A Lightweight Multi-Section CNN for Lung Nodule Classification and Malignancy Estimation

Pranjal Sahu, Dantong Yu, Mallesham Dasari, Fei Hou, and Hong Qin

***Abstract*—The size and shape of a nodule are the essential indicators of malignancy in lung cancer diagnosis. However, effectively capturing the nodule's structural information from CT scans in a computer-aided system is a challenging task. Unlike previous models that proposed computationally intensive deep ensemble models or three-dimensional CNN models, we propose a lightweight, multiple view sampling based multi-section CNN architecture. The model obtains a nodule's cross sections from multiple view angles and encodes the nodule's volumetric information into a compact representation by aggregating information from its different cross sections via a view pooling layer. The compact feature is subsequently used for the task of nodule classification. The method does not require the nodule's spatial annotation and works directly on the cross sections generated from volume enclosing the nodule. We evaluated the proposed method on lung image database consortium (LIDC) and image database resource initiative (IDRI) dataset. It achieved the state-of-the-art performance with a mean 93.18% classification accuracy. The architecture could also be used to select the representative cross sections determining the nodule's malignancy that facilitates in the interpretation of results. Because of being lightweight, the model could be ported to mobile devices, which brings the power of artificial intelligence (AI) driven application directly into the practitioner's hand.**

***Index Terms*—Lung cancer, deep learning, nodule classification, transfer Learning, spherical sampling.**

I. Introduction

AS OF 2017, lung cancer accounts for the most number of deaths in the world [1]. Typically, early diagnosis of lung cancer relies on accurately detecting the presence of lung nodules in CT scans [2]. A lung nodule is a structure with approximately 3 mm in diameter and classified as benign or malignant [3]. For this purpose radiologists read chest CT slice by slice. However, this process is not only laborious but also error-prone.

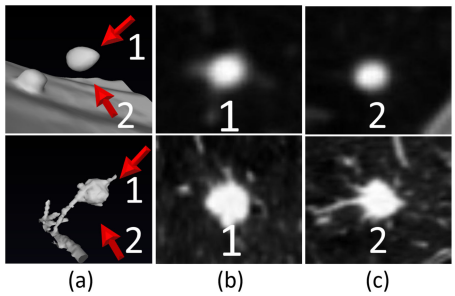


Fig. 1. Figure explaining how different cross-sections of a nodule from view 1 and 2 can capture the spiculations effectively. In Fig. (a), (b) and (c) upper row is benign and lower is malignant. The spiculations on malignant one become clear once the view point is altered as seen in view 2.

Inspecting the 3-D lung nodule slice by slice deems to be insufficient and results in information loss and missing diagnosis. This can be easily understood by looking at the two samples of lung nodules in Figure 1: one benign case and one malignant one, both appear to be similar to each other at one view angle. Once the view angle is altered, their structural differences that determine the diagnosis become evident. Therefore, incorporating the full 3D information of a nodule is necessary.

To help the radiologists in the task of lung nodule classification, numerous computer-aided diagnosis (CAD) methodologies have been proposed in the past [4]–[9]. Traditional approaches used handcrafted features for this purpose. For example Han *et al.* in [5] used Haarlick texture features for nodule classification, while Jacobs *et al.* in [6] used a combination of texture, shape, and context features to classify lung nodules. Similarly, Murphy *et al.* in [8] used a combination of handcrafted features summarizing the structural properties of the nodule with a kNN classifier. These methods predominantly have utilized the 2D slices from CT-scans instead of the available 3-D volumetric data with few exceptions such as in [10] and [4]. In [10], Way *et al.* used active 3D contours for nodule segmentation and extracted morphological features, such as volume, surface area etc. along with texture features from the axial planes of nodule for classification. Similarly in [4], authors utilized Spherical Harmonics for describing lung nodule's shape complexity and used it with a K-nearest classifier for distinguishing malignant and benign lung nodules.

The recent successes of Deep Learning in computer vision showed superior performance of data-driven methods over those

Manuscript received May 16, 2018; revised September 22, 2018; accepted October 26, 2018. Date of publication November 6, 2018; date of current version May 6, 2019. This work was supported by the Brookhaven National Laboratory and National Science Foundation under Grant NSF IIS-1715985. *(Corresponding author: Hong Qin.)*

P. Sahu, M. Dasari, and H. Qin are with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: psahu@cs.stonybrook.edu; mdasari@cs.stonybrook.edu; qin@cs.stonybrook.edu).

D. Yu is with the Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, NJ 07103 USA (e-mail: dtyu@njit.edu).

F. Hou is with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science, Beijing 100190, China (e-mail: houfei@ios.ac.cn).

Digital Object Identifier 10.1109/JBHI.2018.2879834

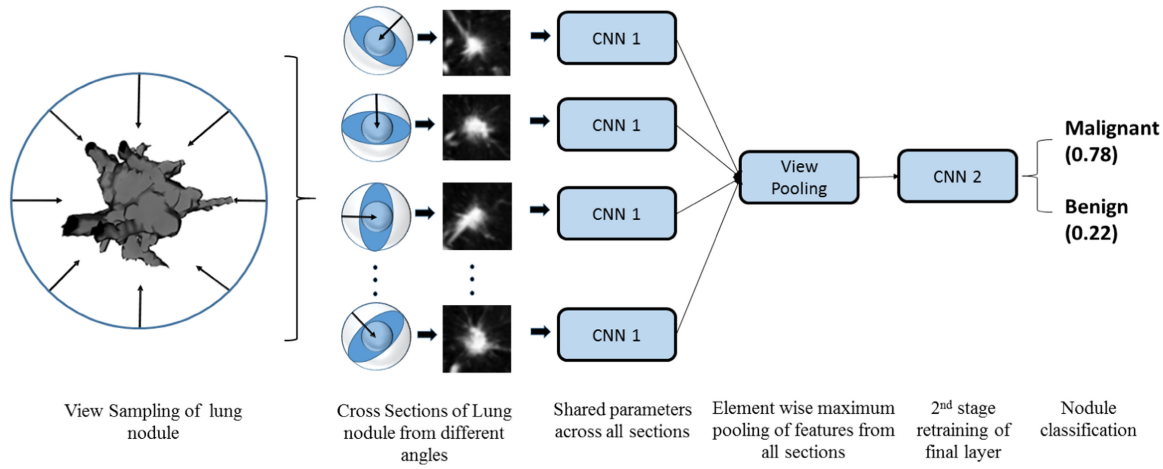


Fig. 2. The pipeline of the Multi-section CNN based on spherical sampling.

