# Journal Pre-proofs

NROI based feature learning for Automated Tumor Stage Classification of pulmonary lung nodules using Deep Convolutional Neural Networks

Supriya Suresh, Subaji Mohan

Please cite this article as: Suresh, S., Mohan, S., NROI based feature learning for Automated Tumor Stage Classification of pulmonary lung nodules using Deep Convolutional Neural Networks, *Journal of King Saud University - Computer and Information Sciences* (2019), doi: https://doi.org/10.1016/j.jksuci.2019.11.013

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# NROI based feature learning for Automated Tumor Stage Classification of pulmonary lung nodules using Deep Convolutional Neural Networks

**Abstract**

Identifying the exact pulmonary nodule boundaries in computed tomography (CT) images are crucial tasks to computer-aided detection systems (CADx). Segregation of CT images as benign, malignant and non-cancerous is essential for early detection of lung cancers to improve survival rates. In this paper, a methodology for automated tumor stage classification of pulmonary lung nodules is proposed using an end-to-end learning Deep Convolutional Neural Network (DCNN). The images used in the study were acquired from the Lung Image Database Consortium and Infectious Disease Research Institute (LIDC-IDRI) public repository comprising of 1018 cases. Lung CT images with candidate nodules are segmented into a 52x52 pixel nodule region of interest (NROI) rectangle based on four radiologists' annotations and markings with ground truth (GT) values. The approach aims in analyzing and extracting the self-learned salient features from the NROI consisting of differently structured nodules. DCNN are trained with NROI samples and are further classified according to the tumor patterns as non-cancerous, benign or malignant samples. Data augmentation and dropouts are used to avoid overfitting. The algorithm was compared with the state of art methods and traditional hand-crafted features like the statistical, texture and morphological behavior of lung CT images. A consistent improvement in the performance of the DCNN was observed using nodule grouped dataset and the classification accuracy of 97.8%, the specificity of 97.2%, the sensitivity of 97.1%, and area under the receiver operating characteristic curve (AUC) score of 0.9956 was achieved with reduced low false positives.

*Keywords:* Deep learning, Convolutional neural network, Segmentation, Data augmentation, Dropouts.

## 1. Introduction

Lung cancer is one of the major widespread diseases worldwide leading to high death rates among other types of cancer. According to statistics 2019 in the United States, nearly 228,150 new lung and bronchus cancer individuals and 142,670 deaths were estimated by the American Cancer Society (Siegel et al., 2019). In order to increase survival rates, the early detection of lung cancer plays a crucial role. Particularly, identifying the pulmonary nodules in the early stages requires radiologist's attention to a greater extent as the nodule densities may have similar anatomical properties to that of the other lung structures (Xiuhua et al., 2011).

However, computer-aided diagnosis/computer-aided detection systems were considered as an alternative approach for automated pulmonary nodule detection to help radiologists overcome the issues with conventional reading (Jacobs et al., 2014). Identifying the exact candidate nodule boundaries in CT images are crucial tasks to CADx due to the similar visualization characteristics of candidate nodules and its surroundings. In addition, extracting the salient features of candidate regions plays a vital role. Usually, features are extracted using traditional hand-crafted features or deep learning methods (deep neural networks) for pulmonary lung nodule classification. Traditional hand-crafted features include statistical, texture, density and morphological behavior and are further classified using few classification techniques. The existing CADx systems are in need to design these features as an essential model. But the process is time consuming and

complicated (Roth et al., 2016). Moreover, the features are to be correlated to obtain expected performance measures. In addition, reducing false positives play a vital role in nodule classification increasing sensitivity rates.

Recently, many researchers have developed architectures to automatically learn and extract feature maps (da Silva et al., 2017; Tajbakhsh & Suzuki, 2017; Yuan et al., 2018) using deep learning techniques. These techniques are used to detect and classify candidate lung nodules without considering manually extracted texture features (Wang et al., 2017b,a; Wikipedia contributors, 2018). Learning features from architectures with multiple layers and hierarchical models of input data like deep neural network/hybrid structures are the trend changing concepts in the recent past. The study results show that deep neural network algorithms can outperform compared to traditional machine learning concepts (Lu et al., 2018; Rastegari et al., 2016). Recent advances in deep neural networks involve the concept of parallel computing with more accessibility and affordability in using graphics processing units for training huge annotated datasets. Many researchers made progress in training and classifying huge datasets using deep learning algorithms for pattern recognition (Zhang et al., 2019; Wang et al., 2018). This led to substantial advancements in using deep neural network algorithms for medical imaging applications as well (Tan et al., 2017).

In the study, we incorporate an end-to-end DCNN architecture for candidate nodule feature extraction and classification according to the malignancy suspiciousness. The NROI is extracted, trained and tested using DCNN and each detected candidate nodule is classified according to stages of malignancy suspiciousness as normal (non-cancerous), benign (level 1 or 2 ), and malignant (level 4 or 5).

The paper is organized as follows. Section 2 describes the related works. Section 3 describes the methodology used to segment and classify candidate nodules as benign, malignant or non-cancerous, using the extracted features from traditional hand-crafted methods and deep convolutional neural networks. Section 4 illustrates the experiments and evaluation of the proposed method. Section 5 discusses the results.

## 2. Related Work

The lung nodule detection systems with nodule segmentation, feature extraction, and nodule classification have certain challenging tasks to overcome as discussed previously. In this section, the previous works related to the proposed method are discussed. Most studies related to lung cancer diagnosis not only relay on extracted features from traditional hand-crafted methods but also use automated end-to-end learned features. In order to organize the related work more precisely, we separated the related works into two groups: Group 1 - Works related to traditional hand-crafted based feature extraction. Group 2 - Works related to deep convolutional neural network-based feature extraction.

