This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Explainable Hierarchical Semantic Convolutional Neural Network for Lung Cancer Diagnosis

Shiwen Shen[1], Simon X Han[2], Denise R Aberle[2], Alex A Bui[2], William Hsu[2]
[1] JD Digits, [2] Medical & Imaging Informatics Group, University of California, Los Angeles

Abstract

While deep learning methods have demonstrated classification performance comparable to human readers in 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." 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 semantic features; and 2) a high-level prediction of nodule malignancy. The low-level outputs reflect diagnostic features often reported by radiologists and serve to explain how the model interprets the images in an expert-interpretable manner. The information from these low-level outputs, along with the representations learned by the convolutional layers, are then combined and used to infer the high-level output. 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 comparable results with the state-of-the-art methods.

1. Introduction and Background

Lung cancer is the leading cause of cancer mortality worldwide. Computed tomography (CT) imaging is the *de facto* modality to early detect and characterize pulmonary nodules. However, some studies indicate that the false positive rate for low-dose CT is upwards of 20%, resulting in unnecessary medical, economic, and psychological costs. Moreover, detection rates vary among less experienced radiologists, particularly in subtle cases, as interpretation heavily relies on past experience. Figure 1 illustrates examples of malignant (top row, R1) and benign (bottom row, R2) nodules. The visual appearance of these nodules is highly varied with subtle differences in size, shape, and texture, underscoring the challenge faced by radiologists in differentiating between the two categories.

In response, computer-aided diagnosis (CADx) systems

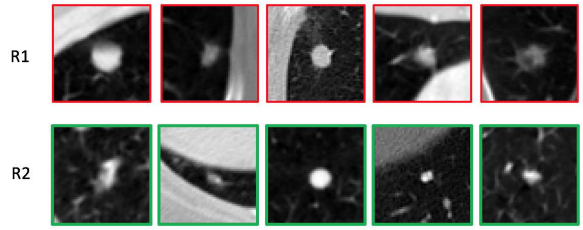


Figure 1. Illustrations of malignant and benign nodules: R1 are malignant nodules; R2 are benign nodules.

[1] are being developed to help distinguish between malignant from benign nodules. Deep learning methods [2, 3, 4, 5], particularly convolutional neural networks (CNNs), have been used for lung nodule classification, with promising results. Markedly, these works use deep learning as a "black box" and do not attempt to explain what representations have been learned or why the model generates a given prediction. This low degree of interpretability arguably hinders target end users, such as radiologists, from understanding how the models work and ultimately impedes model adoption for clinical usage. In contrast, a number of radiologist-interpreted features derived from CT scans have been considered influential when assessing the malignancy of a lung nodule [6]. These features are referred to as *semantic* features in this study. Examples of such semantic features include nodule consistency (texture) and shape. These features are intuitive to radiologists and are moderately robust against perturbations in image resolution and reconstruction kernel.

In this study, we demonstrate a novel interpretable hierarchical semantic convolutional neural network (HSCNN) that predicts whether a nodule is malignant in CT images. The HSCNN 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.

2. Materials and Methods

2.1. Dataset

We trained and evaluated the model using the publicly available Lung Image Database Consortium dataset (LIDC-IDRI) [7]. LIDC-IDRI contains 1,018 CT scans with nodule diameters ranging from 3-30 mm; each CT scan was annotated by four human readers. Nodules $\geq$ 3mm were contoured at the pixel-level in 3D by each radiologist then assigned labels related to likelihood of malignancy and semantic characteristics.

We considered five semantic characteristics: calcification, subtlety, sphericity, margin, and texture. Each feature was rated from 1 to 5 or 6 by each reader. Calcification indicates the presence and patter of calcification in the nodule. The categorical value from 1 to 6 means popcorn, laminated, solid, non-central, central and absent pattern, respectively. Subtlety defines the level of difficulty of detecting the nodule relative to surrounding. Value from 1 to 5 represents the degree from extremely subtle to obvious. Sphericity presents the nodule three dimensional shape in terms of roundness. Value 1, 3 and 5 indicate linear, ovoid and round shape, respectively. Margin feature shows how well defined the margins are. Value 1 means poorly defined margin and value 5 represents shape margin. Finally, texture describes the nodule internal texture consistency. Value 1, 3 and 5 represents non-solid, part solid and solid nodule, respectively.