handcrafted features for a variety of tasks. Several attempts [11]–[16] introduced deep CNNs in nodule classification and showed impressive performance. However, the existing methods are computation-intensive, require considerable manual intervention in the form of spatial annotation, and fail to maximize the strength of a full 3D volumetric data. In many cases, even though these approaches provide the final diagnosis result, they fall short of providing the causal factors that are easily interpretable by practitioners. For example, Shen *et al.* [12] adopted a multi-scale 2D CNN approach for lung nodule classification, where they adopt multi-scale nodule patches and learn class specific features by concatenating feature responses from the last layer for each scale. However, their approach ignored the full 3D volumetric information. In [17] authors proposed a 3D dual path network using 3D CNN for lung nodule detection and feature extraction for classification using GBM (Gradient Boosting Machine). However, these methods fail to provide any causal factors for the obtained results. Arnaud *et al.* in [18] extracted nine different patches from the volume of a lung nodule: three patches from the sagittal, coronal, and axial planes and the remaining six from the diagonal planes of symmetry. However, a severe limitation of the method is that it requires training nine separate deep networks and needs a significant amount of computational resources. Recently, Xie *et al.* [19] proposed an ensemble approach based on transfer learning. Although this method produced promising results, it had significant limitations, since it relied on accurate delineation among lung nodules from CT scans that is costly to obtain because such a delineation requires an expert's supervision.

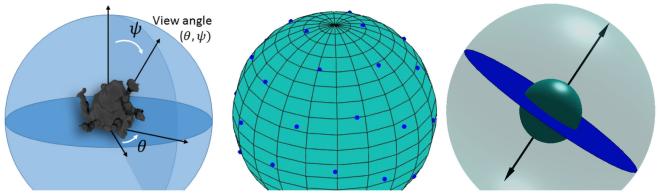
In contrast to the related works and methods, we propose a single-stack CNN that is lightweight, utilizes a full 3D volume of a nodule and generates the interpretable diagnosis. Two previous studies in [20] and [21] heavily inspired our design choices of the neural network. In the first related work [20], the authors introduced a handcrafted feature called "Ipris" that represents the intensity and gradient transitions from the center to peripheral of a nodule and serves as a good indicator of its malignancy. The other related work [21] demonstrated that under a limited number of training samples and computation resources, a CNN for object recognition based on the 2D rendering of 3D objects shows a better discriminative ability than the one with the entire 3D volume. We incorporate these two findings in our model by introducing a Multi-section CNN architecture. Our contributions in this paper are:

- A lightweight Multi-section CNN architecture for obtaining a compact representation of a lung nodule from its volumetric data for classification and malignancy estimation.
- Evaluation of sampling approaches to extract a nodule's cross-sections to capture the intensity and gradient transitions going from internal to external of a nodule in a data-driven automated manner.
- Nodule's salient section detection to assist clinical practitioners in highlighting the causal factor i.e., crucial cross-section and thereby interpreting the diagnosis results.

The proposed model being lightweight can be easily ported to mobile devices such as tablets etc. We also perform experiments on a mobile device to compare the inference latency of Multi-section CNN with other state-of-the-art method and demonstrate the efficient design of our model. The design of Multi-section CNN is described in detail in next section.

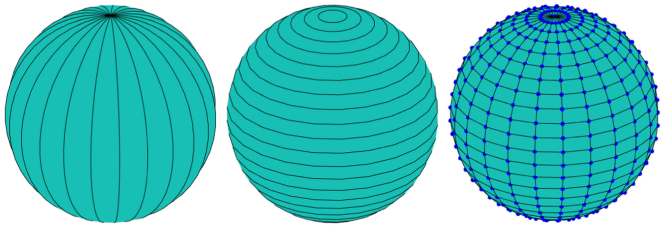
II. Multi-Section CNN

The central idea in this paper is to aggregate information from multiple cross-sections in a data-driven manner to fully exploit the information contained in a nodule's volume. For this purpose, we introduce a Multi-section CNN architecture that is shown in Figure 2. To obtain the Multi-section representations, the first step is to extract multiple cross-sections from a lung nodule for training deep neural networks. Unlike natural objects, a lung nodule is orientation-agnostic; therefore a sampling approach which is rotation invariant is required. Once the cross-sections are obtained from the sampling approach the Multi-section CNN is trained in two stages. We detail the sampling approach, CNN architecture and training, inference methodology in the following sub-sections.



(a) Spherical sampling views (b) t-design views (c) Two opposite view angles.

Fig. 3. Two sampling approaches are used in our experiments to obtain the multiple cross-sections of the lung nodule. (a) θ and ψ are the angle per division in spherical sampling views. (b) The blue dots represent the points in a 40 point spherical t-design. (c) Two opposite view angles will generate the same cross section therefore only [0, π] angular range is taken.



(a) d azimuthal divisions (b) d polar divisions (c) Total d^2 Spherical sampling views

Fig. 4. Figure illustrating how d^2 points are obtained after dividing azimuthal and polar angular range into d equal divisions. Each intersection point (blue dot) represents a view direction used for getting the cross-section.