### 2.1. Group 1 - Works related to traditional hand-crafted based feature extraction

Wu et al. (2013) designed a methodology to differentiate malignant from benign samples using a combination of texture and radiological features. A total of 13 gray level co-occurrence matrix (GLCM) texture

features and 12 radiological features were extracted from 2117 CT slices (116 malignant and 86 benign). The extracted features were classified using back propagation neural network (BPNN) neural networks resulting in a receiver operating characteristic curve (ROC) of 0.91.

de Carvalho Filho et al. (2014) developed a methodology to automatically detect the candidate pulmonary nodules by enhancing the images using quality thresholding. Candidate nodules were segmented using region growing methods. The shape and texture features were extracted using genetic algorithms and further classified using SVM classifiers. A total of 140 new exams were collected from Lung Image Database Consortium (LIDC) dataset achieving an accuracy of 97.55%, a sensitivity of 85.91% and a specificity of 97.70%.

Table 1: Works related to traditional hand-crafted based feature extraction

| Research Papers | Objective | Image Database | No. of Images | Implementation Key difference | Results |
|---|---|---|---|---|---|
| Wu et al. (2013) | Nodule Classification | LIDC | 2117 CT slices 116 - malignant 86 - benign | Combination of 13 texture and 12 radiological features BPNN | ROC curve of 0.91 |
| de Carvalho Filho et al. (2014) | Nodule detection | LIDC | 140 new exams | Quality Thresholding Region Growing segmentation Shape and texture based feature extraction using genetic algorithm SVM classifiers | Accuracy - 97.55% Sensitivity - 85.91% Specificity - 97.70% Time Complexity - 13.56 min |
| Orozco et al. (2015) | Nodule Classification | LIDC and ELCAP | 61 images 36 - nodules 25 - non-nodules | Feature extraction - wavelet based 19 GLCM features SVM classification | Accuracy - 82% Sensitivity - 90.90% Specificity - 73.91% |
| de Carvalho Filho et al. (2017) | Nodule Classification | LIDC | 1403 images 394 - malignant 1011 - benign | Feature extraction: Phylogenetic diversity index and genetic algorithms | Accuracy - 92.52% Sensitivity - 93.1% Specificity - 92.26% |

Orozco et al. (2015) proposed a lung nodule classification scheme using wavelet feature-based descriptor classified using SVM classifiers. A total of 19 GLCM features were extracted from the frequency domain sub-bands. The datasets used in the study were extracted from the Early Lung Cancer Action Program (ELCAP) and LIDC containing 61 images (36 nodules and 25 non-nodule images). The system achieved an accuracy of 82%, sensitivity of 90.90%, specificity of 73.91%.

de Carvalho Filho et al. (2017) further extracted the texture features using phylogenetic diversity index and genetic algorithms for lung nodule classification. A total of 1403 nodule images (394 malignant and 1011 benign) from the LIDC dataset were used in the study resulting in an accuracy of 92.52%, sensitivity of 93.1% and specificity of 92.26%.

*2.2. Group 2 - Works related to deep convolutional neural network-based feature extraction*

de Carvalho Filho et al. (2018) proposed a classification approach differentiating the patterns of benign and malignant samples using topology-based phylogenetic diversity index on CT images. For classifying the extracted features, CNN methods were used. LIDC image dataset comprising 1405 nodules (394 malignant

and 1011 benign nodules) were used in the study achieving an accuracy of 92.63%, the specificity of 93.47%, the sensitivity of 90.7% and Area under ROC of 0.934.

Xie et al. (2018), texture, shape and deep model-based learned information (FuseTDS) were used for classifying the lung nodules. The methodology employed GLCM based texture features, Fourier shape features for characterizing the heterogeneity of lung nodules and DCNN to automatically learn the nodules features slice-by-slice. Each feature is trained using Adaboost BPNN. The three classifiers were fused to differentiate the candidate lung nodules. The algorithm was evaluated against the LIDC-IDRI image dataset consisting of 1972 nodules (1323 benign and 648 malignant), The algorithm achieved an AUC of 96.65%, accuracy of 89.53%, the sensitivity of 84.19% and specificity of 92.02%.

Table 2: Works related to deep convolutional neural network based feature extraction

| Research Papers | Objective | Image Database | No. of Images | Implementation Key difference | Results |
|---|---|---|---|---|---|
| de Carvalho Filho et al. (2018) | Nodule Classification | LIDC | 1404 nodules 394 - malignant 1011 - benign | Topology based Phylogenetic diversity index classification - CNN | Accuracy - 92.63% Sensitivity - 90.70% Specificity - 93.47 ROC curve of 0.934% |
| Xie et al. (2018) | Nodule Classification | LIDC | 1972 nodules images 648 - malignant 1324 - benign | FuseTDS fuses texture, shape and deep model learning crops a 64x64 square region with candidate nodules embedded GLCM based feature descriptor Fourier shape descriptor Feature extraction - DCNN Adaboost BPNN for training | Accuracy - 89.53% Sensitivity - 84.19% Specificity - 92.02% ROC curve of 96.65% |
| Jiang et al. (2018) | Nodule detection | LIDC | 1006 scans | Images enhanced using Frangi filter Nodule detection using 4 channel CNN (Multi - group based learning) | Sensitivity - 80.6% with FP of 4.7 Sensitivity - 94% with FP of 15.1 |
| Xie et al. (2019) | Nodule detection | LUNA 16 | 150414 images 339 - nodules 150075 - non-nodules | Nodule detection - Deconvolutional false positive reduction using 2D CNN | Sensitivity - 86.42% |
| Huang et al. (2019) | Nodule detection | LUNA16 | 888 CT scans 223 nodules | Nodule detection - Faster region CNN CNN for false positive reduction and candidate merging Segmentation - customised fully CNN | Accuracy - 91.4% with FP of 1 Accuracy - 94.6% with FP of 4 |
| Lakshmanaprabu et al. (2019) | Nodule classification | LIDC | 70 images Normal - 27 Benign - 21 Malignant - 22 | Optimal Deep Neural Network (ODNN) Feature extraction - Linear Discriminate Analysis Classification - Modified gravitational search algorithm | Accuracy - 94.56% Sensitivity - 96.2% Specificity - 94.2% |

Jiang et al. (2018) proposed a multi-group based learning system for lung nodule detection. 1006 scans from LIDC datasets were enhanced by the Frangi filter. The nodules were detected through a 4 channel CNN resulting in a sensitivity of 80.06% with false positives of 4.7 per scan and sensitivity of 94% with false positives of 15.1 per scan.

