An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification

Shiwen Shen[a,b], Simon X. Han[a,b], Denise R. Aberle[a,b], Alex A.T. Bui[b], Willliam Hsu[b,*]

[a]*Department of Bioengineering, University of California, Los Angeles, CA, USA*
[b]*Medical Imaging & Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, CA, USA*

Abstract

While deep learning methods are increasingly being applied to tasks such as computer-aided diagnosis, these models are difficult to interpret, do not incorporate prior domain knowledge, and are often considered as a "black-box." The lack of model interpretability hinders them from being fully understood by target users such as radiologists. In this paper, we present a novel interpretable deep hierarchical semantic convolutional neural network (HSCNN) to predict whether a given pulmonary nodule observed on a computed tomography (CT) scan is malignant. Our network provides two levels of output: 1) low-level radiologist semantic features; and 2) a high-level malignancy prediction score. The low-level semantic outputs quantify the diagnostic features used by radiologists and serve to explain how the model interprets the images in an expert-driven manner. The information from these low-level tasks, along with the representations learned by the convolutional layers, are then combined and used to infer the high-level task of predicting nodule malignancy. This unified architecture is trained by optimizing a global loss function including both low- and high-level tasks, thereby learning all the parameters within a joint framework. Our experimental results using the Lung Image Database Consortium (LIDC) show that the proposed method not only produces interpretable lung cancer predictions but also achieves significantly better results compared to common 3D CNN approaches.

Keywords: Lung nodule classification, lung cancer diagnosis, Computed tomography, deep learning, convolutional neural networks, model interpretability

1. Introduction and Background

Lung cancer is the leading cause of cancer mortality worldwide [1, 2]. Computed tomography (CT) imaging is widely used to detect pulmonary nodules and forms the basis for diagnosing lung cancer. The landmark National Lung Screening Trial (NLST) [3] in the United States demonstrated a 20% lung cancer mortality reduction in high-risk subjects who

---

*Medical Imaging & Informatics, 924 Westwood Boulevard, Suite 420, Los Angeles, CA 90024, USA.

Email addresses: shiwenshen@engineering.ucla.edu (Shiwen Shen), simon.x.han@ucla.edu (Simon X. Han), daberle@mednet.ucla.edu (Denise R. Aberle), buia@mii.ucla.edu (Alex A.T. Bui), whsu@mednet.ucla.edu (Willliam Hsu)

underwent screening using low-dose CT (relative to using plain chest radiography). Based on the findings of the NLST, the United States Preventative Services Task Force (USPSTF) went on to recommend low-dose CT lung cancer screening for current and former smokers aged 55-80 with a smoking history of at least 30 pack-years, or former smokers having quit within the past 15 years [4]. However, the potential consequences of implementing lung cancer screening is an increase in false positive screens that result in unnecessary medical, economic, and psychological costs. Indeed, some studies indicate that the false positive rate for low-dose CT is upwards of 20%; less experienced radiologists have highly variable detection rates, particularly in subtle cases, as interpretation heavily relies on past experience [5]. Figure 1 shows examples of malignant and benign nodules, helping to illustrate the challenges of differentiating between the two groups.
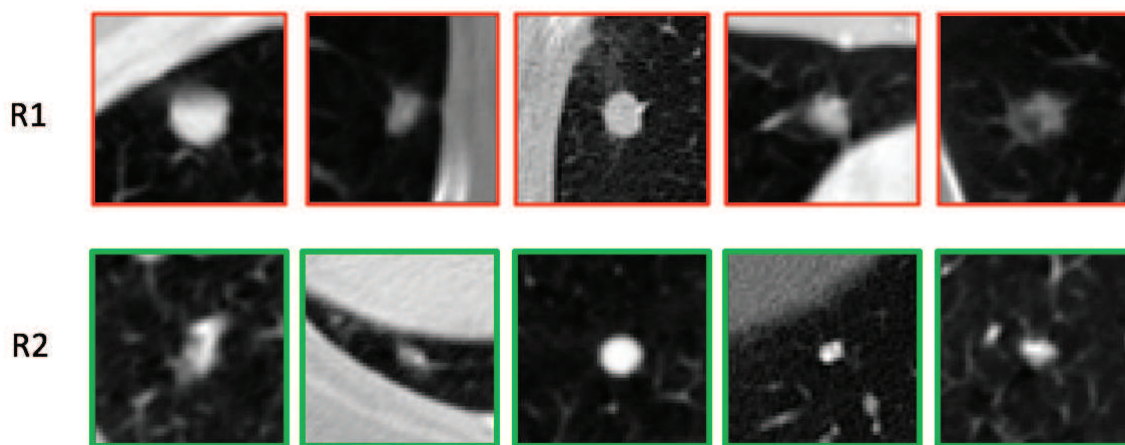

Figure 1: Illustrations of malignant and benign nodules: R1 are malignant nodules; R2 are benign nodules.

In response, computer-aided diagnosis (CADx) systems [6, 7, 8, 9, 10, 11] have been explored to assist radiologists in the interpretation process; these help to separate malignant from benign nodules and show promise in increasing the positive predictive value and reducing the false positive rates in characterizing small nodules [11]. Broadly, a contemporary lung nodule CADx system comprises three components: 1) nodule segmentation; 2) feature extraction from nodule candidates; and 3) classification of the candidate as benign or malignant based on its extracted features. Segmentation of the lung nodule is the first step to identify pertinent regions of interest (ROIs) so as to focus the ensuing analyses. In the second step, computed features such as texture, morphology, and gray-level statistics are extracted to characterize the nodule across various visual and spatial dimensions. Finally, using the computed features, the nodule candidates are categorized into benign or malignant by a classifier such as support vector machine, random forest, or gradient boosting tree. For example, Armato et al. [6] segmented the lung nodule using multilevel thresholding techniques; extracted morphological and gray-level features; and classified nodules as benign or malignant using linear discriminant analysis. Zinovev et al. [12] employed both texture and intensity features using belief decision trees and a multi-label approach to perform lung nodule classification. Way et al. [13] segmented lung nodules using k-means clustering, combined nodule surface features together with texture and morphological features, and used linear

discriminant analysis to diagnose malignant lung cancers. Notably, the prevalent portion of these extracted features, such as volume and shape, are sensitive to the underlying variability arising from the lung nodule segmentation results. This variability may be caused by: 1) the challenges inherent in nodule delineation due to the range of nodule morphology and connectivity complexity [14]; and 2) the variation posed by the computational segmentation models themselves in capturing quantitative characteristics [15, 16]. Thus, using segmented regions may lead to inaccurate features that are subsequently used as inputs into downstream classifiers [15]. Another critical question raised by this type of CADx design is how to define the "optimal" subset of features that can best encode characteristics of the lung nodule [17]. The optimal feature set is dependent on the characteristics of the training dataset, and feature selection and classification methods, which makes achieving comparable results using different datasets difficult.