A. View Sampling Approach

We experimented with two rotation invariant sampling approaches, namely spherical sampling views and t-design views. They are are described as follows:

Spherical Sampling Views: A view point on a sphere is described as a pair of angles (θ, ψ) where θ is the azimuthal angle, and ψ is the polar angle as shown in Figure 3(a). Each view angle acts as the direction of the normal to the plane of cross-section. Since we take a cross-section view of a nodule which passes through nodule's center, two viewing directions exactly opposite to each other on a sphere will generate the same cross-section as shown in Figure 3(c). Hence, we limit the range of θ and ψ to $[0, \pi]$. The view angles are generated by dividing each of the two angular domains into d intervals. Then the number of cross-sections generated for a nodule with the spherical sampling approach becomes d^2 as shown in Figure 4.

t-Design Views: One disadvantage of taking Spherical sampling views is that they are non-uniformly distributed over the sphere. The number of views along the equator is sparse (under-sampled) compared to the number of views taken from the polar views (over-sampled) on a sphere. The asymmetrical sampling results in non-optimal view sampling and information loss around the equator. An alternative approach is to adopt the view sampling of spherical t-design. A spherical t-design is a set of points X on sphere S^2 iff for any polynomial f with a degree less than t, it is possible to determine its average precisely over S^2 by sampling f only at the points in X. In other words a t-design can accurately integrate a polynomial of order t and below. This can also be represented by the following equation:

$$\frac{1}{Vol(S^2)} \int_{S^2} f(\epsilon) d\epsilon = \frac{1}{|X|} \sum_{x \in X} f(x), \tag{1}$$

where in Equation 1, the left hand side represents the average of function f over the entire sphere S^2 and the right hand side represents the average of the values of function f sampled only over points in set X.

Spherical t-designs produce an isotropic distribution of points on S^2 and results in a better sampling of view points. We only consider the points (x, y, z) where $x > 0$. Figure 3(b) shows the placement of t-design points on a sphere S^2. It should be noted in Figure 3(b) that each point is equidistant to all its neighbours unlike spherical sampling points shown in Figure 4(c).

TABLE I
MOBILENET ARCHITECTURE USED IN MULTI-SECTION CNN. CNN_1 COMPRISE LAYERS UP TO AVG-POOL/S1 WHILE CNN_2 COMPRISE LAYERS FROM FC/S1 UP TO THE SOFTMAX LAYER. HERE CONV-DW INDICATES DEPTH-WISE SEPARABLE CONVOLUTION

Type/Stride	Filter Shape	Input Size
Conv/s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw/s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv/s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw/s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv/s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw/s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv/s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw/s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv/s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw/s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv/s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 128$
Conv dw/s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv/s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5 \times$ Conv dw/s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5 \times$ Conv/s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw/s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv/s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw/s1	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv/s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool/s1	Pool 7×7	$7 \times 7 \times 1024$
FC/s1	1024×1000	$1 \times 1 \times 1024$
Softmax/s1	Classifier	$1 \times 1 \times 1000$

B. Network Architecture

It requires a large number of images to train deep network models from scratch while our available cancer dataset only contains a few thousand of nodule samples. To avoid the over-fitting problem and accelerate training process, we applied the transfer learning approach and adopted a lightweight network, called MobileNet [22], that is pre-trained with the ImageNet and employs a depth-wise separable convolution for reducing the number of parameters (weights) see Table I. The last fully connected layer (FC) in the MobileNet network (FC/s1) is replaced with a fully connected layer that is randomly initialized for the task of binary classification. This new final layer is denoted by CNN_2 and the layers before the FC/s1 are denoted as CNN_1 (shown in Figure 2). We used the sigmoid activation function in the final layer and chose the categorical cross-entropy as the loss function of the network. We introduced a View Pooling

Layer between CNN_1 and CNN_2 that acts as an information aggregator and performs an element-wise max-pooling operation on the feature representations across all the sections of a nodule.

C. Training and Inference

The network training process has two stages: during the first stage, the training is nodule-independent and uses the cross-sections of all the nodules obtained from spherical sampling to fine-tune the network end-to-end; the second stage of training is nodule dependent where the weights in the first half of the network are fixed while the second half of the network (CNN_2) is re-trained. The View Pooling Layer performs an element-wise max-pooling operation across all cross-sections and generates a compact representation for each nodule. The compact feature is then used to fine-tune the CNN_2 again. Similarly, the inference process also consists of two steps. All the cross-sections of a nodule at first are passed through the network to obtain the CNN_1 representations. During the second inference step, the View Pooling Layer aggregates the representations of all cross-sections into a compact feature and sends to CNN_2 that subsequently generates the final classification result. We will detail the hyper-parameters for training and the cross-section generation in the next section.

III. Experiment Details

Here we describe the major steps, i.e., preparing training data and selecting hyper-parameters required for training Multi-section CNNs.

A. Data Pre-Processing

1) Lung image database consortium (LIDC) and image database resource initiative (IDRI) Dataset: We utilize the LIDC-IDRI [23] benchmark dataset that has been extensively used in several studies [9], [19], [24]–[26]. The malignancy in the dataset are marked with five levels, namely, Highly Unlikely (1), Moderately Unlikely (2), Indeterminate (3), Moderately Suspicious (4) and Highly Suspicious (5). Similar to the earlier works, we only use the nodules having diameter $\geq$3 mm to ensure the consistency in the evaluation and model comparison. For the same nodule, multiple annotations from up to four radiologists are present in LIDC-IDRI dataset. However, the dataset does not provide any attribute for determining whether two different annotations belong to the same nodule. Therefore, to associate each annotation to some nodule, we form the adjacency matrix for the annotations. The annotations are then clustered following an iterative approach where in each iteration annotations are grouped when the minimum distance between the contour boundary points is smaller than a certain threshold τ, where τ is initialized with the pixel spacing of the CT scan. In each iteration the threshold τ is reduced by multiplying it with a factor $\alpha = 0.9$ and the process is continued till each group of annotation (cluster) has $\leq$4 annotations. The mean of the nodule center in a cluster is defined as the final location for the nodule and mode of the malignancy values in a cluster is taken as the final malignancy level for that nodule. Finally, we labeled 448 nodules with the malignancy level of 4 or 5 "Malignant" and 857 nodules with the malignancy level 1 or 2 as "Benign" and converted the multi-class classification to a binary one. Our experiments used the nodules from 649 patients. Figure 5 shows the distribution of nodules' diameter vs malignancy in the dataset used in our experiments.

