way, the input image is of fixed size (256×256) causing nodules with different diameters to have a uniform size. To solve the data imbalance issue, we perform data augmentations on both benign and malignant lung nodules. We randomly crop a 224×224 region out of the 256×256 image and further randomly flip it horizontally or vertically to increase and the total number of both types of lung nodules. In the meantime, we perform more data augmentations on malignant nodules than benign ones to balance the total number of benign and malignant lung nodules.

B. Experiment Settings

*1) **Training Parameters**:* To employ transfer learning from pretrained models on ImageNet, we initialize the parameters of the feature extraction module in our network by using the parameters from the pretrained 152-layer residual network. Moreover, we utilize "Xavier" initialization method to initialize the parameters in both classification and regression modules. Adam optimizer was employed in all experiments for updating all parameters within the entire proposed network. Learning rate was initially set to 3e-3 for the feature extraction part and 3e-5 for both classification and regression modules. To accelerate the training process, the learning rate was periodically annealed by 0.1. We trained our model for 150 epochs using the Pytorch library with a batch size of 32. Moreover, we use grid-search to find out the suitable hyper-parameters. We set 3 hyper-parameters for controlling the weight of λ, β, η as 1, 5e-1, 1e-3, respectively, and the marginal parameter γ was chosen as 1e-1. The details of each layer within the multi-task architecture can be found in Table I.

*2) **Inference**:* In the inference stage, we abandoned half of the Siamese network architecture and keep only one multi-task deep learning model, which means we only need one image as input. Since the two branches of the Siamese network are trained in a weight-shared manner and there is no need to backpropagate the loss during testing. Hence, during the inference process, we only take one lung nodule as input and go through one branch to get the binary benign-malignancy label and other eight attribute scores. On average, our method takes less than three seconds to process a single lung nodule volume with the diameter between 3-30 mm in a single TITAN Xp GPU.

*3) **Evaluation Metrics**:* For classification problem, accuracy is the primary evaluation metric. However, only considering accuracy cannot comprehensively evaluate the performance of a model. We assigned a binary label to nodule based on its 0-5 averaged malignancy score by setting dividing line as 3. It is often argued why not using 2 or 4 as the dividing line. To alleviate the adverse effect of setting a specific dividing line, we need to consider model's sensitivity, specificity and especially area under the curve (AUC) similar to all other lung nodule classification works [11], [39]. The four evaluation metrics were defined as follow:

$$\mathcal{A}ccuracy = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{tn} + N_{fp} + N_{fn}} \tag{7}$$

$$\mathcal{S}encitivity = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{8}$$

$$\mathcal{S}pecificity = \frac{N_{tn}}{N_{tn} + N_{fp}} \tag{9}$$

$$\mathcal{A}UC = \frac{\sum_{i \in p} rank_i + N_p * (N_p + 1)}{N_p * N_n} \tag{10}$$

where N_{tp}, N_{tn}, N_{fp}, N_{fn} denote the number of true positive, true negative, false positive and false negative samples, respectively. N_p, N_n are the numbers of positive samples and negative samples, respectively. $rank_i$ is the rank of the ith positive example. $i \in p$ denotes the ith example from the positive sample[40].

For regression task, we used the absolute distance error (ABE) solely to evaluate the effectiveness of our model following [13], which was defined as follow:

$$\mathcal{A}BE = \frac{1}{N} \sum_i |\hat{y}_i^r - y_i^r| \tag{11}$$

where y_i^r is ith output of a single attribute score, while $\hat{y}_i^r$ is the ground truth of attribute score annotated by experts.

TABLE II
PERFORMANCE OF LUNG NODULE CLASSIFICATION METHODS ON LIDC-IDRI DATASET. ACC, SEN, SPE, AUC DENOTES ACCURACY, SENSITIVITY, SPECIFICITY AND AREA UNDER CURVE, RESPECTIVELY.