To overcome these issues, deep learning methods [14, 15, 17, 18, 19], particularly convolutional neural networks (CNNs), have recently been used for lung nodule classification, with some promising results. These deep learning models prove capable of adaptively learning image representations in a fully data-driven way, taking raw image data as input without relying on *a priori* nodule segmentation masks or handcrafted feature designs. For instance, Kumar et al. [18] first trained an unsupervised deep autoencoder to extract latent features from 2D CT patches. These extracted deep features were then used together with decision trees to predict lung cancer. Hua et al. [19] employed supervised techniques with a deep belief network and CNN to train models to classify lung nodules as benign or malignant. Their models outperformed two baseline methods: using scale-invariant feature transform (SIFT) features and local binary patterns (LBP) [20]; and using fractal analysis [21]. Ciompi et al. [17] used pre-trained CNN models to classify candidates as peri-fissural nodules (PFNs) or non-PFNs. Deep features were extracted from the pre-trained model for three 2D image patches in axial, coronal, and sagittal views. An ensemble of the deep features and a bag of frequency features were then used to train supervised binary classifiers for the PFN classification task. Shen et. al. [14] designed a multi-scale CNN using 3D nodule patches at three scales to perform the lung cancer diagnosis task. This work is further extended in [15] by adding a multi-crop pooling strategy to improve model performance. Markedly, these cited works use deep learning as a "black-box" and are not able to explain what representations have been learned or why the model generates a given prediction. This low degree of interpretability arguably hinders domain experts, such as radiologists, from understanding how the models work and ultimately impedes model adoption for clinical usage. As discussed in [22], improved interpretability is helpful to improve the radiologist-CAD interaction to allow radiologists to calibrate their trust in the CAD system. Moreover, human domain knowledge performs well across a wide range of observed nodule morphologies to differentiate benign from malignant nodules. Nonetheless, these domain knowledge are presently not incorporated into deep learning frameworks.

A number of radiologist-interpreted features derived from CT scans have been considered influential when assessing the malignancy of a lung nodule [23, 24]. These features are referred to as diagnostic *semantic* features in this study. Examples of such semantic features include nodule spiculation, lobulation, consistency (texture) and shape. Although qualitative in nature, studies have shown that these semantic features can be characterized numerically using low-level image features [25]. Hancock et al. [26] demonstrated that ma-

chine learning can achieve high prediction accuracy for lung cancer malignancy using only semantic features as inputs. In addition, semantic features are intuitive to radiologists and are moderately robust against perturbations in image resolution and reconstruction kernel. An opportunity exists to incorporate these semantic features into the design of deep learning models, combining the advantages of both.

In this study, we propose a novel interpretable hierarchical semantic convolutional neural network (HSCNN) to predict whether a nodule is malignant in CT images. The HSCNN takes the raw CT image cubes centered at nodules as input and generates two levels of outputs. The first predictive level provides intermediate outputs in terms of diagnostic semantic features, while the second level represents the final lung nodule malignancy prediction score. Jump connections are employed to feed the information learned from the first level semantic features to the final malignancy prediction. As such, our first level outputs provide explanations about what the HSCNN model has learned from the raw image data and correlates semantic features with the specific malignancy prediction; it also provides additional information to improve the final malignancy prediction task through the jump connections. This unified model is trained as a whole by minimizing a global cost function, where both first- and second-level task losses are included.

The contributions of this paper are fourfold:

1. We describe an approach to build a radiologist-interpretable deep convolution neural network. The intermediate outputs from the model give predictions of diagnostic semantic features associated with the final classification, helping to explain the prediction. To the best of our knowledge, this is the first example of a network architecture emphasizing the interpretability of the results.

2. We provide a hierarchical design that integrates both semantic features and deep features to predict malignancy. Shared convolution modules in the HSCNN are used to learn generalizable features across tasks. The information learned for each specific low-level semantic feature is then fed into the final high-level malignancy prediction task.

3. We present a new global cost function to train the whole model jointly, taking both first- and second-level outputs into consideration simultaneously. The new objective function concurrently handles data imbalance issues for both tasks.

4. We show how semantic features can be quantified using a deep CNN model from reconstructed CT images in a data-driven fashion. The produced model generates labels for all semantic features on previously unseen cases.

The remainder of this paper is organized as follows. In Section 2, we describe the dataset used in this study and the proposed HSCNN model. In Section 3, we present results and compare the proposed method with a traditional 3D CNN. In Sections 4 & 5, we discuss the findings and limitations of the work.

2. Materials and Methods

2.1. Lung Image Database Consortium Dataset

The Lung Image Database Consortium image collection (LIDC-IDRI) [27] is a publicly available dataset, which we used to train and test our proposed methods. LIDC-IDRI contains both screening and diagnostic CT scans collected from 7 academic centers and 8 medical imaging companies. Inclusion criteria for CT scans were: 1) having a collimation and reconstruction interval no greater than 3 mm; and 2) each scan approximately containing no more than 6 lung nodules with the longest dimension ranging from 3-30 mm, as determined by a cursory review during case selection at the originating institution [27]. For the whole dataset, the slice thicknesses varied from 0.6 to 5 mm, and the in-plane pixel size varied from 0.461 to 0.977 mm. LIDC-IDRI comprises 1,018 cases (representing 1,010 different patients, 8 patients having 2 distinct scans), with each including images from a clinical CT scan and an associated eXtensible Markup Language (XML) file. The XML files record the reference standard for lung nodule locations, as manually annotated by four radiologists following a two-phase image annotation process. Pixel-level 3D contour segmentations, panel opinions on the assessment of nodule likelihood for malignancy, and eight nodule characteristics were generated only on lesions categorized as nodules $\geq$ 3 mm. The eight nodule characteristics are semantic diagnostic features, including: calcification, subtlety, lobulation, sphericity, internal structure, margin, texture, spiculation, and malignancy. Each feature is rated from 1 to 5 or 6 by radiologists. Table 1 lists the description and definitions for each of the labels from [28].

Table 1: Detailed nodule characteristics labels in LIDC dataset.

Semantic Feature	Description	Ratings
Malignancy	Likelihood of malignancy	1. Highly unlikely 2. Moderately unlikely 3. Indeterminate 4. Moderately suspicious 5. Highly suspicious
Margin	How well defined the margins are	1. Poorly defined 2. 3. 4. 5. Sharp
Sphericity	Three dimensional shape in terms of roundness	1. Linear 2. 3. Ovoid 4. 5. Round
Subtlety	Difficulty of detection relative to surround	1. Extremely subtle 2. Moderately subtle 3. Fairly subtle 4. Moderately obvious 5. Obvious

Semantic Feature	Description	Ratings
Spiculation	Degree of exhibition of spicules	1. Marked 2. 3. 4. 5. None
Radiographic solid-ity (texture)	Internal texture (consistency) of nodule	1. Non-solid 2. 3. Part Solid 4. 5. Solid
Calcification	Presence and pattern of calcification	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Internal structure	Expected internal composition of the nodule	1. Soft tissue 2. Fluid 3. Fat 4. 5. Air
Lobulation	The presence and degree of lobula-tion of the nodule margin	1. Marked 2. 3. 4. 5. None

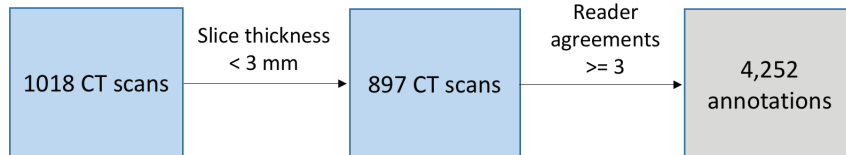## 2.2. Our Usage of the LIDC dataset

Figure 2: Lung nodule annotation selecting criteria.

In the LIDC, nodules could be associated with 1-4 annotations, depending on how many of the four readers completed the annotation task. To determine which annotations referred to the same nodule, we used an annotation list provided in [29]. Only nodules identified by at least three radiologists on CT scans with slice thicknesses smaller than 3 mm were included in this study, as shown in Figure 2. The inclusion criteria resulted in sub-selecting 4,252 nodule annotations. Each annotation is considered independently (e.g., one object might be marked by all four radiologists as a nodule and was considered as four independent nodules), the rationale for this assumption being that the exact boundaries of a given nodule

Table 2: Summary of generating binary labels from LIDC rating scales for nodule characteristics.