Xie et al. (2019) developed an automated candidate nodule detection scheme using a faster region-based convolutional neural network. Deconvolutional layers were used to detect the candidate nodules. In order to reduce false positives, 2D CNNs were employed in the study. Experiments were conducted on total candidates of 150414 ( 339 nodule images and 150075 non-nodule images) from LUNA 16 datasets for training. Candidates were labeled with class 0 for non-nodules and 1 for nodule images. The system exhibited a nodule detection sensitivity of 86.42%.

Huang et al. (2019) designed an end-to-end fully automated framework to precisely segment candidate nodule contours from CT images. The methodology involves 3 major phases. 1) Faster regional CNN to detect the candidate nodules 2) CNN for false-positive reduction and candidate merging 3) nodule segmentation using customized fully CNN. The experiments were conducted on 888 CT scans having 223 nodules from LUNA 16 image datasets and resulted in an accuracy of 91.4% with false positives 1 per scan and 94.6% with false positives of 4 per scan.

Lakshmanaprabu et al. (2019) proposed an optimal automated classification method using Optimal Deep Neural Network (ODNN) and Linear Discriminate Analysis (LDA). The deep extracted features from lung nodules were reduced using LDA. The candidate nodules were further classified using Modified gravitational search algorithm (MGSA). 70 images were used for training (Normal - 27, Benign - 21, Malignant - 22) and 30 images were used for testing (Normal - 8, Benign - 11, malignant - 11) resulting in an accuracy of 94.56%, the specificity of 94.2% and sensitivity of 96.2%.

From the concept of unique feature learning from hierarchical neural network layers, CNN models are beneficial for image segmentation and classification exhibiting encouraging results in medical imaging. All the above methods exhibit promising results related to accuracy and sensitivity but few approaches result in high false positives either per scan or per patient or per-image basis, in turn, affecting the performance of classification accuracy. However, there do exist some uncertainties to CNN approaches in modeling heterogeneous lung CT volumes acquiring datasets with enough samples for training. Most of the methods discussed do not have an equal number of image samples for each class of malignancy categorized resulting in overfitting. Designing the neural network hierarchy for capturing 2D features of candidate nodule regions are to be explicitly addressed. In summary, several methods exhibit potential progress in the field of lung nodule detection, segmentation, and classification but still needs improvement to overcome the challenging issues like detection of irregularly structured nodules based on the shape, varying size, and location identification, high sensitivity rates by reducing low false positives, providing robust technologies applicable across multiple databases.

In summary, Table 1 and 2 represents the works related to traditional hand-crafted based feature extraction and deep convolutional neural network methods respectively.
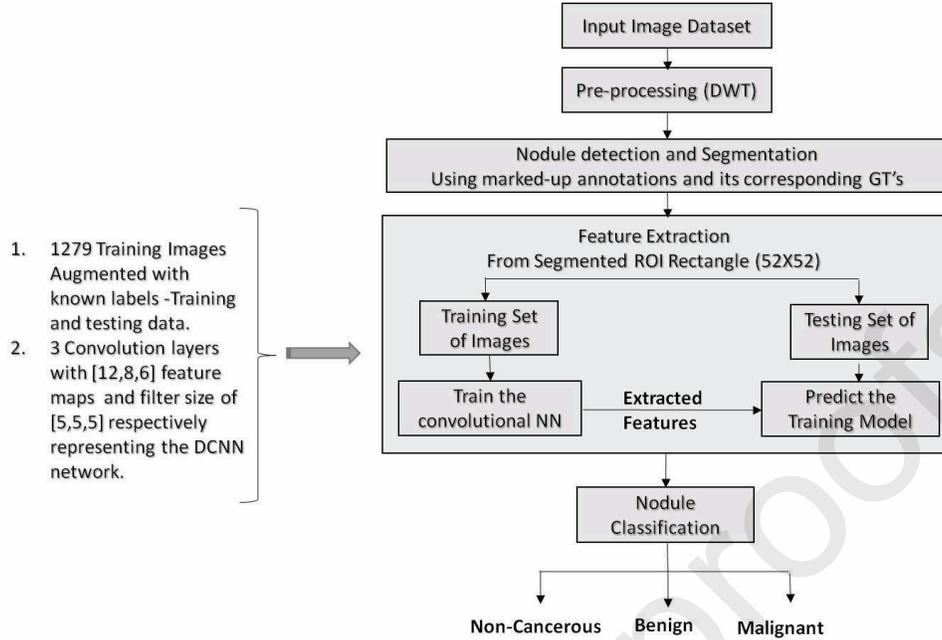
5

1. 1279 Training Images Augmented with known labels -Training and testing data.
2. 3 Convolution layers with [12,8,6] feature maps and filter size of [5,5,5] respectively representing the DCNN network.

Figure 1: Architecture of the proposed method

## 3. Methodology

In this paper, a novel approach for automated tumor stage classification of pulmonary nodules differentiating malignant from benign and non-cancerous lung nodules is proposed. The algorithm is summarised in Figure 1. For comparison, four variants of features are extracted from each lung nodule CT image. Three are from traditional hand-crafted methods - statistical, texture and morphological based features classified using SVM classifiers. One learned from 8 layered deep convolutional neural networks. The self-learned extracted features are trained and tested using CNN's to reduce false positives.