	Acc(%)	Sen(%)	Spe(%)	AUC
Anand et al. 2015 [41]	86.3	89.6	86.7	-
Xie et al. 2016 [11]	93.4	91.4	**94.1**	0.978
Shen et al. 2017[10]	87.1	77.0	93.0	0.930
MTMR-Net(w $\mathcal{L}_{reg}$)	90.1	90.5	88.5	0.974
MTMR-Net(w $\mathcal{L}_{rank}$)	90.9	91.2	91.6	0.961
MTMR-Net	**93.5**	**93.0**	89.4	**0.979**

TABLE III
PERFORMANCE OF ATTRIBUTE SCORE PREDICTION. IB INDICATES THE INTER-OBSERVER VARIATION, WHICH IS CALCULATED BASED ON ALL POSSIBLE PAIRS OF THE GIVEN SCORES FROM THE RADIOLOGISTS. MTR, LASSO, EN ARE MULTI-TASK REGRESSION MODEL [13], LASSO REGRESSION MODEL AND ELASTIC NETWORK, RESPECTIVELY. THE SUB, INT, CAL, SPH, MAR, LOB, SPI, TEX DENOTES SUBTLETY, INTERNAL STRUCTURE, CALCIFICATION, SPHERICITY, MARGIN, SPICULATION, LOBULATION, TEXTURE, RESPECTIVELY. THE SCORE IS CALCULATED ON THE ORIGINAL UNSCALED DATA.

	Sub	Int	Cal	Sph	Mar	Lob	Spi	Tex
IB	0.90	0.09	0.21	0.91	0.81	0.84	0.76	0.054
MTR [13]	0.75	0.04	**0.48**	0.81	0.86	0.87	0.80	0.58
LASSO [13]	1.25	**0.02**	2.18	1.25	1.13	0.95	0.89	1.04
EN [13]	1.20	0.14	1.44	1.09	0.98	0.96	0.86	1.24
MTMR-Net	**0.54**	0.03	0.56	**0.59**	**0.54**	**0.54**	**0.49**	**0.44**

C. Experiment Results

*1) **Benign-Malignant Classification**:* To utilize the multi-task framework for accurate benign-malignant classification, we employed four commonly-used metrics for the comparison: accuracy, specificity, sensitivity and area under the curve (AUC). Compared with state-of-the-art methods, our method achieved the best accuracy, sensitivity, AUC, and comparable specificity,which demonstrates the effectiveness of exploiting the relatedness between classification task and attribute prediction task as well as the margin ranking loss in improving the classification accuracy. In order to carefully scrutinize the contributions of different components of the final model, we further compared the proposed MTMR-Net with MSE Loss, and the MTMR-Net with MR Loss (two ablation studies). It is observed that the MTMR-Net with MR Loss achieved better performance in accuracy, sensitivity and specificity, compared to MTMR-Net only with MSE Loss. This proved that our margin ranking loss got the ability to enhance discriminating capability pf the proposed model. However, the accuracy of MTMR-Net with MSE Loss and MTMR-Net with MR Loss was very close. In multiple experiments, MTMR-Net with MSE Loss outperformed the MTMR-Net with MR Loss several times. This is because the multi-task framework with additional malignancy-related information can also increase the final classification accuracy. Moreover, MSE Loss enables the model to output eight attribute scores along with the malignancy label. Hence, this module is indispensable noting that the benefits from both losses can be superimposing. Thus, benefiting from the multi-task architecture and the margin ranking loss, our proposed MTMR-Net achieved the best accuracy, sensitivity, AUC, and comparable specificity. This also further proved the feasibility and correctness of our model.

*2) **Nodule Attribute Score Regression**:* To provide solid proofs for benign-malignancy classification results, we further compared the predicted results of attribute score regression module of our model with two commonly used models, lasso regression model and elastic network, as well as a state-of-the-art method, MTR [13]. We employed the metric of absolute distance error to evaluate the prediction results. As shown in Table III, the regression performance on the five semantic features is compared w.r.t multi-task regression model, lasso regression model and elastic network. We provided comparison results for all attributes. However, the prediction results for the "internal structure" task is extremely good. This is because most nodules were rated as score 1 which means there is no obvious pattern between the attribute and malignancy. These results affirm the superiority of our proposed model for distinguishing highly-correlated features. Our model provides an intuitive cascading benefit as follows: In comparison with previous methods, our model achieved a significantly lower absolute distance error on most of the attributes while the attribute prediction task can also improve the performance of the classification task. The classification task, in turn, enhanced the attribute prediction accuracy through the multi-task framework training based on the relatedness between these two tasks. To the best of our knowledge, compared to existing state-of-the-art methods, our model can reach the lowest absolute distance error on most of these eight attributes as shown in Table. III.