Only nodules identified by at least three radiologists were included in this study. CT scans with slice thickness larger than or equal to 3 mm were also excluded, resulting in 897 LIDC scans with 4,252 nodule annotations. The LIDC annotation process employed one ordinal feature (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 [4] and then binarizing each feature: average scores between 1-3 were assigned Label 0 while 4-5 were assigned Label 1.

2.2. Hierarchical Semantic Convolutional Neural Network

The proposed HSCNN utilizes a 3D patch centered on the lung nodule as input and outputs two levels of predictions, as shown in Figure 2. 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 module adaptively learns the image features that are generalizable across different tasks. The low-level task predicts five semantic diagnostic features: margin, texture,

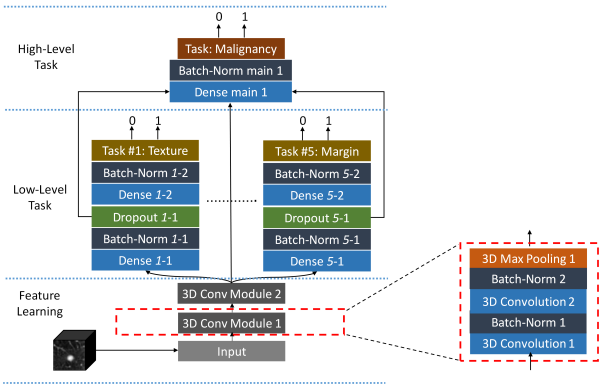


Figure 2. Model architecture of the hierarchical semantic convolutional neural network.

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 module (Figure 2, feature learning) 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 $3 \times 3 \times 3$. Rectified linear units (ReLUs) are used as the nonlinear activation functions. 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 to progressively reduce the spatial size of the feature maps by a half in all three dimensions.

After the last convolutional module, output features are fed simultaneously into the low- and high-level task modules. The low-level task module (Figure 2, low-level task) consists of five branches, each with the same architecture, representing a distinct semantic feature (i.e., texture, margin, sphericity, subtlety, or calcification). Fully-connected layers (densely-connected) are used 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 is applied. The dropout method are employed between fully-connected layers. 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 module (Figure 2, high-level task) predicts whether the nodule is malignant. This module combines the output features from the feature learning module and each of the low-level task branches as its input. As shown in Figure 2, the output feature maps from the last convolution module 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 all tasks. It also makes use of the information learned from each related semantic 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.

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

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

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

$$L_{j,i} = -\log\left(e^{f_{y_i,j}} / \sum_n e^{f_{y_n,j}}\right) \cdot \omega_{y_i,j} \quad (2)$$

where y_i is true label for the i^{th} 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.

3. Experimental Results

We performed model training, validation, and testing using the 897 LIDC scans. A 4-fold cross validation study design was employed to obtain the final assessment of the model performance. Within each fold, we split these cases into four subsets, where each subset had a similar number of nodules. 2 subsets are used for training, 1 subset for validation, and 1 subset for holdout testing. To better control for model overfitting, 3D data augmentation was applied during the training process.