### 3.1. Data Acquisition

All the images used in the study are collected from the publically available repository Lung Image Database Consortium and Infectious Disease Research Institute dataset consisting of 1018 CT scans with marked-up annotations by 4 expert radiologists (Armato III, 2015; Jacobs et al., 2016). Images from the LIDC-IDRI datasets are of 512 x 512 dimensions in size. Each CT slice thickness varies from 1.25 mm to 3mm with nodule diameters varying from 3mm to 30mm indicating the malignancy suspiciousness from levels 1 to 5. CT scan images are pre-processed to uniquely segment the NROI in correspondence to four Radiologists' annotations and markings. An XML file is associated with each case of CT scans and a two-phase annotation process was held by four radiologists to distinguish the suspicious nodules. Each radiologist reviewed and labeled the nodules/lesions to one of the three key categories: nodule greater than or equal to 3mm, non-nodule larger than 3mm, nodule less than 3mm. Each individual annotations are read from the XML files and their corresponding locations in DICOM images are traced and cropped. The dataset images

6

are segmented based on these traced annotations in correspondence to the malignancy levels, extracting the nodule area in each slice into a 52x52 pixel rectangle converted into a TIF image format for easier processing Lampert et al. (2016). Nodules with ambiguous 'IDs' and samples with malignancy levels 3 are eliminated in the study to distinguish the pulmonary nodules better. Overall, a total of 1279 samples are used in the experiments with 278 benign (LMNs), 432 malignant (HMNs) and 569 non-cancerous samples.
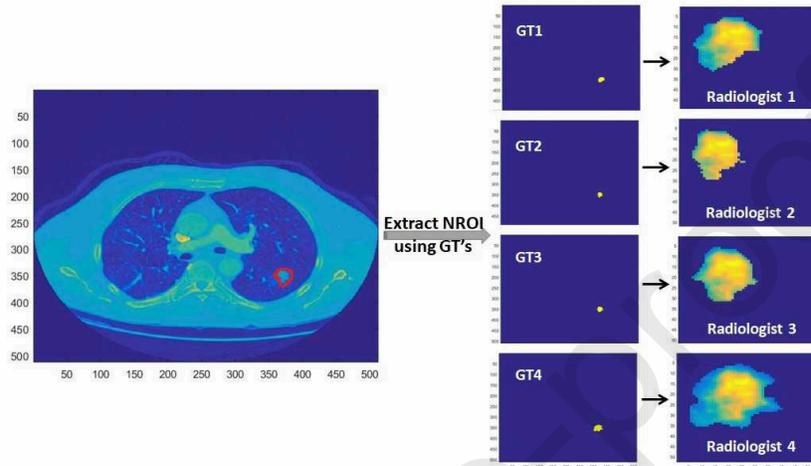


Figure 2: CT image with small bottom left lung nodule highligted in 'red' annotated by four expert radiologists and their corresponding GTs

### 3.2. Nodule extraction and Segmentation

Candidate nodule regions are extracted slice-by-slice using the marked-up annotations and their corresponding GT values from datasets. Figure 2 shows the CT image with lung nodule highlighted in 'red' annotated by four expert radiologists and their corresponding GTs. Each annotation in the XML file corresponding to the candidate nodule is segmented according to their malignancy levels. The candidate nodule's pixel values are retained using the masks meanwhile the rest are padded with zero, forming NROI 'Tiff image' of size 52x52 pixel rectangle. All the information concerning to the nodule structures along with its shapes and sizes are extracted efficiently fitting the pulmonary nodule in the rectangle frame. In case the nodule size exceeds the 52x52 rectangle size, down-sampling is applied to the large nodules to fit into the rectangle. Figure 3 shows few NROI images and their corresponding candidate nodules highlighted in 'red'.

### 3.3. Feature Extraction

#### 3.3.1. Deep convolutional neural network-based feature extraction

In order to automatically learn the features from NROI, an 8 layered Convolutional Neural Network was constructed. Since each candidate nodules have varying sizes and shapes, a template was generalized with each region of interest (ROI) resized into a 52 x 52-pixel rectangle as input to DCNN. Rather than considering the entire dataset images of size 512x512 for training the neural network, only the NROI rectangle was used to improve the processing time, storage capability, and extract relevant features maps to
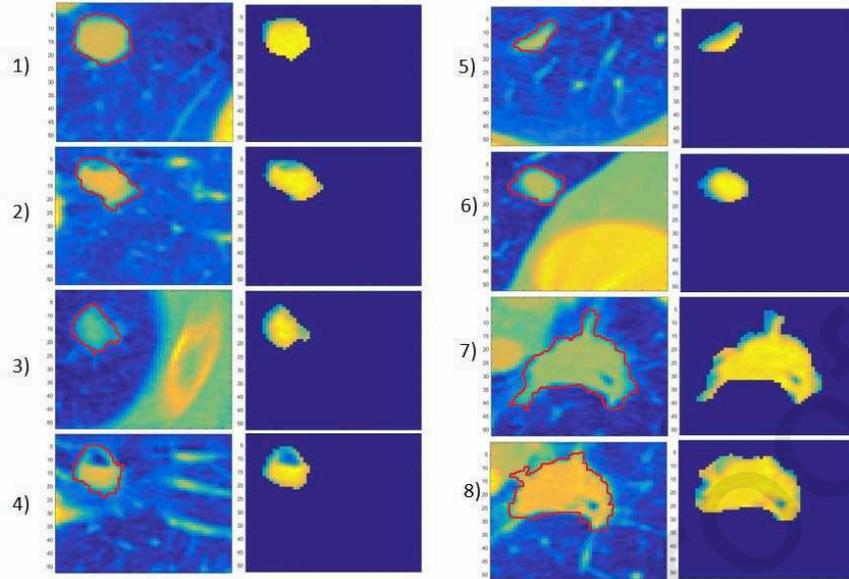
7