*3) **Lung Nodule Anaysis Results**:* To utilize our proposed MTMR-Net for explicitly exploring the relatedness between benign-malignancy and attribute scores, we present an intuitive lung nodule analysis results in Fig. 4. The left part showed the typical results from other state-of-the-art methods [10]–[12] which only consider the benign-malignancy classification task and output a binary label. The right part showed lung nodule analysis results from our proposed model which contains both the classification results and the corresponding attribute prediction results.

Inspiringly, we found our analysis results are quite consistent with those of previous clinical studies. As mentioned in [16], typically, benign nodules usually had well-defined and smooth margins whereas malignant nodules had spiculated margins and a lobular or irregular contour. This statement proved the correctness of the output of our multi-task model: the first and second benign nodule in the top row had higher margin score while all three malignant nodules in the bottom row had higher spiculation score as showed in Fig. 4. The right-most benign nodule in the top row had a very low margin score which means the margin of this lung nodule is blurred, however, the classification still stands correct. That was due to its subtlety score of 0.41 (subsolid), low spiculation score and low lobulation score. Another statement in [16] demonstrated that persistent subsolid nodules were more likely to be malignant. Hence, this nodule was correctly classified to benign because its subtlety, spiculation and lobulation score even

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2934577, IEEE Transactions on Medical Imaging

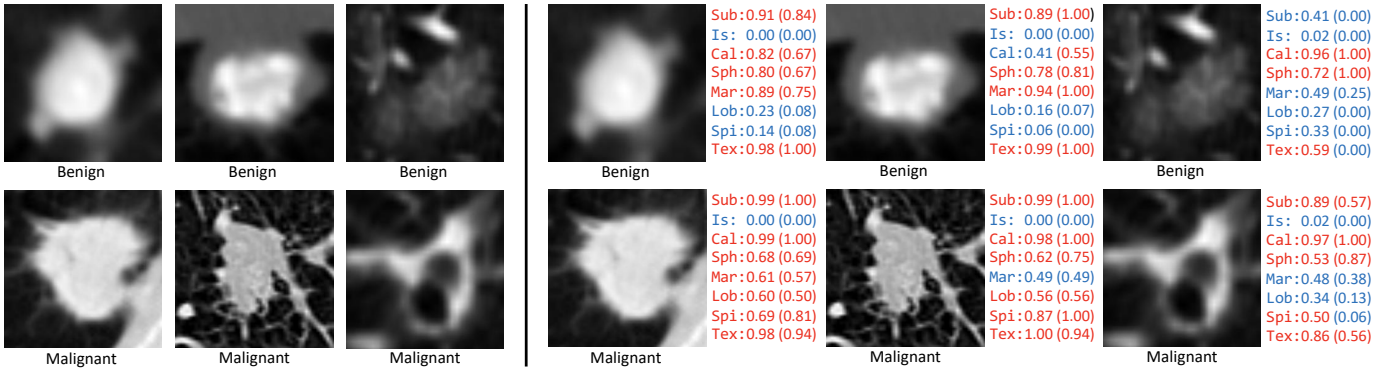

Fig. 4. **(Left)** classification outputs from previous work's model [11], [12]. **(Right)** classification outputs with attribute score from MTMR-Net. The bracket behind the predicted score contains the ground truth score for each attribute. Sub, Is, Cal, Sph, Mar, Lob, Spi, Tex denotes subtlety, internal structure, calcification, sphericity, margin, spiculation, lobulation, and texture, respectively. The score for each attribute is rescaled to the range of 0-1. The higher the score is, the more obvious the characteristic is.

though it had a very low margin score. These examples also demonstrated that we cannot classify the nodules based solely on one or two attributes. However, we should comprehensively consider more attributes simultaneously.

In comparison with previous methods without explicitly exploring the relatedness of benign-malignancy and other eight attribute scores, see Fig 4 (left part), our proposed model can provide more clues and evidence for diagnosis by simultaneously outputting the attribute scores besides better classification accuracy. Our proposed MTMR-Net can not only be used in automated lung nodule diagnosis systems, but also it can be employed as a tool for investigations that aim at revealing the underlying yet complicated relationship between the malignancy of a nodule and its highly-correlated attributes.