3.1. Malignancy Prediction Results

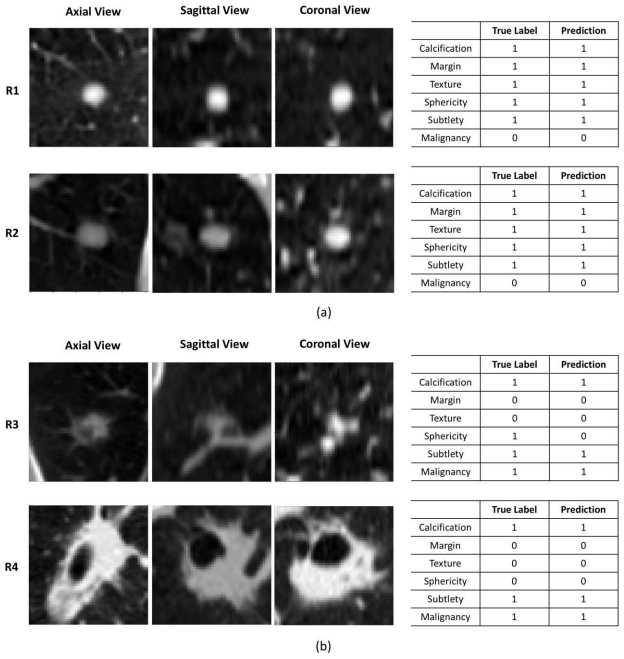
To evaluate and compare the HSCNN performance on lung nodule malignancy prediction, a 3D convolutional neural network (3D_CNN) was implemented as a baseline model. This 3D CNN uses the same feature learning and high-level task modules as the HSCNN but do not include the low-level subtask module. The baseline model was trained and evaluated using the same data and settings. 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. The metric assessments show that the proposed HSCNN achieved better performance for malignancy prediction compared with the conventional 3D CNN approach.

We also compared our results with other deep learning models for lung nodule malignancy prediction that utilized the LIDC dataset reported in literature to date. Kumar et al. [3] developed a deep autoencoder-based model with 4,323 nodules from LIDC, achieving model accuracy of 0.7501. Hua et al. [2] presented a CNN model and deep belief network (DBN) model using 2,545 LIDC nodule samples. 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 [4], Shen et al. used a model based on multi-scale 3D CNN. Developed with 1,375 LIDC nodule samples, the average accuracy is reported 0.84. All of these previously reported methods were evaluated with only training and validation data splits without an independent holdout test dataset. Generally, our model achieved better or similar performances compared with these reported methods. However, direct comparison of these models is difficult given that each model was trained and tested on different subsets of the LIDC dataset.

3.2. Semantic Feature Prediction Results and Model Interpretability

For the classification performance for each of the low-level tasks, 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 3 demonstrates the interpretability of the HSCNN



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

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

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

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

Figure 3. Illustrating the HSCNN model interpretability: lung nodule central slices, 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) Two malignant nodule samples.

model by visualizing the central slices of the 3D nodule patches in axial, coronal, and sagittal projections while presenting the predicted interpretable semantic labels along with the malignancy classification results. Figure 3a-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 its predictions. Similarly, in Figure 3b-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 3b-R3.

4. Conclusion

In summary, we developed a novel radiologist-interpretable HSCNN model for predicting whether an (indeterminate) nodule is malignant. This model simultaneously predicts nodule malignancy and five 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 model's prediction of nodule malignancy. Our network architecture provides a way to create a mapping between semantic features with which radiologists are familiar and deep features that are learned by the model from the data. Results from the low-level tasks can be used to automatically pre-populate a radiologist report or provide context to the radiologist on the model's overall prediction of malignancy.

References

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

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

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

[4] 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*, pp. 588–599, Springer, 2015.

[5] S. A. El-Regaily, M. A. Salem, M. H. Abdel Aziz, and M. I. Roushdy, "Survey of computer aided detection systems for lung cancer in computed tomography," *Current Medical Imaging Reviews*, vol. 14, no. 1, pp. 3–18, 2018.

[6] H. Kim, C. M. Park, J. M. Goo, J. E. Wildberger, and H.-U. Kauczor, "Quantitative computed tomography imaging biomarkers in the diagnosis and management of lung cancer," *Investigative radiology*, vol. 50, no. 9, pp. 571–583, 2015.

[7] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, ..., and E. Kazerooni, "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.