Nodule characteristics	Label 0	Label 1
Malignancy	Scale 1 - 3 Benign	Scale 4 - 5 Malignant
Sphericity	Scale 1 - 3 Lesser roundness	Scale 4 - 5 High degree of roundness
Margin	Scale 1 - 3 Poorly defined margin	Scale 4 - 5 Sharp margin
Subtlety	Scale 1 - 3 Poor contrast between nodule and surroundings	Scale 4 - 5 High contrast between nodule and surroundings
Texture	Scale 1 - 3 Non-solid internal density	Scale 4 - 5 Solid internal density
Calcification	Scale 1 - 5 Present of calcification	Scale 6 Absent of calcification

lacks universal agreement. Furthermore, following the conventions used by others [30, 26, 31], we wished to maximize use of existing annotations. Uniform labels are assigned to all annotations referring to the same nodule for each feature. As shown in Table 1, the LIDC annotation process employed ordinal scales to label likelihood of malignancy and four semantic features: margin, sphericity, nodule subtlety, and texture (consistency). Scores for these five nodule characteristics were binarized by averaging the scores for each nodule as in [14] and assigning a threshold at 4 to distinguish scores of 1-3 (Label 0) from scores of 4-5 (Label 1). Label 0 typically indicates a benign nodule, poorly defined margin, lesser roundness, poor conspicuity between nodule and surroundings, and a non-solid (ground-glass-like) consistency. Conversely, Label 1 more typically denotes a malignant nodule, sharp margins, higher sphericity, high conspicuity between nodule and surroundings, and solid consistency. Calcification was annotated using a categorical scale from 1 to 6 in which 6 refers to the absence of calcification; here, we averaged ratings for each nodule by majority vote per [30]. Nodules with average ratings of 6 were labeled as absent of calcification pattern (Label 1); all other ratings represented the presence of calcification (Label 0).

Table 3: Label counts for nodule characteristics.

Nodule characteristics	Label 0 (#)	Label 1 (#)	Total (#)
Malignancy	3212	1040	4252
Sphericity	2304	1948	4252
Margin	1640	2612	4252
Subtlety	1570	2682	4252
Texture	518	3734	4252
Calcification	496	3756	4252

The feature "internal structure" was overwhelmingly annotated as soft tissue, thus provided little discriminative information [26] and was excluded from our analysis. Moreover,

the Cancer Imaging Archive (TCIA) reported that an indeterminate subset of cases in the dataset were inconsistently annotated with respect to spiculation and lobulation [32]. As such, we did not consider these two features in our model. Finally, it should be noted that the actual diagnosis of the nodules is not known in the LIDC dataset. For the purposes of this work, the likelihood of malignancy served as the proxy for truth. Table 2 summarizes the generation of the binary labels from LIDC rating scales as described above. Table 3 lists the data counts for each label of the nodule characteristics.

2.3. Data Preprocessing

The LIDC dataset contains a heterogeneous set of scans obtained using various acquisition and reconstruction parameters. To normalize pixel values, all CT scans were first transformed to Hounsfield (HU) scales using the information in the DICOM (Digital Imaging and Communication in Medicine) series header and converted to a range of $(0,1)$ from (-1000, 500 HU). A 3D cube of $40 \times 40 \times 40$ mm were extracted for each candidate. Each cube was centered around the candidate. 40 mm was chosen so that all candidates would be fully contained in this cube as the largest nodules in our subset were maximally 30 mm. We then rescaled each cube to a fixed size of pixels in all three dimensions, resulting in isotropic cubes for all cases. This preprocessing method retained the original relative nodule size information, which is considered useful domain knowledge for the the following tasks.

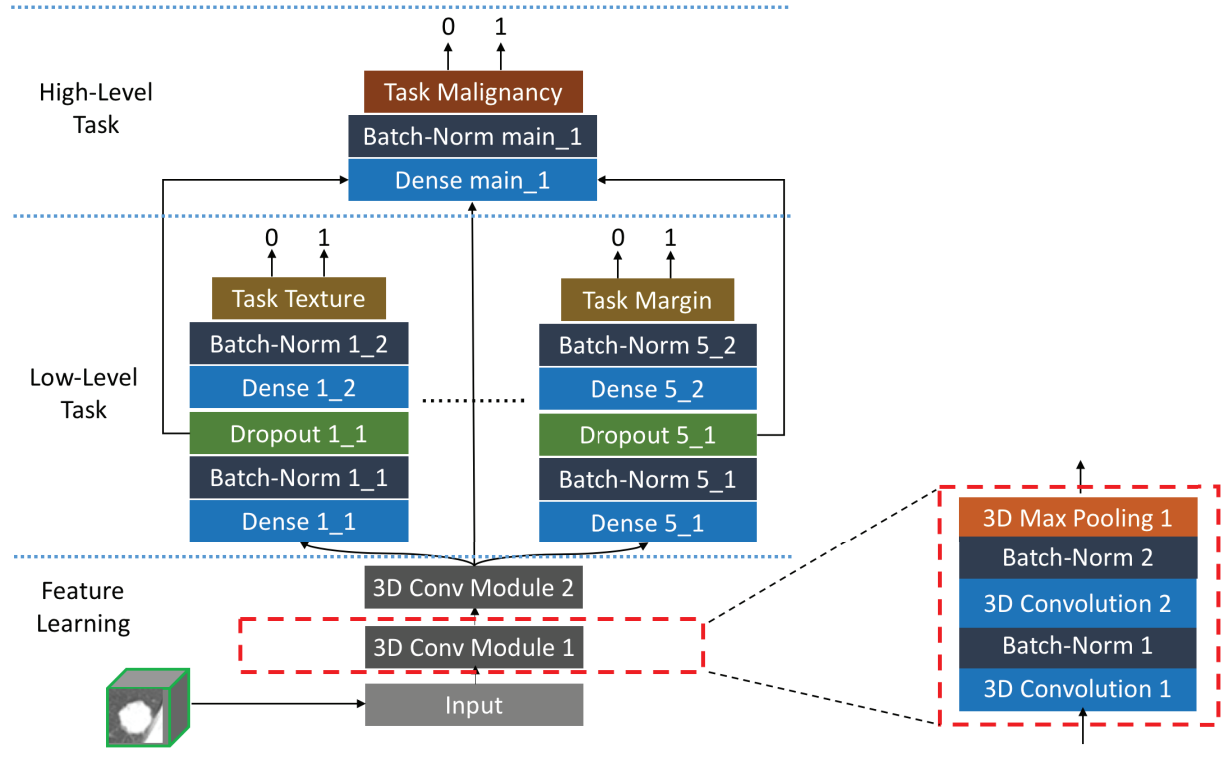## 2.4. Hierarchical Semantic Convolutional Neural Network

Figure 3: Model architecture of the hierarchical semantic convolutional neural network.

The proposed HSCNN utilizes a 3D patch capturing the lung nodule as input and outputs two levels of predictions, as shown in Figure 3. This architecture comprises three parts: 1) a feature learning module; 2) a low-level task module; and 3) a high-level task module. The feature learning component adaptively learns the image features that are generalizable across different tasks. The low-level task predicts five semantic diagnostic features: margin, texture, sphericity, subtlety, and calcification. The high-level task incorporates information from both the generalizable image features and the low-level tasks to produce an overall prediction of lung nodule malignancy.