D. Recursive Feature Elimination and t-SNE Visualization

To further validate the correctness of our model and the correlations between two tasks, we utilize the recursive feature elimination method to iteratively rank the attributes among the eight attributes. Based on the ranking results, we further perform two t-SNE visualization experiments to intuitively present the relationship between the malignancy and other eight attributes.

1) ***Recursive Feature Elimination****:* Recursive feature elimination is a commonly-used feature selection method which can measure the importance of each feature to the model [42]. It iteratively eliminates unimportant features by calculating the correlation coefficients and its important ratio. In our work, we build a logistic regression classifier [43] as our model to validate the relationship between malignancy and other eight attributes. In each step of feature elimination, we kept the correlation coefficients and its important ratio of the removed feature. After the recursive elimination process is done, we obtain the final importance for each attribute in a nested manner based on the kept correlation coefficients and important ratio from each step.

As showed in Table IV, using recursive feature elimination, we found out 3 attributes have a relative low important score, e.g., internal structure, sphericity and texture have an important score lower than average. Those attributes had relatively lower correlation coefficients, compared to the other five attributes (subtlety, calcification, margin, spiculation, lobulation). We also highlight all the features whose final correlation coefficients were higher than average in Table IV.

As mentioned in [16], subtlety, margin, spiculation, lobulation were the four attributes highly correlated to malignancy, while calcification may not be a determinant factor for benign-malignancy diagnosis, it is still an important indicator for being benign. More Interestingly, the ranking results showed us that the relationship between benign-malignancy and attributes with high ranking relatedness score met some similar patterns in clinical use mentioned above. The listed malignancy factors in [16] contained all the attributes with relatedness score higher than the average score from the recursive feature elimination method. Besides, the most unrelated feature (first one to be removed in the recursive feature elimination method) is internal structure which also, in turn, proved that our assumption and observation were correct. While texture has a relatively low score, it is reported correlated to benign-malignancy of lung nodule in [39], [44]. Hence, instead of discarding attributes with low importance, we keep all attributes in our final proposed model for further exploring between benign-malignancy and the eight attributes.

2) ***t-SNE Visualization****:* We further conducted two t-SNE [45] experiments to present intuitive visualization results of the relationship between benign-malignancy and other eight attributes. As shown in Fig 5 (left), in the left experiment, we selected attributes with importance score lower than average score as inputs, and benign-malignancy as outputs. The t-SNE visualization results showed us that it is hard to make a definite classification based only on these attributes. Since we only need to prove the correctness of the model, instead of choosing highly-correlated attributes, we set the second experiment based on all eight attribute scores. As shown in Fig. 5 (right), the second visualization result showed us that consider the eight attributes simultaneously can make a more accurate benign-malignant classification on lung nodules, hence, it is reasonable to make a prediction based on all eight attributes.

TABLE IV
FEATURE SELECTION RESULTS. IMPORTANCE INDICATE THE CORRELATION COEFFICIENTS, PRESENTED AS MEAN±STD.

Attributtes	Sub	Int	Cal	Sph	Mar	Lob	Spi	Tex
Relatedness	0.251±0.19	0.007±0.001	0.133±0.020	0.072±0.006	0.125±0.009	0.171±0.027	0.164±0.030	0.067±0.003
Ranking	1	8	4	6	5	2	3	7
Above Average	✓		✓		✓	✓	✓	

Fig. 5. t-SNE visualizations. Red dots and blue dots denote malignant and benign lung nodules, respectively. **(Left)** Features are discarded attributes from after-trained network. **(Right)** Features are selected attributes from after-trained network.

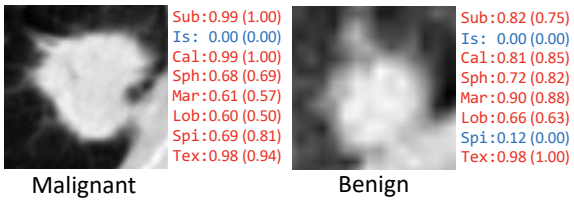

Fig. 6. Two similar lung nodules with different benign-malignant labels.