Figure 3: Few NROI images and their corresponding candidate nodules

understand the characteristics captured by DCNN better. The NROI with candidate nodules identifies the majority of contributions towards prediction by considering the importance of the tissues surrounding the candidate nodules increasing the classification accuracy with low false positives. The network filters, feature maps, input image sizes are tuned by DCNN architecture for better pulmonary nodule classification accuracy. Thus the augmented NROI of size 52 x 52 image patches is passed as input to the input layer of DCNN. CNN's have three layers 1) Convolutional layers 2) Pooling layers 3) Fully connected layer. The architecture incorporated in the study consists of three convolutional layers and three sub-sampling layers intercepted with max-pooling, Rectified linear unit (ReLu) and batch normalization for salient feature extraction and finally a fully connected layer connected to 3 neurons classifying input patterns to one of the categories of classes as benign, malignant, and non-cancerous. Figure 4 represents the structure of the DCNN.

The first convolutional layer has 12 filters of size 5x5 feature maps connected to the input layer. The second layer has 8 filters of size 5x5 connected to the previous layer (12x8 = 96 5x5 filters). The third layer has 6 filters of size 5x5 (8x6 = 48 5x5 filters) used from the previous layer. Each filter produces a 2D image output of 12x48x48 images from the first convolutional layer as shown in Figure 4. The number of filters used may be varied to optimize the classification accuracy during training. The last layer before the output layer(eight layers), which is the fully connected layer has the input shrunk to 3x3 matrices using softmax non-linear functions having 3 output neurons that fall into one of the 3 categories of classes benign, malignant or non-cancerous nodules. Fully connected implies that every neuron in one layer is connected to the same location at the other layers. As a result, each neuron receives input as linear combinations from its corresponding neurons with a set of input weights and bias in the previous layer. Finally, the output layer provides the strength of the network prediction for each possible category of classes. The output of each
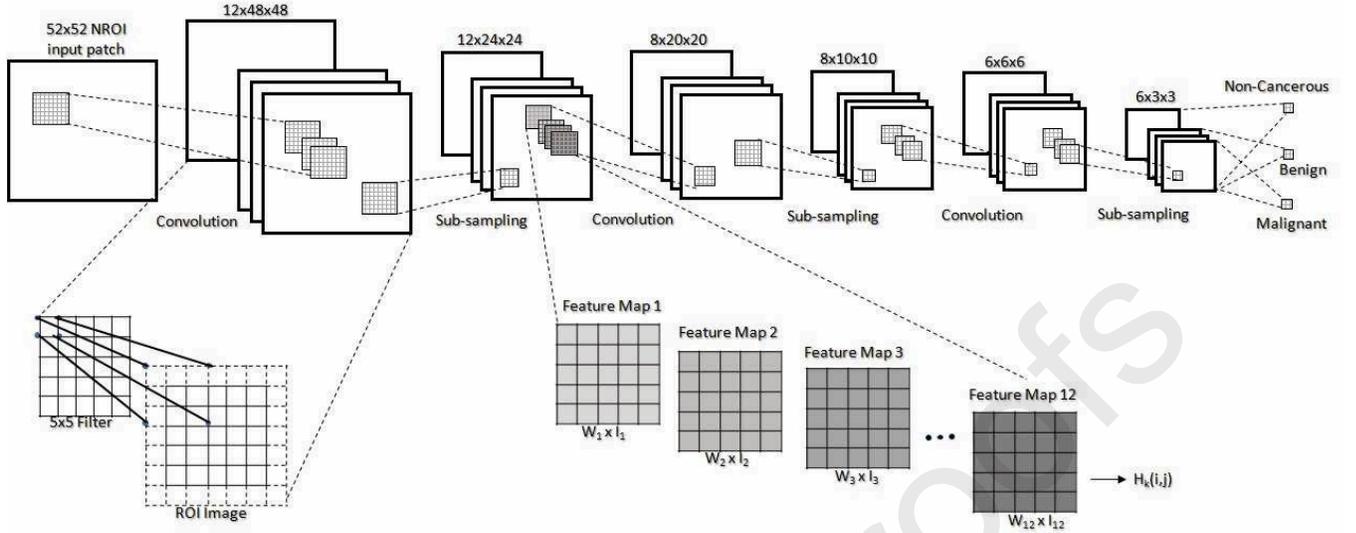
Figure 4: Structure of the DCNN

Table 3: Tested combinations of input parameters for the DCNN Architecture.

| Input Parameters | Values |
|---|---|
| # of Layers | 4, 6, 8, 10, 12 |
| # of kernels | 32, 16, 12, 8, 6 |
| Learning rate(Alpha) | 0.01, 0.1 |
| Kernel size | 3, 5 |

layer in the CNN architectures was normalized, whitened to enhance the contrast before it was sent to the next layer Hyvärinen & Oja (2000).

Each convolutional layer is followed by a max-pooling layer of size 2x2 (stride of 2) that reduces the size of the input patch by half with maximum intensity values in the non-overlapping window thus reducing the computational cost and helps maintain the neural network unchanged to any translations, transformations, and distortions on the input patches. Rectified linear unit in the sub-sampling layer is applied to each pixel in the input patch replacing all negative values in the feature maps to 0 and preserving only non-negative pixel values for further processing. Dropouts are used in architecture to prevent overfitting. Also while training the network, a change in the distribution of data was observed with gradient values (trending to decrease), hence batch normalization was incorporated in the algorithm making each layer independently learn more and align the data distribution more efficiently for the other layers. To optimize the performance of DCNN, adaptive moment estimation (Adam) optimizer (Wikipedia contributors, 2019; Kingma & Ba, 2014) was employed with a learning rate of 0.1. The number of iterations was set to 50 with a batch size of 100 and a sub-sampling rate constantly set to 2 in the study. A systematic analysis was conduct against varying input parameters like the convolution filter sizes, learning rates and the number of layers in the architecture. Table 3 shows the tested combinations of input parameters for the DCNN Architecture.