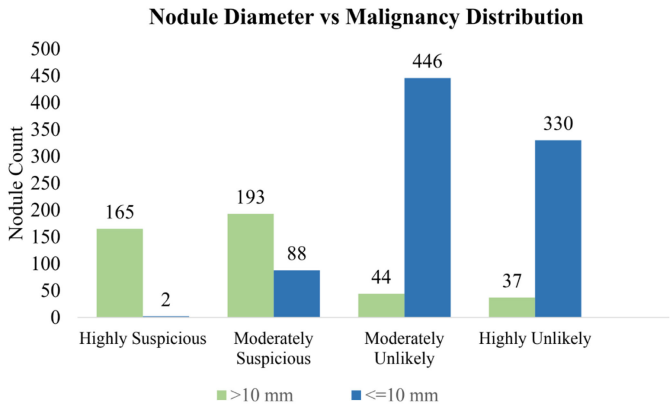


Fig. 5. Nodules' diameter vs malignancy distribution in LIDC-IDRI dataset for nodules having diameter $\geq$3 mm. Number of Benign samples are more as compared to Malignant Samples.

2) Pre-Processing: All the volume samples are obtained by doing the bi-cubic interpolation of the CT scans to obtain 1mm spacing in each dimension (xy, yz and xz). In our experiments all the nodules had diameter less than 50 mm, therefore we limited the volume size to $50 \times 50 \times 50$ mm^3. The extracted volume of size $50 \times 50 \times 50$ mm^3 is centered around the centroid of the nodule. Finally, for each volume, the view direction d_v is obtained using the selected view sampling approach and a plane of size 50×50 centered at nodule is rotated such that d_v is the normal to it. Grid sampling (using bi-cubic interpolation) is done on the planes. This process results into multiple cross-sections from different view angles. Each cross-section is then resized to 224×224. We replicated the obtained sections to form a three-channel image because the MobileNet model takes as input only three-channel images. The model is then trained independently 10 times by a 10-fold cross-validation. In each fold, we divide the dataset based on the nodule id so that all cross-sections of a nodule belong to the same split i.e., either training or testing.

B. Parameter Selection

1) Spherical Sampling Views: The dataset is imbalanced between the benign and malignant cases as shown in Figure 5, and thereby, we sample more views from malignant nodules than from benign nodules during the training step to ensure the total number of views for malignant and benign classes to be the same to each other. To equalize the view count for both classes, we set

$$n_m d_m^2 = n_b d_b^2, \tag{2}$$

where $n_m = 448$, $n_b = 857$ and d_b, d_m are the number of divisions for a benign and malignant sample respectively. After

solving for this equation and rounding the values of d_m and d_b to the nearest integers, we obtain $d_b = 5$ and $d_m = 7$. Note that there are more than one solutions for this equation. We show the view sensitivity experiments later to explore the impact of more number of views. The class label is not known for prediction in prior. Therefore we generate the view angles with $d_t = d_b$ for each nodule sample in the test split.

2) Spherical t-Design Views: Similar to the spherical sampling, we need to ensure that the number of cross-sections/views from Malignant and Benign samples in the training dataset remains same. Therefore we use the following equation to obtain the number of the sampled views:

$$n_m v_m = n_b v_b, \tag{3}$$

where $n_m = 448$, $n_b = 857$ and v_b, v_m are the number of views using the t-design sampling approach. We used the precomputed values of t-design with $v_b = 20$ and $v_m = 40$ points. Here also since the class label is not available while doing prediction we use $v_t = v_b$ for nodules in the test split.

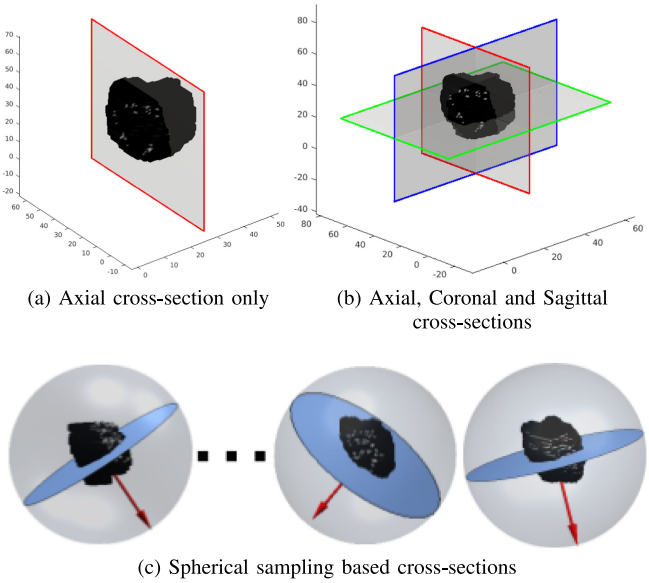
3) Training Hyper-Parameters: The network is already pretrained on ImageNet and needs to be fine-tuned with the cancer dataset at a low learning rate. Therefore we set the learning rate of 0.0001 for a stochastic gradient descent optimizer. In each fold, 10% of the training sections are reserved for validation in the first stage training, while 10% of nodules from the training split are reserved for validation in the second training stage. We reduce the learning rate by a factor of 0.2 once the validation loss reaches a plateau. A patience of five epochs and a minimum delta of 0.001 are used as the criteria for the early stopping. Each epoch has 400 steps, each step comprising 32 mini-batch training samples and the network is trained for a maximum of 50 epochs. We implemented the classification network with Keras library and trained it on a Titan Xp Nvidia GPU.