IV. DISCUSSION

Deep convolutional neural network (CNN) methods have obtained extraordinary discrimination capability in medical image classification tasks by extracting and adapting the highly semantic features. However, existing CNN-based methods focus on further improving the classification accuracy by applying insightful architectures and modules. Considering the lack of interpretability for CNN-extracted features, it is difficult to directly connect the classification results with the morphological attributes of a lung nodule, which has clinical relevance to benign-malignancy. To satisfy the requirement of high accuracy and interpretability of an automatic diagnosis CAD system, we propose a MTMR-Net which employ the multi-task framework to explore the relationship between benign-malignancy and other eight morphological attributes in a cause-and-effect manner which further boosts the performances for both tasks. Our proposed multi-task deep model is a foundational and efficient tool for lung nodule analysis, in which the two tasks are simultaneously considered (benign-malignancy classification and attribute score regression) . Besides the intuitive classification results, the method can also output eight attribute scores as supportive indicators for the classification results.

Multi-task framework is a commonly-used technique which applies different modules to tackle multiple tasks simultaneously within the same framework. By sharing the same basic module (i.e., ResNet backbone in the proposed method), this framework enables information interaction between different tasks. Based on the shared module and unshared modules (i.e., classification and regression modules in the proposed method), multi-task framework presents a mutual influence process and can further obtain performance gains for different tasks. Normally, the multi-task framework is presented in a parallel manner, where information interaction mainly based on the shared module; see Fig. 2(b). In our study, to increase the interpretability of the classification results, we add a connection from the regression module to the classification module. During the training process, we pass the attribute feature to the classification module through this connection, and concatenate it with the malignancy feature for the final label prediction. The operation establishes a cause-and-effect relationship between the benign-malignancy and other eight attributes. Moreover, in the backward propagation process, the classification loss will flow back to the regression module to influence the regression result, which further enhances the cause-and-effect logic between two modules. Since the cause-and-effect logic (the connection) in the multi-task framework is built based on the highly semantic features which lacks interpretability, the attribute scores can only be treated as indicators instead of definite proofs for the benign-malignancy. Moreover, Siamese architecture with the marginal ranking loss also helps to further distinguish similar nodules with different labels. In Fig. 6, we show two cases which have similar morphological features while having different benign-malignancy labels. By employing our Siamese network which enables comparison between nodules, our proposed model can successfully classify these two lung nodules into the correct class. While the model trained without Siamese network cannot make a correct prediction. Hence, our proposed model can successfully classify these two lung nodules into the correct class. The comparison between nodules from Siamese network by capturing the attributes scores, see Figure. In the further, we will further combine the probabilistic graphical model with the multi-task to explore a more definite relationship between benign-malignancy and other attributes.

In our experiments, we systematically perform ablation studies to evaluate the correctness of different parts of the proposed methods. Nevertheless, we only utilize the recursive feature elimination method to rank each attribute to validate the relationship with benign-malignancy. The attributes with high ranking are often regarded as supportive indicators to determine benign-malignancy of a lung nodule in clinical

usage, which proves that correctness of the relationship between attributes and benign-malignancy. However, our method currently considers the eight attributes comprehensibly, instead of exploring different patterns between only a few highly-correlated attributes (e.g., internal structure is still considered during classification even though it is less related to benign-malignancy). Even though some attributes might show lower correlations to malignancy, it could still play a significant role in assisting classification regarding extreme lung nodule cases [44]. Hence, in the future, we will consider the correlations between benign-malignancy with different attributes solely or different combinations of attributes to further improve the interpretability of the results of our models.

Besides classification and regression, extensive studies on LIDC-IDRI dataset for lung nodule analysis tasks (i.e., segmentation and detection) have achieved state-of-the-art performance benefited from CNN techniques. Existing CNN-based methods for cancerous lung nodule detection are often grouped into two categories: two-stage based methods and Faster RCNN based methods. Two-stage methods firstly detect candidate nodules regions and then reduce false positive candidates by using another classification module [46]–[49]. While Faster RCNN [50] architecture boosted the detection accuracy while maintaining a low number of false positive candidates in an end-to-end manner [51], [52]. Different from detection, segmentation tasks proposed to sketch out the detailed appearance of the lung nodule instead of region of interest by giving pixel-level label. Existing methods focus on encoding the shape information of lung nodules into highly semantic features and further decode segmentation masks from the high-level features, such as FCN [53], U-Net [54], V-Net [55] and Mask RCNN [56]. Accurate localization of lung nodules with CT scans (i.e., segmentation and detection tasks) works as the first step to perform benign-malignant classification and attribute score regression, which largely enhances the survival of the patient. In the feature, we will further explore to integrate the localization tasks and analysis tasks to boost each task's performance.