The feature learning component (Figure 3, feature learning) is the first processing unit of the proposed network. It consists of two convolution module blocks where each block shares the same structure and contains two stacked 3D convolution layers followed by batch normalization and one 3D average pooling layer. Each convolution layer has a kernel size of $3 \times 3 \times 3$. These layers perform the convolution operation on input feature maps along all three dimensions of the input cube to produce an output feature map defined by:

$$f^j = \sum_i c^j * f^i + b^j \tag{1}$$

where f^j and f^i are the jth output feature map and i^{th} input feature map, respectively. And c^j is the jth convolution kernel and $*$ represents the 3D convolution operation between the convolution kernel and input feature map. b^j is the jth bias corresponding to the jth convolution kernel. After convolution, batch normalization is applied to all output feature maps to accelerate the training process and reduce the internal covariate shift by normalizing the feature maps [33]. Rectified linear units (ReLUs) [34] are used as the nonlinear activation functions to take the output from batch normalization. 16 feature maps are used for both convolution layers in the first convolution module, and 32 feature maps are adopted for both convolution layers in the second convolution module. A 3D max pooling layer is used in the end for each convolution module block to progressively reduce the spatial size of the feature maps to reduce the number of parameters and control for overfitting. This layer is defined as:

$$\begin{aligned}
\hat{f}^i_{x,y,z} = max\{&f^i_{x',y',z'}; x' \in [x \cdot s_x,\ x \cdot s_x + d_x - 1], \\
&y' \in [y \cdot s_y,\ y \cdot s_y + d_y - 1], \\
&z' \in [z \cdot s_z,\ z \cdot s_z + d_z - 1]\}
\end{aligned} \tag{2}$$

where x (the row index), y (the column index), and z (the depth index) start from zero. Here, s is the stride size (downscale factor) and d is the size of the max pooling window. We employ a pooling window size of $d = (2, 2, 2)$ and stride size of $s = (2, 2, 2)$. This design downsamples the input feature maps by a factor of 2 across all three cube dimensions. This pooling layer has no learnable parameters.

After the last convolutional module, output features are fed simultaneously into the low- and high-level task components. The low-level task components (Figure 3, low-level task) consist of five branches, each with the same architecture, addressing a distinct semantic feature task (i.e., texture, margin, sphericity, subtlety, or calcification). A fully-connected layer

(densely-connected) is the major basic building block for each of these branches. One fully-connected layer connects each input unit to each output unit, designed to capture correlations from all input feature units to the output. Batch normalization and dropout techniques are both used to control model overfitting. The dropout method randomly removes connections between input and output units during network training to prevent units from co-adapting too much [35]. Two fully-connected layers are employed before the final binary prediction with 256 neurons and 64 neurons for the first and second layer, respectively.

The high-level task component (Figure 3, high-level task) predicts the lung nodule malignancy as the final task. This module concatenates as input the output features from the feature learning component and each of the low-level task branches. As shown in Figure 3, the output feature maps from the last convolution module of the feature learning component are used, along with the output from the last second fully-connected layer of each subtask branch. This design makes the final prediction utilize the basic features learned from the shared convolution modules, and forces the convolution blocks to extract representations that are generalizable across tasks. It also makes use of the information learned from each related explainable subtask to ultimately infer nodule malignancy. The last fully-connected layer in each subtask branch is trained to extract representations more specific to the corresponding subtask compared to the second to last fully-connected layer. Thus, the second to last layer of the subtask branch is chosen to provide less specific but salient information for the final malignancy prediction task. The concatenated features are inputted into a fully-connected layer with 256 neurons, followed by a batch normalization operation before the final malignancy prediction.

To jointly optimize the the HSCNN during the network training, a global loss function is proposed to maximize the probability of predicting the correct label for each task by:

$$L_{global} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{j=1}^{5} \lambda_j \cdot L_{j,i} + L_{M,i}) \tag{3}$$

where N is the total number of training samples and i indicates the ith training sample. j is the jth subtask and $j \in [1,5]$. λ_j is the weighting hyperparameter for the jth subtask. $L_{j,i}$ represents the loss for sample i and task j. $L_{M,i}$ is the loss for the malignancy prediction task for the ith sample. Each loss component is defined as weighted cross entropy loss by:

$$L_{j,i} = -\log (e^{f_{y_i,j}} / \sum_n e^{f_{y_n,j}}) \cdot \omega_{y_i,j} \tag{4}$$

where y_i is true label for the ith sample (x_i, y_i). Here, y_i equals 0 or 1. $f_{y_i,j}$ is the prediction score of the true class y_i for task j and $f_{y_n,j}$ represents a prediction score for class y_n. We use $\omega_{y_i,j}$ to represent the weight of class y_i for task j. The use of $\omega_{y_i,j}$ is important because the labels are imbalanced in all the tasks and $\omega_{y_i,j}$ is helpful in reducing the training bias introduced by such data imbalance. Specifically, $\omega_{y_i,j}$ weights each class loss proportional to the reciprocal of the class counts in the training data. For instance, $\omega_{y_i=0,j} = N_{y_i=1,j}/(N_{y_i=0,j} + N_{y_i=1,j})$ and $\omega_{y_i=1,j} = N_{y_i=0,j}/(N_{y_i=0,j} + N_{y_i=1,j})$. $N_{y_i=1,j}$ represents the total count of samples in the training data for task j, where the true class label equals 1. The global loss function is minimized during the training process by iteratively computing the gradient of L_{global} over the learnable parameters of HSCNN and updates the parameters through back-propagation.

During training, model learnable parameters are initialized using the Xavier algorithm [36] and are updated using the Adam stochastic optimization algorithom [37].

2.5. Training

We performed model training, validation, and testing using 897 LIDC cases, selected as described in Section 2.2. We split these cases into four subsets, where each subset had a similar number of nodules. A 4-fold cross validation study design was employed to obtain the final assessment of the model performance. For each fold, 2 subsets are used for training, 1 subset for validation, and 1 subset for holdout testing. The validation set is used to tune the hyperparameters and test set is employed as external holdout to report the final model performance. Each subset is used as the test set once during the cross validation. This design ensures that the test set is independent of model training and parameter optimizations, and should better reflect the true model performance without information leakage. We note that earlier studies in [14, 15, 18, 19] only use training and validation splits during the cross validation process, without consideration for holdout test sets; such designs arguably have information leakage, and thus tend to over-estimate model performance.

To better control for model overfitting, 3D data augmentation was applied during the training process. Data augmentation artificially inflates the dataset by using label-preserving transforms to add more invariant data examples and is considered as a model regularization scheme [34]. One or more random operations are applied on each training dataset to generate artificial samples. The spatial affine operations used in this study included translating the position of the nodule within 4 mm or flipping the 3D nodule cube along one of the three axes. The translation limit was set to 4 mm to ensure that the boundaries of the largest nodules were captured properly in the 3D cube ($40 \times 40 \times 40$ mm).

3. Experimental Results

This section first describes how we trained the models. We then compared our model to a traditional 3D CNN model and other state-of-art methods. We then assessed the accuracy of semantic feature predictions, providing illustrations of correct and incorrect predictions.

3.1. Model Training

Models were trained for 300 epochs during each fold of cross-validation. After 100 epochs of training, the model loss on the validation set became stable. The best model for each fold was chosen to be the one that achieved the lowest malignancy prediction loss on the validation dataset. Only the independent test dataset was used to calculate end model performance. An online augmentation scheme was employed during model training: during each training epoch, additional artificially created training samples were generated by randomly picking one or multiple augmentation operations, as described in Section 2.5. The same augmentation process was also applied to the validation dataset. To capture a majority of nodule morphology while reducing the input data dimensions, the input nodule cube size was set to be $52 \times 52 \times 52$ voxels. The learning rate was set to be 0.001. The convolution kernel size, number of feature maps, pooling window size, downscale factor, and number of neurons for each fully-connected layer were reported in Section 3. The choices of these parameters have been commonly used, as shown in [34, 38]. The hyperparameters presented in Equation

Table 4: Results comparison: HSCNN versus 3D CNN.