IV. Results

We performed multiple experiments using the Multi-section CNN to obtain its performance under various settings. Two experiments were done to quantify its performance for classification and malignancy estimation along with two other experiments to understand its sensitivity to the number of views and choice of CNN. To compare the performance of 2D and 3D approaches, we performed two more experiments 1) using only Axial cross-section and 2) using the three orthogonal cross-sections, see Figure 6. Results of the experiments are discussed below.

A. Classification Results

We evaluated the performance of the models using the standard metrics for evaluating binary classification models namely, Accuracy, Sensitivity, Specificity and AUC (Area Under the ROC Curve). Results of the binary classification experiments are shown in Table II. We observed that the 3D information from multiple cross-sections increases the sensitivity and specificity considerably as compared to that of cross-sections using axial only. Table III shows that the proposed Multi-section CNN method outperforms all of the previous fully automated methods



(a) Axial cross-section only (b) Axial, Coronal and Sagittal cross-sections

(c) Spherical sampling based cross-sections

Fig. 6. Sampling approaches considered in our experiments to compare the performance of 2D (Axial) vs 3D (Axial, Coronal, Sagittal) approaches. Our spherical sampling based approach is shown in (c). Results of the experiments are shown in Table II.

[5], [12], [24]–[27] in term of the accuracy in binary classification (malignant vs benign). The ensemble method proposed by Xie *et al.* in [19] achieves slightly better accuracy (93.40%) than ours. However, their method relies on the manual spatial annotation of nodules in the CT slices and is not fully automated. The Multi-section CNN does not require this manual annotation and still achieves comparable performance. Moreover, while the ensemble method in [19] takes 12 hours to train, our MobileNet based model trains within 1 hour.

We also conducted experiments using classical features such as SIFT using a Bag of Words model. We use 16×16 neighborhood around the key-points obtained from SIFT for feature extraction. Then we divided each block into 16 sub-blocks of size 4×4 and created an eight-bin orientation histogram for each sub-block. This generates 128 bin values for each block. We applied K-means clustering with $k = 20$ to cluster these 128-bin features and generate a code-book.

Finally, each nodule is represented by a 20-D feature vector which represents the frequency of each code-word. Three classifiers Logistic Regression, k-Nearest Neighbor and Support Vector Machine (SVM) are trained on these features for the binary classification task. We applied the grid sampling to obtain the parameters for each classifier. The results listed in Table III show that the performance of the SIFT feature based classifiers is considerably lower than that of our Multi-section CNN. ROC curves obtained from some representative models for fold-1 of Test split is shown in Figure 7.

B. Malignancy Estimation Results

Beyond simply classifying into the malignant and benign classes, few methods in the past attempted to quantitatively estimate the malignancy level of a nodule i.e., Highly Suspicious (5), Moderately Suspicious (4), Moderately Unlikely (2)

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT SAMPLING APPROACHES FOR BINARY CLASSIFICATION TASK

View sampling strategy	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Axial section only, MobileNet	81.14	80.22	82.08	0.87
Axial section only, InceptionV3	81.68	82.54	82.01	0.87
Axial, Coronal and Sagittal sections, MobileNet	84.75	83.33	85.46	0.89
Axial, Coronal and Sagittal sections, InceptionV3	86.02	85.13	86.94	0.91
Spherical sampling, MobileNet, $(d_b, d_m) = (2, 3)$	93.01	88.15	**96.04**	**0.98**
Spherical sampling, MobileNet, $(d_b, d_m) = (5, 7)$	**93.18**	89.40	95.61	0.98
Spherical sampling, MobileNet, $(d_b, d_m) = (10, 14)$	92.41	89.01	95.01	0.97
t-Design sampling, MobileNet, $(d_b, d_m) = (42, 80)$	93.10	89.53	95.02	0.98
t-Design sampling, MobileNet, $(d_b, d_m) = (62, 120)$	91.95	86.98	94.85	0.98
t-Design sampling, MobileNet, $(d_b, d_m) = (93, 180)$	91.73	**89.55**	92.96	0.97
t-Design sampling, InceptionV3, $(d_b, d_m) = (42, 80)$	91.57	86.60	94.59	0.97
t-Design sampling, InceptionV3, $(d_b, d_m) = (62, 120)$	91.88	87.76	94.25	0.98
t-Design sampling, InceptionV3, $(d_b, d_m) = (93, 180)$	91.41	87.81	93.65	0.97

TABLE III
COMPARISON OF BINARY CLASSIFICATION METHODS ON LIDC-IDRI DATASET

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Shen *et al.* 2017 [12]	87.14	77.00	93.00	0.93
Dhara *et al.* 2016 [27]	-	89.73	86.36	0.95
Han *et al.* 2015 [5]	-	89.35	86.02	0.94
Anand [24]	86.3	89.60	86.70	-
Hua *et al.* 2015 [26]	-	73.40	82.20	-
Han *et al.* 2013 [9]	-	-	-	0.94
Xie *et al.* 2016 [25]	86.79	60.26	95.42	-
Xie *et al.* 2017 [19] (Single ResNet model)	91.65	88.35	93.34	0.97
Xie *et al.* 2017 [19] (Ensemble ResNet model)*	**93.40**	**91.43**	94.09	0.97
SIFT Bag of Words (Support Vector Machine)	76.09	74.44	76.11	0.79
SIFT Bag of Words (Nearest Neighbor)	70.88	70.54	73.22	0.71
SIFT Bag of Words (Logistic Regression)	74.11	73.66	70.33	0.75
Multi-section CNN (mean $\pm$ standard deviation) spherical sampling, $(d_b, d_m) = (5, 7)$	93.18$\pm$0.03	89.40 $\pm$0.02	**95.61** $\pm$0.07	**0.98** $\pm$0.01

* Ensemble model in [19] requires spatial annotation of nodule which is hard to obtain and requires considerable human intervention.