V. Conclusion

In this paper, we presented the MTMR-Net under a multi-task deep learning framework with margin ranking loss for automated lung nodule analysis. The relatedness between lung nodule classification and attribute score regression was explicitly explored with multi-task deep learning model, which contributed to the performance gains of both tasks. Furthermore, Siamese network was employed in a weight-shared manner with a margin ranking loss to model the nodule heterogeneity and encourage the discrimination capability on ambiguous nodule cases. To further explore the relatedness between benign-malignancy and attribute scores and the efficacy of the proposed model, we utilized a commonly-used feature selection method (recursive feature elimination) to iteratively rank the most malignancy-related attributes and intuitively present the relationship by applying t-SNE visualization experiments. Extensive experiments on the benchmark dataset verified the efficacy of our method and achieved competitive performance over the state-of-the-arts. In the future, we will explore our method on more heterogeneous tasks in medical image computing.

References

[1] A. del Ciello, P. Franchi, A. Contegiacomo, G. Cicchetti, L. Bonomo, and A. R. Larici, "Missed lung cancer: When, where, and why?" *Diagnostic and Interventional Radiology*, vol. 23, no. 2, pp. 118–126, 2017.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21551

[3] D. J. Brenner and E. J. Hall, "Computed tomography—an increasing source of radiation exposure," *New England Journal of Medicine*, vol. 357, no. 22, pp. 2277–2284, 2007.

[4] J. W. Gurney, "Missed lung cancer at ct: imaging findings in nine patients." *Radiology*, vol. 199, no. 1, pp. 117–122, 1996.

[5] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in ct images using deep convolutional neural networks," in *International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 243–250.

[6] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in ct images," in *Conference on Computer and Robot Vision (CRV)*, 2015, pp. 133–138.

[7] P. B. Bach, J. N. Mirkin, T. K. Oliver, C. G. Azzoli, D. A. Berry, O. W. Brawley, T. Byers, G. A. Colditz, M. K. Gould, J. R. Jett *et al.*, "Benefits and harms of ct screening for lung cancer: a systematic review," *Jama*, vol. 307, no. 22, pp. 2418–2429, 2012.

[8] M. Seemann, A. Staebler, T. Beinert, H. Dienemann, B. Obst, M. Matzko, C. Pistitsch, and M. Reiser, "Usefulness of morphological characteristics for the differentiation of benign from malignant solitary pulmonary lesions using hrct," *European Radiology*, vol. 9, no. 3, pp. 409–417, 1999.

[9] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[10] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognition*, vol. 61, pp. 663–673, 2017.

[11] Y. Xie, Y. Xia, J. Zhang, D. D. Feng, M. Fulham, and W. Cai, "Transferable multi-model ensemble for benign-malignant lung nodule classification on chest ct," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017, pp. 656–664.

[12] J. L. Causey, J. Zhang, S. Ma, B. Jiang, J. A. Qualls, D. G. Politte, F. Prior, S. Zhang, and X. Huang, "Highly accurate model for prediction of lung nodule malignancy with ct scans," *Scientific Reports*, vol. 8, no. 1, p. 9286, 2018.

[13] S. Chen, D. Ni, J. Qin, B. Lei, T. Wang, and J.-Z. Cheng, "Bridging computational features toward multiple semantic features with multi-task regression: A study of ct pulmonary nodules," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 53–60.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[16] M. T. Truong, J. P. Ko, S. E. Rossi, I. Rossi, C. Viswanathan, J. F. Bruzzi, E. M. Marom, and J. J. Erasmus, "Update in the evaluation of the solitary pulmonary nodule," *Radiographics*, vol. 34, no. 6, pp. 1658–1679, 2014.

[17] S. Hussein, K. Cao, Q. Song, and U. Bagci, "Risk stratification of lung nodules using 3d cnn-based multi-task learning," in *International Conference on Information Processing in Medical Imaging*, 2017, pp. 249–260.