Table 4: Traditional handcrafted features used in the study for comparison.

| Category | | Features extracted |
|---|---|---|
| Texture | GLCM | 1) Contrast, 2) Correlation, 3) Energy, 4) Entropy, 5) Homogeneity, 6) Autocorrelation, 7) Cluster Prominence, 8) Cluster shade, 9) Inverse difference, 10) Difference in entropy, 11) Sum entropy, 12) Difference in variance, 13) information measure on correlation1, 14) information measure on correlation2, 15) Dissimilarity, 16) Max probability, 17) Sum average, 18) Sum variance |
| Statistical | | 19) Mean, 20) Standard deviation, 21) Variance, 22) Skewness, 23) Kurtosis |
| Morphological | | 24) Area, 25) Perimeter, 26) Eccentricity |

### 3.3.2. Traditional hand-crafted based feature extraction

For comparison purposes, we adapted statistical, texture and morphological based descriptors as described in Table 4 for characterizing the heterogeneity of candidate nodule voxel values from the same NROIs. A total of 26 features were extracted manually from traditional hand-crafted methods defining discriminative features. All features were combined and trained using SVM classifiers for the same categories of classes specified as non-cancerous, benign or malignant. The features extracted like area, perimeter, eccentricity, contrast, correlation, energy, entropy, homogeneity, sum average, and sum entropy plays a prominent role in nodule classification. These features are correlated to obtain expected performance measures.

### 3.3.3. Data Augmentation

Data augmentation is one of the techniques used to overcome the problems related to database overfitting. To improve classification accuracy, the database images were rotated with a fixed angle (90,180,270) and translated in a range of [-3 3] forming training datasets of image samples to overcome the limitations in terms of limited labeled datasets by radiologists. Also, in order to have an equal number of image samples for each class of malignancy suspiciousness - Data augmentation techniques were employed. Further, on the balanced dataset of candidate nodules, the DCNN were well-trained on features extracted as per the learning parameters.

### 3.3.4. Nodule Classification

Based on the nodule marked-up annotations by expert radiologists, each suspicious nodule greater than or equal to 3mm is categorized into one of the 5 stages of malignancy levels ranging from 1 to 5. Malignancy levels 1 and 2 samples are combined forming benign samples labeled as low malignancy nodules (LMNs) with "Nodules >= 3mm with Malignancy level 1 and 2". Malignancy levels 4 and 5 samples are combined forming malignant samples labeled as high malignancy nodules (HMNs) with "Nodules >= 3mm with Malignancy level 4 and 5". Samples with non-nodules greater than 3 mm and nodules less than 3mm are categorized as non-cancerous samples. Further, all the augmented NROI images are trained and tested using DCNN and are then classified into one of the 3 categories of classes specified according to the probability scores for each input as non-cancerous, benign or malignant samples.

Table 5: Performance measure on 3 LIDC datasets

| Dataset | Category | Images with Malignancy levels | No. of Nodules | Total Nodules | Accuracy with DCNN | Accuracy with Traditional hand crafted method |
|---|---|---|---|---|---|---|
| D1 | Benign | 1,2 | 278 | 1279 | 93.46% | 91.10% |
| | Malignant | 4,5 | 432 | | | |
| | Non-Cancerous | - | 569 | | | |
| D2 | Benign | 1,2,3 | 507 | 1508 | 87.91% | 85.43% |
| | Malignant | 4,5 | 432 | | | |
| | Non-Cancerous | - | 569 | | | |
| D3 | Benign | 1,2 | 278 | 1508 | 71.90% | 68.86% |
| | Malignant | 3,4,5 | 661 | | | |
| | Non-Cancerous | - | 569 | | | |

## 4. Experiments and Evaluation metrics

A detailed analysis of modeling the LIDC datasets based on the nodule malignancy levels was integrated into the study. The methodology was evaluated on 3 LIDC datasets using 10-fold cross-validation as shown in Table 5. The classification of nodules with D1 dataset was easy compared to the other 2 datasets as malignancy level 3 images were eliminated to distinguish the pulmonary nodules better. Also, the table depicts a high classification accuracy of 93.46% with DCNN compared to the traditional hand-crafted statistical and texture-based descriptors. For datasets, D2 and D3 having intermediate malignancy level 3 nodules either labeled as benign (LMNs) or malignant (HMNs) increased the difficulty in classifying the nodules.

Table 6: DCNN architecture performance measure using Adam Optimizer.

| # of Layers | Architecture | Alpha | Kernel size | Accuracy | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | epoch 20 | epoch 30 | epoch 40 | epoch 50 | epoch 100 |
| 8 | **12, 8, 6** | 0.1 | 5, 5, 5 | 0.913 | 0.919 | 0.919 | **0.934** | 0.919 |
| 8 | 12, 8, 4 | 0.1 | 5, 5, 5 | 0.904 | 0.906 | 0.910 | 0.914 | 0.915 |
| 8 | 12, 8, 6 | 0.1 | 5, 5, 3 | 0.904 | 0.906 | 0.906 | 0.911 | 0.903 |
| 8 | 12, 8, 4 | 0.1 | 5, 5, 3 | 0.885 | 0.910 | 0.900 | 0.899 | 0.921 |
| 10 | 12, 8, 6, 4 | 0.1 | 5, 5, 5, 5 | 0.915 | 0.900 | 0.899 | 0.910 | 0.900 |

Table 7: DCNN architecture performance measure using Sgdm Optimizer.