Model	AUC (SD)	Accuracy (SD)	Sensitivity (SD)	Specificity (SD)
3D CNN	0.847 (0.024)	0.834 (0.022)	0.668 (0.040)	0.889 (0.022)
HSCNN	0.856 (0.026)	0.842 (0.025)	0.705 (0.045)	0.889 (0.022)

3 were chosen by using a randomized coarse-to-fine grid search with the validation dataset in the first 20 epochs of each fold [39].

The proposed HSCNN model was implemented in Python 2.7 with TensorFlow [40] and the Keras toolkit [41]. All experiments were performed on a server with 6-core Intel Xeon E5-2630 processor, 32GB memory, and one NVIDIA TITAN Xp GPU (12GB on-board memory). The training of one HSCNN model takes about 5 hours for 300 epochs.
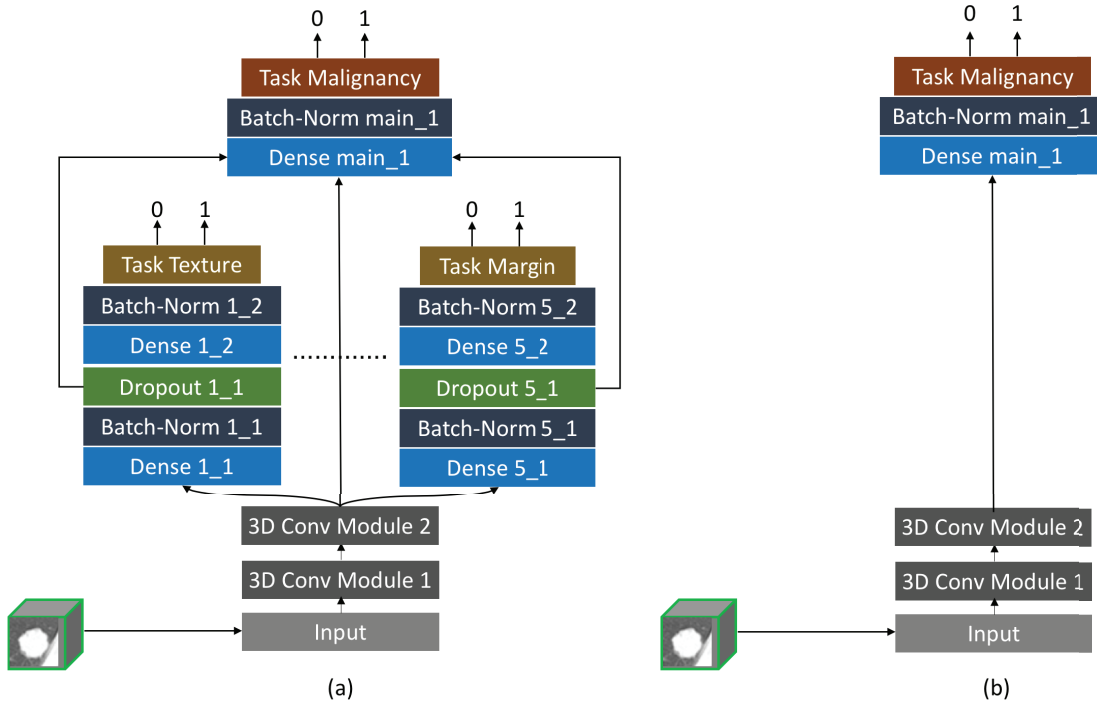
3.2. Malignancy Prediction Results

Figure 4: Framework comparison between proposed HSCNN and baseline 3D CNN. (a) The proposed HSCNN architecture; (b) a baseline 3D CNN architecture. The baseline model has the same structure as the HSCNN but without the low-level semantic task component.

To evaluate and compare the HSCNN performance on lung nodule malignancy prediction, a 3D convolutional neural network (3D_CNN) was first implemented as a baseline model, shown in Figure 4b. This 3D CNN uses the same feature learning and high-level task components architecture as the HSCNN but has the low-level task component removed. The baseline model was trained and evaluated using the same 4-fold cross validation process and with the same data splitting for each fold (using the same randomization seed).

Figure 5 shows the receiver operating characteristic (ROC) curve plots comparing HSCNN versus 3D CNN performance. These plots represent the intuitive trade-off between sensitivity
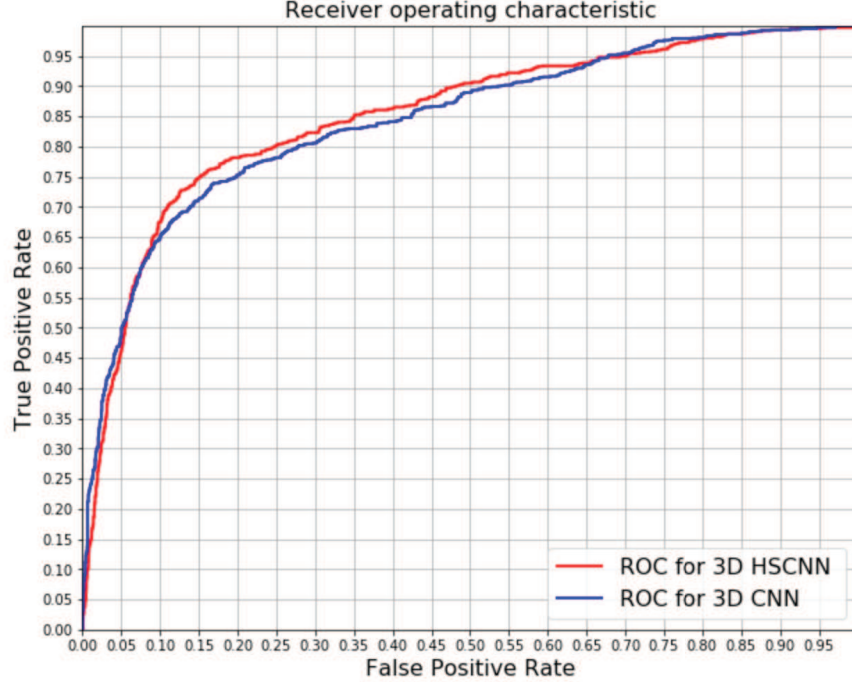
Figure 5: Receiver operating characteristic curve comparison: HSCNN versus 3D CNN. The AUC of 3D HSCNN is significantly higher than 3D CNN according to a paired T-test as shown in Table 5.

and specificity. By visual inspection of the ROC curves, HSCNN performs better than the traditional 3D CNN model. The area under the ROC curve (AUC) quantitatively compares the overall performance of a classification model and is frequently used as a metric to access performance in nodule classification [15, 17, 26, 30, 31]. Table 5 summarizes the mean AUC score, accuracy, sensitivity, and specificity for both models. The HSCNN model achieved a mean AUC 0.856, mean accuracy 0.842, mean sensitivity 0.705 and mean specificity 0.889; while the 3D CNN model achieved a mean AUC 0.847, mean accuracy 0.834, mean sensitivity 0.668 and mean specificity 0.889. Both ROC plots and metric assessments show that the proposed HSCNN achieved better performance for malignancy prediction compared with the conventional 3D CNN approach.