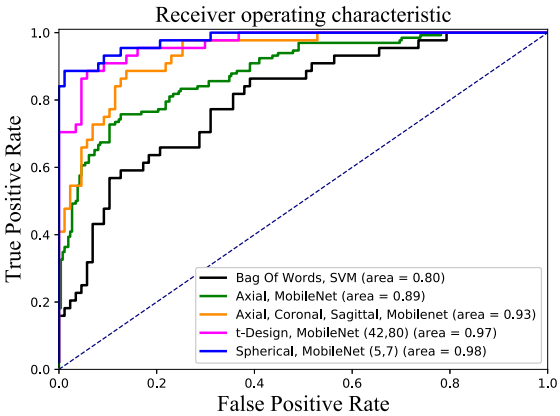


Fig. 7. ROC curves for some representative models for fold-1 of Test split.

and Highly Unlikely (1). We tested the effectiveness of the Multi-section representation in estimating the severity of malignancy by obtaining the class probability using a logistic regression model. The compact feature is used along with the nodule's malignancy level for training the logistic regression model. We performed a 10-fold cross-validation on the same folds used for the classification task. While doing inference the malignancy level with maximum probability is assigned to a nodule. To account for the variability in the malignancy annotations from radiologists, methods in the past adopted the Off-By-One accuracy metric that discounts the error within the range of ± 1 from the actual value [11], [13], [28]. The formulation of Off-By-One accuracy and mean score difference is shown in Equation 4 and 5 as follows.

$$Off - By - One Accuracy = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & \text{if } absolute(p_i - g_i) \leq 1,\\ 0, & \text{otherwise,}\end{cases} \quad (4)$$

$$Mean\ score\ difference = \frac{1}{n}\sum_{i=1}^{n} absolute(p_i - g_i), \quad (5)$$

where in Equation 4 and 5, n is the total number of test samples, p_i is the predicted malignancy level for nodule i obtained from the logistic regression model and g_i is the ground truth malignancy value.

Table IV shows the performance of the Multi-section representation in terms of the mean score difference and Off-By-One accuracy. Our proposed method outperforms those earlier methods and attained lower mean score difference and higher Off-By-One accuracy than the Multi-task learning method using the 3D CNN in [28].

TABLE IV
COMPARISON OF MALIGNANCY ESTIMATION METHODS

Algorithms	Off-By-One Accuracy (%)	Mean Score Difference
Sarfaraz *et al.* (transfer learning) [28]	80.08	0.6259
Sarfaraz *et al.* (multi-task learning) [28]	91.26	0.4593
Sarfaraz [13]	82.47	0.6200
Mario *et al.* [11]	82.40	-
MultiSection CNN spherical sampling, $(d_b, d_m) = (5, 7)$	**93.79**	**0.2713**

C. View Sensitivity Results

We also performed experiments to determine the sensitivity of Multi-section CNN to the number of views. Ideally the large the number of views, the better should be the accuracy. For the spherical view sampling and t-design view sampling, we vary the values of (d_m, d_b) and (v_m, v_b) satisfying equation 2 and 3 respectively to obtain more cross-sections. The sensitivity of the model's performance to the number of cross-sections is shown in Table II. From the results, we observed that having too many cross-sections incur a negative impact on the model's performance. We also observe that the best results are obtained using the spherical sampling approach with $d_m = 7$ and $d_b = 5$.

D. Model Sensitivity Results

In our model, we used MobileNet as the base CNN that is optimized for Mobile devices and has a small number of parameters. Ideally, a model with more parameters or more complex architecture should perform better than a smaller network. To understand the impact of the choice in CNNs in our prediction model, we conducted experiments using another CNN model, i.e., InceptionV3 [29]. For the InceptionV3 model, CNN_1 comprises multiple layers before fully connected and CNN_2 comprise several layers from the fully connected layer to the softmax layer. We conducted experiments with only the t-design views for the model sensitivity study.

The performance variation with the choice of CNN is shown in Table II. We observed that MobileNet, in spite of having a lesser number of parameters, performs better than InceptionV3. One possible explanation for this is that InceptionV3 has about five times more parameters than the MobileNet does and therefore it might experience an over-fitting problem on the dataset with a fixed size.

V. APPLICATIONS AND DISCUSSION

In addition to the diagnosis result, a practitioner always needs to locate the malignancy and visually inspect its formation. One advantage of our method is that it naturally leads to the salient view selection. We use the gradient-based approach to evaluate the saliency of each view. Given a trained network, its output score (denoted by F) and an input cross-section S, the impact of the cross-section to the final output score is determined by taking the gradient of predicted class c score (F_c) with respect to the input cross-section S, i.e., $\partial F_c / \partial S$. Algorithm 1 describes how to select the salient section. Figure 8 shows the salient sections obtained by Algorithm 1. From Figure 8 we note that the model

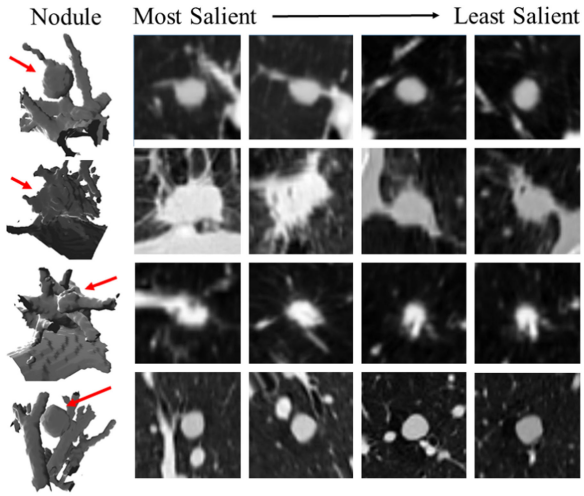


Fig. 8. Salient cross-sections of Malignant nodules. Spiculations are captured clearly in the most salient cross-sections.