| # of Layers | Architecture | Alpha | Kernel size | Accuracy | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | epoch 20 | epoch 30 | epoch 40 | epoch 50 | epoch 100 |
| 8 | 12, 8, 6 | 0.1 | 5, 5, 5 | 0.885 | 0.890 | 0.893 | 0.899 | 0.899 |
| 8 | 12, 8, 4 | 0.1 | 5, 5, 5 | 0.895 | 0.898 | 0.894 | 0.893 | 0.894 |
| 8 | **12, 8, 6** | 0.1 | 5, 5, 3 | 0.885 | 0.905 | 0.899 | 0.904 | **0.919** |
| 8 | 12, 8, 4 | 0.1 | 5, 5, 3 | 0.881 | 0.883 | 0.885 | 0.885 | 0.889 |
| 10 | 12, 8, 6. 4 | 0.1 | 5, 5, 5, 5 | 0.881 | 0.890 | 0.896 | 0.885 | 0.889 |

The DCNN architecture with Adam training optimizer was tested upon different configurations of input parameters with performance values all above 88.5%, with a maximum classification accuracy of 93.46% as shown in Table 6. We also evaluated if there is any drastic change in performance measurements when the number of neurons varies in the hidden layers. But the variations in performance were quite stable and less than 0.4% for certain kernel size in correspondence to a particular epoch value. Finally, based on the

experiments, we fixed the hidden layer's neurons as [12, 8, 6] with the kernel size of [5, 5, 5] due to their relative stability while continuing our experiments.

In addition, the experiments were repeated to train CNN using Stochastic gradient descent momentum (sgdm) optimizer (Wikipedia contributors, 2019) having the configurations to evaluate its performance as shown in Table 7. The classification accuracy for all the tested combinations of input parameters using sgdm optimizer were all above 88.1%, with a maximum value of 91.9% as highlighted.



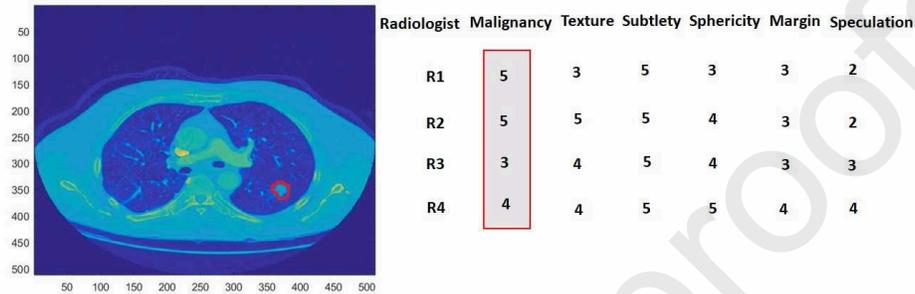| Radiologist | Malignancy | Texture | Subtlety | Sphericity | Margin | Speculation |
|---|---|---|---|---|---|---|
| R1 | 5 | 3 | 5 | 3 | 3 | 2 |
| R2 | 5 | 5 | 5 | 4 | 3 | 2 |
| R3 | 3 | 4 | 5 | 4 | 3 | 3 |
| R4 | 4 | 4 | 5 | 5 | 4 | 4 |

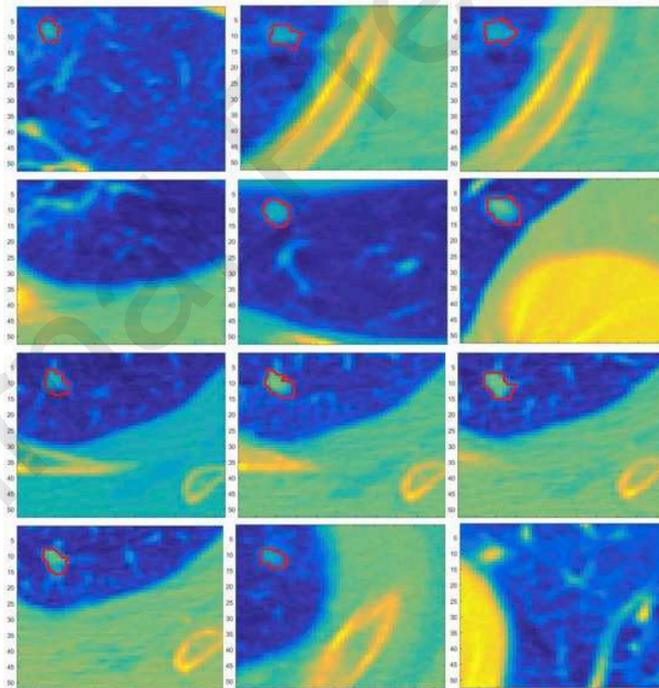Figure 5: Nodule diagnosis by 4 expert radiologists



Figure 6: Images false positively accepted as HMNs

To improve the performance of the existing methodology, we interpreted the inconsistency raised in the nodule classification. The features extracted from the statistical and texture based descriptors were observed with overlapping values. In order to identify the inconsistency raised with candidate nodules having malignancy levels ranging from 1 to 5, a detailed study on marked-up annotations was held. While