To assess the statistical significance of model performance improvements, we conducted a paired sample t-test to evaluate the mean differences in AUC scores between the HSCNN and 3D CNN model. Group 1 consists of the AUC score of the HSCNN model for each holdout test fold during the cross validation. Group 2 consists of the corresponding AUC score for the 3D CNN for the same fold. The null hypothesis is that the mean difference of AUC score between these two models equals 0. Table 5 summarizes the AUC scores for these groups and results of a paired t-test. The test obtained a p-value of 0.005 and confidence interval of $[0.0051, 0.0129]$, thus rejecting the null hypothesis and indicating that the HSCNN model achieved a statistically significantly better AUC relative to the 3D CNN. The mean improvement of the AUC score was 0.009. This finding demonstrates that adding a low-level task component on an existing CNN structure may improve the prediction of malignancy in a lung nodule.

We also compared our results with current deep learning models for lung nodule malig-

13

Table 5: Paired T-Test summarizes for AUC scores between HSCNN and 3D CNN model on test set of each fold. CI represents for confidence interval.

Test Fold	HSCNN AUC	3D CNN AUC	AUC Difference (HSCNN - 3D_CNN)	Paired T-Test
Fold 1	0.878	0.869	0.009	P-value=0.005, Mean_difference=0.009, CI = [0.0051, 0.0129]
Fold 2	0.813	0.807	0.006	
Fold 3	0.874	0.862	0.012	
Fold 4	0.860	0.851	0.009	

Table 6: Classification performance for semantic feature predictions.

Semantic Features	Accuracy (SD)	AUC (SD)	Specificity (SD)	Sensitivity (SD)
Calcification	0.908 (0.050)	0.930 (0.034)	0.763 (0.092)	0.930 (0.067)
Margin	0.725 (0.049)	0.776 (0.033)	0.632 (0.109)	0.758 (0.091)
Subtlety	0.719 (0.019)	0.803 (0.015)	0.796 (0.045)	0.673 (0.044)
Texture	0.834 (0.086)	0.850 (0.042)	0.636 (0.199)	0.855 (0.108)
Sphericity	0.552 (0.027)	0.568 (0.015)	0.554 (0.076)	0.552 (0.095)

nancy prediction, as reported in the literature to date. Kumar et al. [18] developed a deep autoencoder-based model with 4,323 nodules of the LIDC dataset, achieving model accuracy of 0.7501. Hua et al. [19] presented a CNN model and deep belief network (DBN) model. Both models were trained and validated using 2,545 lung nodule samples from LIDC. The CNN model had specificity of 0.787 and sensitivity 0.737; and the DBN model obtained specificity of 0.822 and sensitivity 0.734. In [14], Shen et al. used a model based on multi-scale 3D CNN. Developed with 1,375 LIDC nodule samples, the average accuracy is reported above 0.84 with different configurations. In [15], Shen et al. extended this multi-scale model using a multi-crop approach and achieved accuracy of 0.839, 0.8636, and 0.8714 with 340, 1030 and 1375 nodules of LIDC, respectively. All of these previously reported methods were evaluated with only training and validation data splits without an independent holdout test dataset as discussed in Section 2.5. This evaluation design might lead to information leakage due to the use of the validation data for optimal model parameters selection and may overestimate the model performance. In general, our model achieved better or similar performances compared with these reported methods.

3.3. Semantic Feature Prediction Results and Model Interpretability

Table 6 presents the classification performance for each of the low-level tasks (i.e., semantic features). We achieved mean accuracy of 0.908, 0.725, 0.719, 0.834 and 0.552; mean AUC score of 0.930, 0.776, 0.803, 0.850 and 0.568; mean sensitivity of 0.930, 0.758, 0.673, 0.855 and 0.552; and mean specificity of 0.763, 0.632, 0.796, 0.636 and 0.554 for calcification, margin, subtlety, texture, and sphericity, respectively. These results suggest that the HSCNN model is able to learn feature representations that are predictive of semantic features while simultaneously achieving high performance in predicting nodule malignancy.

Figure 6 demonstrates the interpretability of the HSCNN model by visualizing the central slices of the 3D nodule patches in axial, coronal, and sagittal projections while presenting
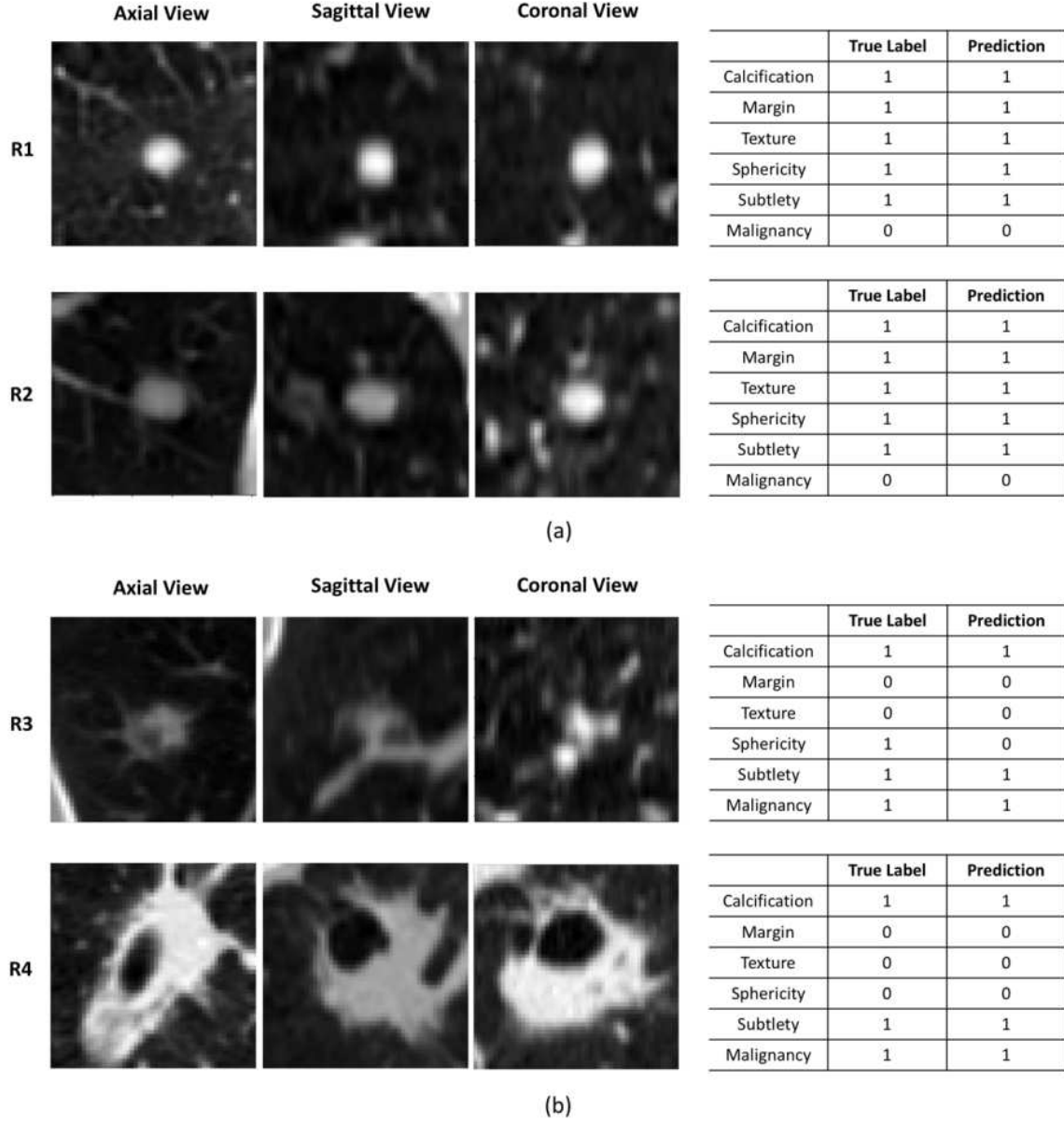
Figure 6: Illustrating the HSCNN model interpretability: lung nodule central slices, interpretable semantic feature prediction and malignancy prediction. R1, R2, R3 and R4 are four different nodules. (a) Central slices of axial, coronal and sagittal view of two benign nodule samples; true and predicted labels for interpretable semantic features and malignancy. (b) Central slices of axial, coronal and sagittal view of two malignant nodule samples; true and predicted labels for interpretable semantic features and malignancy.