Algorithm 1: Salient Sections.

$N \leftarrow$ Number of sections
$S \leftarrow$ Section images
$K \leftarrow$ Output Salient Sections Count
procedure GETSALIENTSECTION(N, K)
  for i `in` N **do**
    $Gradient_i \leftarrow \partial F_c / \partial S_i$
    $AbsGradient_i \leftarrow absolute(Gradient_i)$
  end for ▷ Sort views on absolute gradient score
  $sortedSections \leftarrow sort(AbsGradient, S)$
  return K sections from $sortedSections$
end procedure

can highlight the discriminative features, such as spiculations on the nodule. Those less salient cross-sections are devoid of these spiculations and appear to be quite benign in some cases. This observation validates the requirement for a Multi-section approach.

To demonstrate the discriminative ability and effectiveness of the aggregated information from multiple sections in the Multi-section model, we visualized the feature embeddings using the t-SNE algorithm [30]. Figure 9 shows the comparison between the t-SNE embeddings of the axial cross-section's representation and the aggregated feature representation. We observed that the nodules are cleanly separated between two classes with the Multi-section compact representation, while the two classes are significantly overlapping with each other under the representation of the axial cross-section.

We also ported the Multi-section CNN model to an Android tablet (Google Pixel C) which comes with an 8-core ARM processor and 3 GB of RAM. Figure 10 shows a prototype of the android application where a practitioner can select the suspicious nodule location on a CT scan. The model then outputs the malignancy probability. We also performed experiments to measure the inference time for a nodule using Multi-section CNN. For the inference on ten cross-sections with the

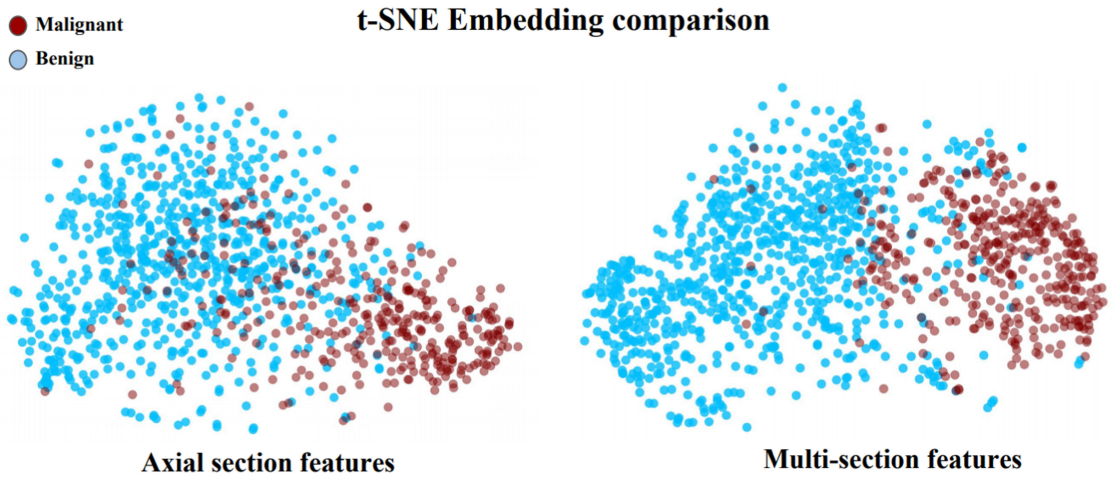


Fig. 9. t-SNE embedding (perplexity = 70) comparison for fold-1 of Train split. We observe that the overlapping is significantly less in case of Multi-section features compared to only axial section features.

TABLE V
COMPARISON OF MEMORY FOOTPRINT AND INFERENCE TIME. HERE, MULTI-SECTION CNN USES MOBILENET AS BASE CNN

Model	Accuracy	Max RAM Requirement (in MB)	Inference Time (in sec)	Size (in MB)	Parameters (in millions)
Xie *et al.* [19], Ensemble Model	93.40	517	1.37±0.25	313 MB	76.9
Ours, Spherical Sampling, $(d_b, d_m) = (5, 7)$	93.18	131	0.89±0.06	43	4.2
Ours, Spherical Sampling, $(d_b, d_m) = (2, 3)$	93.01	101	0.45±0.05	43	4.2

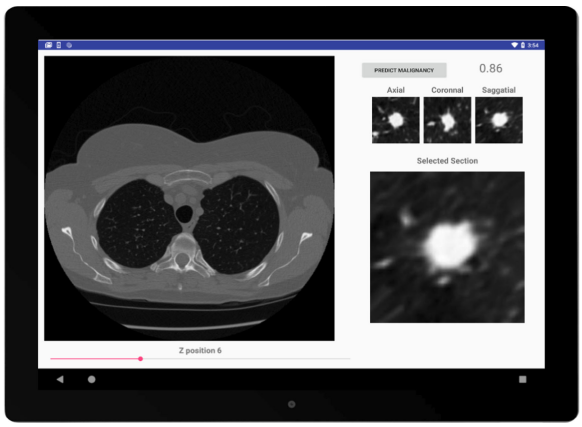


Fig. 10. Screen capture of an Android application running on a Google Pixel C tablet. User selects the slice location by varying the z position and picks a nodule location. The Multi-section CNN model does the on-device prediction of malignancy.