[18] S. Chen, J. Qin, X. Ji, B. Lei, T. Wang, D. Ni, and J.-Z. Cheng, "Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in ct images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 3, pp. 802–814, 2017.

[19] Q. Wang, X. Zhou, C. Wang, Z. Liu, J. Huang, Y. Zhou, C. Li, H. Zhuang, and J.-Z. Cheng, "Wgan-based synthetic minority over-sampling technique: Improving semantic fine-grained classification for lung nodules in ct images," *IEEE Access*, vol. 7, pp. 18 450–18 463, 2019.

[20] Q. Dou, H. Chen, Y. Jin, H. Lin, J. Qin, and P.-A. Heng, "Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 630–638.

[21] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 630–645.

[24] L. Liu, Q. Dou, H. Chen, I. E. Olatunji, J. Qin, and P.-A. Heng, "Mtmr-net: Multi-task deep learning with margin ranking loss for lung nodule analysis," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 74–82.

[25] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[26] W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary nodule classification with deep convolutional neural networks on computed tomography images," *Computational and Mathematical Methods in Medicine*, vol. 2016, 2016.

[27] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *International Conference on Information Processing in Medical Imaging*, 2015, pp. 588–599.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[29] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[30] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] D. J. Rao, S. Mittal, and S. Ritika, "Siamese neural networks for one-shot detection of railway track switches," *arXiv preprint arXiv:1712.08036*, 2017.

[32] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 662–679.

[33] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[34] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[35] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 149–157.

[36] H. Chen, Q. Dou, L. Yu, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation," *arXiv preprint arXiv:1608.05895*, 2016.

[37] A. Riccardi, T. S. Petkov, G. Ferri, M. Masotti, and R. Campanini, "Computer-aided detection of lung nodules via 3d fast radial transform, scale space representation, and zernike mip classification," *Medical Physics*, vol. 38, no. 4, pp. 1962–1971, 2011.

[38] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[39] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal, "A combination of shape and ure features for classification of pulmonary nodules in lung ct images," *Journal of Digital Imaging*, vol. 29, no. 4, pp. 466–475, 2016.

[40] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[41] S. V. Anand, "Segmentation coupled textural feature classification for lung tumor prediction," in *2010 IEEE International Conference on Communication Control and Computing Technologies (ICCCCT)*, 2010, pp. 518–524.

[42] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[43] P. McCullagh and J. A. Nelder, *Generalized linear models*. CRC press, 1989, vol. 37.

[44] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, 2015.

[45] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[46] S. Zheng, J. Guo, X. Cui, R. N. Veldhuis, M. Oudkerk, and P. van Ooijen, "Automatic pulmonary nodule detection in ct scans using convolutional neural networks based on maximum intensity projection," *arXiv preprint arXiv:1904.05956*, 2019.

[47] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.

[48] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[49] Q. Wang, F. Shen, L. Shen, J. Huang, and W. Sheng, "Lung nodule detection in ct images using a raw patch-based convolutional neural network," *Journal of digital imaging*, pp. 1–9, 2019.

[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99.

[51] X. Huang, W. Sun, T.-L. B. Tseng, C. Li, and W. Qian, "Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic ct scans using deep convolutional neural networks," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, vol. 74, p. 25—36, March 2019.

[52] N. Sun, D. Yang, S. Fang, and H. Xie, "Deep convolutional nets for pulmonary nodule detection and classification," in *International Conference on Knowledge Science, Engineering and Management*, 2018, pp. 197–208.

[53] A. Yaguchi, K. Aoyagi, A. Tanizawa, and Y. Ohno, "3d fully convolutional network-based segmentation of lung nodules in ct images with a clinically inspired data synthesis method," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, 2019, p. 109503G.

[54] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[55] A. Hatamizadeh, S. P. Ananth, D. Terzopoulos, X. Ding, and N. Tajbakhsh, "Automatic, fast, and reliable lung lobe segmentation through a 3d progressive dense v-network."

[56] M. Liu, J. Dong, X. Dong, H. Yu, and L. Qi, "Segmentation of lung nodule in ct images based on mask r-cnn," in *2018 9th International Conference on Awareness Science and Technology (iCAST)*, 2018, pp. 1–6.