the predicted interpretable semantic labels along with the malignancy classification results. Figure 6a-R1 shows that the HSCNN model classifies the lung nodule as benign (the true label is also benign). This decision correlated to predictions of this nodule as having no calcification, sharp margins, roundness, obvious contrast between nodule and surroundings, and solid consistency. The predictions of these five semantic characteristics are the same as the true label and corresponds to our knowledge about benign lung nodules. Compared to a 3D CNN malignancy prediction model, the HSCNN provides more insight for interpreting
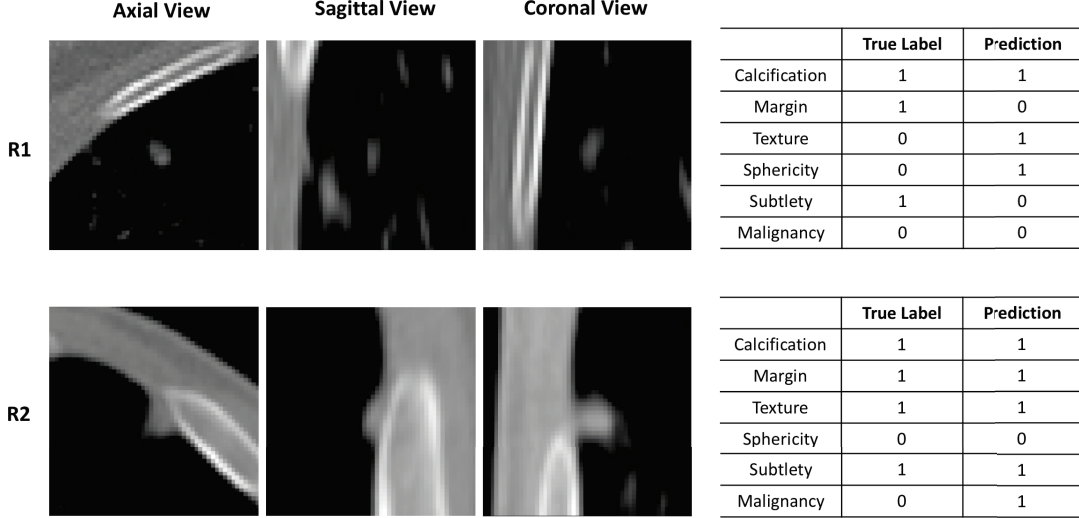
	Axial View	Sagittal View	Coronal View		True Label	Prediction
				Calcification	1	1
				Margin	1	0
R1				Texture	0	1
				Sphericity	0	1
				Subtlety	1	0
				Malignancy	0	0

			True Label	Prediction
		Calcification	1	1
		Margin	1	1
R2		Texture	1	1
		Sphericity	0	0
		Subtlety	1	1
		Malignancy	0	1

Figure 7: Example cases where the HSCNN model incorrectly predicts semantic features or cancer malignancy. R1 and R2 are two different nodules. R1: This case has four incorrect semantic feature predictions yet a correct malignancy prediction. R2: This case has all correct semantic predictions yet an incorrect malignancy prediction.

its predictions. Similarly, in Figure 6b-R3, the proposed model predicts the lung nodule as malignant (true label is also malignant). Different from the benign case, the HSCNN model predicts this nodule having poorly defined margins, ground glass consistency, and non-round shape. This partly explains why the HSCNN makes a malignancy classification with such nodule characteristics corresponding to our expert knowledge about typical malignant nodules. We note that the sphericity predictions made by the model are different from the true label. This result is explained by the fact that while the nodule has a more regular round shape in axial view, the shape is actually more elongated in the two other projections, as shown in Figure 6b-R3.

Figure 7 shows two representative cases where the HSCNN fails to predict either one or more semantic features or cancer malignancy. Figure 7-R1 shows that the HSCNN model classifies the lung nodule correctly as benign but incorrectly for four semantic features of this nodule (margin, texture, sphericity and subtlety). In Figure 7-R2, the HSCNN model incorrectly classifies the lung nodule as malignant (the true label is benign). However, all semantic features of this nodule are predicted correctly. These two cases present the situation where the correctness are inconsistent between the malignancy and semantic predictions. Section 4 provides further discussion about how the HSCNN model can be augmented with more semantic features.

4. Discussion

We present a HSCNN model that incorporates domain knowledge into the model architecture design, predicting semantic nodule characteristics along with the primary task of nodule malignancy diagnosis. Five semantic features were considered: calcification, margin, subtlety, texture, and sphericity. Our results in Section 3.3 suggest that the HSCNN

16

model quantifies these nodule characteristics in a fully data-driven way yet in one integrated model that predicts nodule malignancy. The semantic labels are useful in interpreting the model's predictions for malignancy, mapping the features used by the network to predict the high-level task with established domain knowledge about pulmonary nodules. Section 3.2 shows that this design of the HSCNN also improves the model performance for lung cancer malignancy classification. This study focuses mainly on exploring the values that are added by incorporating the low-level semantic task component design into the CNN architecture. Further optimization of the network architecture to achieve higher prediction performance can be performed. For instance, densely connected designs [42] and residual designs [43] could be used to potentially improve model performance. But due to limitations of computation power, not all designs are optimally searched; we will investigate these as part of future work.

There are some limitations to this study. Our semantic labels did not include those of known higher association with malignancy, such as nodule size, margin spiculation, lobulation, and anatomic location, which have previously been reported as informative [44, 45]. In the case of lobulation and spiculation, known labeling errors in the LIDC dataset made them unsuitable for our use. Additionally, semantic labels are subject to moderate inter-reader variability; performance might be enhanced by limiting semantic labels to those on which there is high reader agreement. Third, the malignancy labels provided in the LIDC dataset do not reflect pathological diagnosis but rather, suspicion levels of the interpreting radiologists. Finally, the original semantic features have scales of 5 or 6; binarizing the labels may lose some of the semantic information. Changing the threshold for binary classification would also affect results. Our rationale for binary labels in this case was to overcome data sparsity, where the number of cases labeled for certain scales might be very small compared with the other scales (e.g., only 11 cases are labeled as linear for sphericity out of total 4252 cases). Moreover, analysis shows that inter-reader agreement is much lower for 5 or 6 scales compared with the proposed binary labels. Thus, binary labeling helps to reduce labeling noise caused by inter-reader variability. These limitations may be circumvented by training on large datasets that have been systematically annotated using a shared lexicon that includes discriminating features. In future work we will explore model improvement by including discriminating semantic features, and investigate model variability using different semantic labeling schemes. Although only five subtask modules are presented for the HSCNN architecture in this paper, the HSCNN framework and global loss function could be easily extended to increase or decrease the number of low-level semantic features. This study also paves the way to apply this idea in other disease domain to build interpretable models.