MobiletNet Multi-Section CNN it took an average 974 ± 28 ms (for 10 cross-sections and the average taken over 5 runs). To demonstrate the efficient design of our model, we compared its memory requirements and inference latency with the Ensemble model proposed by Xie *et al.* in [19]. The testing device in this experiment is Google Pixel 2 that comes with an 8-core ARM processor and 4 GB of RAM. We calculated the mean inference time for the classification of a nodule over five runs. Also, the maximum memory (RAM) requirement during the inference was recorded by the Android Profiler Tool. The results of the experiments in Table V confirmed that our model has considerably less inference time and memory requirement than the model for comparison and will run efficiently even on mobile devices.

Currently Android does not support the complete TensorFlow [31] functionalities, therefore we can not implement the salient section algorithm on a mobile device. In future, we expect that a lightweight Tensorflow with the entire functionalities will be available for mobile devices. Moreover, determining salient sections requires to calculate the gradient of the output score with respect to the input image and incurs intensive computation as compared to the inference only machine learning model. Therefore, an alternative is to adopt a hybrid approach using both on-device and cloud capabilities. Because the practitioner might need to observe the malignancy in multiple nodules, an on-device malignancy prediction will provide a smooth experience. However, for the salient section determination, a cloud based paradigm is needed where a model exploits GPU's capabilities to determine and fetch the salient sections quickly. Another future extension to our approach is to design an automated nodule detection algorithm to assist the practitioner in identifying and locating a suspicious nodule from a large image of CT scan.

VI. CONCLUSION

In this paper, we introduced a novel Multi-section CNN for classifying lung nodules and estimating the probability of malignancy. The experiment results showed that our proposed model outperforms several state-of-the-art classification methods. Our Multi-section CNN method does not require the tedious manual spatial annotation, is lightweight, and can be easily ported

to Mobile devices, such as tablets and embedded system. This portability will lead to a wide adoption by practitioners. In addition, we envision a hybrid solution that utilizes the cloud-based computing capacity to highlight the causal factors of prediction results and delivers them back to the mobile device, making our method a compelling choice for the clinical settings.

Acknowledgment

The authors would like to acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] B. W. K. P. Stewart and C. P. Wild, *World Cancer Report*. Lyon, France: Int. Agency Res. Cancer, 2017.
[2] Nat. Lung Screening Trial Res. Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England J. Med.*, vol. 365, no. 5, pp. 395–409, 2011.
[3] S. Swensen *et al.*, "Lung nodule enhancement at CT: Multicenter study," *Radiology*, vol. 214, no. 1, pp. 73–80, 2000.
[4] A. El-Baz *et al.*, "3D shape analysis for early diagnosis of malignant lung nodules," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, 2011, pp. 772–783.
[5] F. Han *et al.*, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *J. Digit. Imag.*, vol. 28, no. 1, pp. 99–115, 2015.
[6] C. Jacobs *et al.*, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Med. Image Anal.*, vol. 18, no. 2, pp. 374–384, 2014.
[7] E. L. Torres *et al.*, "Large scale validation of the m5l lung CAD on heterogeneous CT datasets," *Med. Phys.*, vol. 42, no. 4, pp. 1477–1489, 2015.
[8] K. Murphy, B. van Ginneken, A. M. R. Schilham, B. J. De Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," *Med. Image Anal.*, vol. 13, no. 5, pp. 757–770, 2009.
[9] F. Han *et al.*, "A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database," in *Proc. IEEE Int. Conf. Med. Imag. Phys. Eng.*, 2013, pp. 14–18.
[10] T. W. Way *et al.*, "Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours," *Med. Phys.*, vol. 33, no. 7, pp. 2323–2337, 2006.
[11] M. Buty, Z. Xu, M. Gao, U. Bagci, A. Wu, and D. J. Mollura, "Characterization of lung nodule malignancy using hybrid shape and appearance features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2016, pp. 662–670.
[12] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2015, pp. 588–599.
[13] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci, "Tumornet: Lung nodule characterization using multi-view convolutional neural network with Gaussian process," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, 2017, pp. 1007–1010.
[14] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[15] F. Ciompi *et al.*, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Sci. Rep.*, vol. 7, 2017, Art. no. 46479.
[16] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.
[17] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: 3d deep convolutional nets for automated pulmonary nodule detection and classification," pp. 673–681, Mar. 2017, doi: 10.1109/WACV.2018.00079.
[18] A. A. A. Setio *et al.*, "Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.
[19] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention.*, 2017, pp. 656–664.
[20] M. Alilou *et al.*, "Intra-perinodular textural transition (IPRIS): A 3D descriptor for nodule diagnosis on lung CT," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 647–655.
[21] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 945–953.
[22] A. G Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, arXiv:1704.04861.
[23] S. G. Armato *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.
[24] S. K. V. Anand, "Segmentation coupled textural feature classification for lung tumor prediction," in *Proc. IEEE Int. Conf. Commun. Control Comput. Technol.*, 2010, pp. 518–524.
[25] Y. Xie *et al.*, "Lung nodule classification by jointly using visual descriptors and deep features," in *Proc. Bayesian grAph. Models Biomed. Imag. Int. MICCAI Workshop Med. Comput. Vis.*, 2016, pp. 116–125.
[26] 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 Therapy*, vol. 8, pp. 2015–2022, 2015.
[27] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal, "A combination of shape and texture features for classification of pulmonary nodules in lung ct images," *J. Digit. Imag.*, vol. 29, no. 4, pp. 466–475, 2016.
[28] S. Hussein, K. Cao, Q. Song, and U. Bagci, "Risk stratification of lung nodules using 3D CNN-based multi-task learning," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 249–260.
[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proc. IEEE Conf. Comput. Vision Pattern Recogt.*, pp. 2818–2826, 2016.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[31] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement.*, 2016, pp. 265–283.