5. Conclusion

In this paper, we have developed a novel radiologist-interpretable HSCNN model for predicting lung cancer in CT-detected indeterminate nodules. This model is able to simultaneously predict nodule malignancy while classifying five nodule semantic characteristics, including calcification, margin, subtlety, texture, and sphericity of nodules. These diagnostic semantic features predictions are intermediate outputs associated with the final malignancy prediction, and are useful to explain the diagnosis prediction. Information from each low-

level semantic feature prediction is incorporated into the malignancy prediction task by employing jump connections. This framework is able to enforce the shared basic convolution modules in the HSCNN to learn features that are generalizable across tasks. This unified model is trained by minimizing a joint global loss function, where the losses of both malignancy and semantic feature prediction tasks are incorporated. Extensive experiments and statistical tests show that the proposed HSCNN model is able to significantly improve the classification performance for nodule malignancy prediction and the semantic characteristics predictions have improved the model interpretability. This trained model could also serve as a lung nodule semantic feature generator.

Author Contributions

All authors contributed to the development of the project. SS developed the methodology, conducted the experiments and wrote the manuscript. SXH contributed to the experiments. AATB and DRA provided valuable advice and domain input. WH provided oversight over the project and contributed to its design. All authors reviewed the manuscript.

Acknowledgement

References

References

[1] L. A. Torre, R. L. Siegel, A. Jemal, Lung cancer statistics, in: Lung Cancer and Personalized Medicine, Springer, 2016, pp. 1–19.

[2] S. Shen, S. X. Han, P. Petousis, R. E. Weiss, F. Meng, A. A. Bui, W. Hsu, A bayesian model for estimating multi-state disease progression, Computers in biology and medicine 81 (2017) 111–120.

[3] N. L. S. T. R. Team, et al., Reduced lung-cancer mortality with low-dose computed tomographic screening, N Engl J Med 2011 (2011) 395–409.

[4] K. ten Haaf, J. Jeon, M. C. Tammemägi, S. S. Han, C. Y. Kong, S. K. Plevritis, E. J. Feuer, H. J. de Koning, E. W. Steyerberg, R. Meza, Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study, PLoS medicine 14 (2017) e1002277.

[5] B. Zhao, Y. Tan, D. J. Bell, S. E. Marley, P. Guo, H. Mann, M. L. Scott, L. H. Schwartz, D. C. Ghiorghiu, Exploring intra-and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on ct scans reconstructed at different slice intervals, European journal of radiology 82 (2013) 959–968.

[6] S. G. Armato, M. B. Altman, J. Wilkie, S. Sone, F. Li, A. S. Roy, et al., Automated lung nodule classification following automated nodule detection on ct: A serial approach, Medical Physics 30 (2003) 1188–1197.

[7] S. Shen, A. A. Bui, J. Cong, W. Hsu, An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy, Computers in biology and medicine 57 (2015) 139–149.

[8] N. Duggan, E. Bae, S. Shen, W. Hsu, A. Bui, E. Jones, M. Glavin, L. Vese, A technique for lung nodule candidate detection in ct using global minimization methods, in: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, pp. 478–491.

[9] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, R. Valentim, Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy, Biomedical engineering online 15 (2016) 2.

[10] G. J. Amir, H. P. Lehmann, After detection:: The improved accuracy of lung cancer assessment using radiologic computer-aided diagnosis, Academic radiology 23 (2016) 186–191.

[11] P. Huang, S. Park, R. Yan, J. Lee, L. C. Chu, C. T. Lin, A. Hussien, J. Rathmell, B. Thomas, C. Chen, et al., Added value of computer-aided ct image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study, Radiology 286 (2017) 286–295.

[12] D. Zinovev, J. Feigenbaum, J. Furst, D. Raicu, Probabilistic lung nodule classification with belief decision trees, in: Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE, IEEE, pp. 4493–4498.

[13] T. W. Way, B. Sahiner, H.-P. Chan, L. Hadjiiski, P. N. Cascade, A. Chughtai, N. Bogot, E. Kazerooni, Computer-aided diagnosis of pulmonary nodules on ct scans: Improvement of classification performance with nodule surface features, Medical physics 36 (2009) 3086–3098.

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

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

[16] E. A. R. Piedra, R. K. Taira, S. El-Saden, B. M. Ellingson, A. A. Bui, W. Hsu, Assessing variability in brain tumor segmentation to improve volumetric accuracy and characterization of change, in: Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on, IEEE, pp. 380–383.

[17] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, B. van Ginneken, Automatic classification of pulmonary perifissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box, Medical image analysis 26 (2015) 195–202.

[18] D. Kumar, A. Wong, D. A. Clausi, Lung nodule classification using deep features in ct images, in: Computer and Robot Vision (CRV), 2015 12th Conference on, IEEE, pp. 133–138.

[19] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, Y.-J. Chen, Computer-aided classification of lung nodules on computed tomography images via deep learning technique, OncoTargets and therapy 8 (2015).

[20] A. Farag, A. Ali, J. Graham, A. Farag, S. Elshazly, R. Falk, Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose ct scans of the chest, in: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE, pp. 169–172.

[21] P.-L. Lin, P.-W. Huang, C.-H. Lee, M.-T. Wu, Automatic classification for solitary pulmonary nodule in ct image by fractal analysis based on fractional brownian motion model, Pattern Recognition 46 (2013) 3279–3287.

[22] W. Jorritsma, F. Cnossen, P. van Ooijen, Improving the radiologist–cad interaction: designing for appropriate trust, Clinical radiology 70 (2015) 115–122.

[23] H. Kim, C. M. Park, J. M. Goo, J. E. Wildberger, H.-U. Kauczor, Quantitative computed tomography imaging biomarkers in the diagnosis and management of lung cancer, Investigative radiology 50 (2015) 571–583.

[24] J. J. Erasmus, J. E. Connolly, H. P. McAdams, V. L. Roggli, Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions, Radiographics 20 (2000) 43–58.

[25] A. Kaya, A. B. Can, A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics, Journal of biomedical informatics 56 (2015) 69–79.

[26] M. C. Hancock, J. F. Magnan, Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods, Journal of Medical Imaging 3 (2016) 044504.

[27] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffmoan, 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 38 (2011) 915–931.

[28] M. F. McNitt-Gray, S. G. Armato, C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, P. H. Bland, et al., The lung image database consortium (lidc) data collection process for nodule detection and annotation, Academic radiology 14 (2007) 1464–1474.

[29] A. P. Reeves, A. M. Biancardi, The lung image database consortium (lidc) nodule size report, http://www.via.cornell.edu/lidc/, 2011. Release: 2011-10-27-2.

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., The cancer imaging archive (tcia): maintaining and operating a public information repository, Journal of digital imaging 26 (2013) 1045–1057.

[31] B. R. Froz, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, M. Gattass, Lung nodule classification using artificial crawlers, directional texture and support vector machine, Expert Systems with Applications 69 (2017) 176–188.

[32] The Cancer Imaging Archive, Lung image database consortium - reader annotation and markup - annotation and markup issues/comments, https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI, 2017.

[33] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[34] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (2014) 1929–1958.

[36] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

[37] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014).

[39] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning., in: OSDI, volume 16, pp. 265–283.

[41] F. Chollet, et al., Keras, 2015.

[42] G. Huang, Z. Liu, K. Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, arXiv preprint arXiv:1608.06993 (2016).

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

[44] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, et al., Probability of cancer in pulmonary nodules detected on first screening ct, New England Journal of Medicine 369 (2013) 910–919.

[45] S. J. Swensen, M. D. Silverstein, D. M. Ilstrup, C. D. Schleck, E. S. Edell, The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules, Archives of internal medicine 157 (1997) 849–855.

Conflict of Interest Statement

None Declared.