interpreting certain cases, the same candidate nodule region may be evaluated as malignancy level ranging from 1 to 5 by four different radiologists for the same CT scan. Example: One radiologist evaluates the candidate region as malignancy level 5 while the other radiologists have evaluated the same candidate region as malignancy level 3 and 4 correspondingly as shown in Figure 5. Due to incorrect malignancy suspiciousness, false categorizing the segmented images as benign or malignant samples resulted in low-performance accuracy overall exhibiting false positive classification of candidate nodule regions. To overcome the limitation, nodule grouping was manually performed to reduce the misclassification by considering the statistical and texture based descriptors. The images with overlapping texture and statistical features values were recognized and segregated according to their malignancy levels manually by replacing the candidate nodules in their corresponding categories.
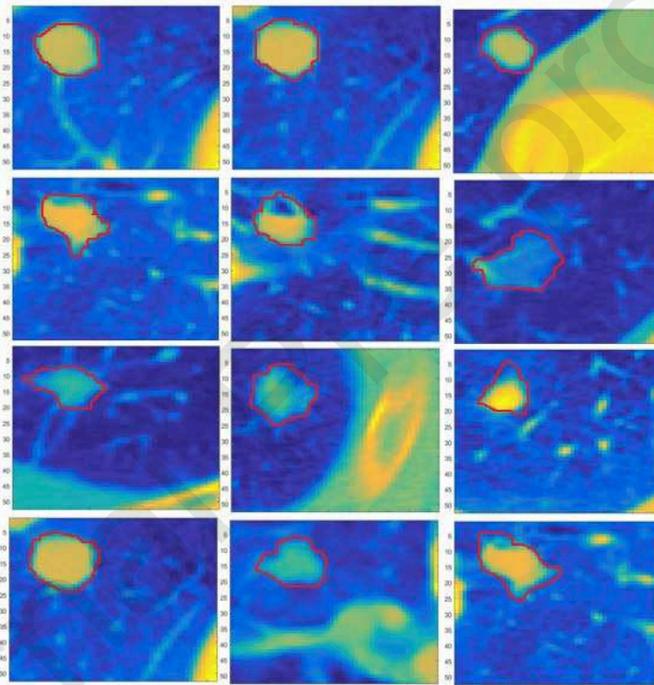


Figure 7: Images false positively accepted as LMNs

In addition, some candidate regions may be false positively considered as nodules by radiologists in their reviews with images close to the ribs, vessels or walls of the chest. The procedures in identifying features are time-consuming and may not guarantee good results if correlations between features are not properly considered. In most of the cases, machine learning models evaluate nodules by considering the "size" as a characteristic feature to distinguish between benign or malignant but not taking into consideration large variations of nodule patterns. However, dependence on the nodule size led to misclassification of small candidate nodules as benign and large candidate nodules as malignant in some cases. However, few candidate nodule regions were true positively considered as HMNs which were supposed to be true positively accepted as LMNs as depicted in Figure 6. Similarly, Figure 7 represents a few candidate nodule regions true positively

considered as LMNs but were supposed to be accepted as HMNs with high probability. Thus by grouping the nodules according to their patterns and malignancy levels, a total of 70 misclassified samples were potentially identified and tested in phases. In addition, the experiments were repeated to train the DCNN using stochastic gradient descent with momentum (sgdm) optimizer with the same input size, the number of layers (depth), same convolution filters to evaluate its performance. As a result, the methodology achieved good promising results with Adam optimizer compared to Sgdm as shown in Figure 8.
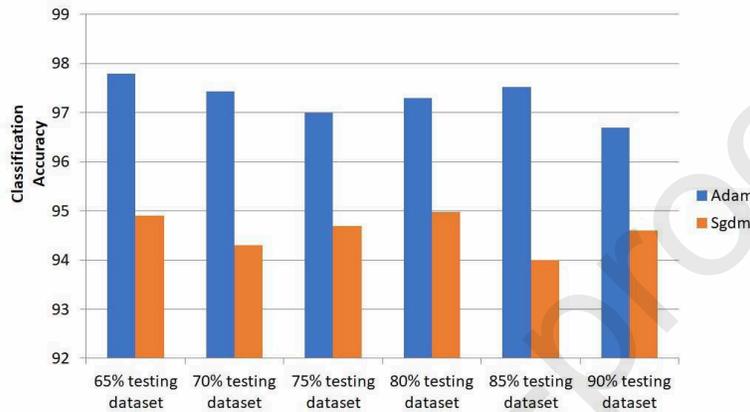


Figure 8: Proposed methodology tested using 'Adam' and 'Sgdm' optimizer

Table 8: Experiments conducted on the nodule grouped LIDC dataset

| LIDC Dataset | Methodology | Classification Accuracy |
|---|---|---|
| Initial Dataset | Traditional hand-crafted features | 91.10% |
| | DCNN | 93.46% |
| Nodule Grouped Dataset | Traditional hand-crafted features | 95.47% |
| | DCNN | **97.8%** |

A consistent improvement in the performance of the DCNN was observed using a nodule grouped dataset resulting in the highest classification accuracy of 97.8% when compared to the initial dataset images as shown in Table 8. Also, the performance of the DCNN is compared with the traditional hand-crafted methods for both datasets. In addition, the methodology is highlighted with datasets used and the results obtained from DCNN method are compared with related works as shown in Table 9. An accurate comparison is evaluated in terms of performance metrics with de Carvalho Filho et al. (2018), Xie et al. (2018), and Lakshmanaprabu et al. (2019). The proposed DCNN algorithm achieved the highest accuracy, sensitivity, specificity, and ROC on a balanced dataset resulting in better classification of nodules with reduced low false positives compared to other methods. Few methods use smaller datasets or may do not have an equal number of image samples for each class, which reduces the reliability for comparison. For performance evaluation, the balanced augmented image datasets with equal number of image samples are used in the study. The training

14

Table 9: Comparison of results with related works

| Research papers | Image Database | No. of images | Accuracy | Specificity | Sensitivity | ROC |
|---|---|---|---|---|---|---|
| de Carvalho Filho et al. (2018) | LIDC | 1404 nodules<br>394 - malignant<br>1011 - benign | 92.63% | 93.47% | 90.70% | 0.934 |
| Xie et al. (2018) | LIDC | 1972 nodules<br>648 - malignant<br>1324 - benign | 89.53% | 92.02% | 84.19% | 0.966 |
| Jiang et al. (2018) | LIDC | 1006 scans | - | - | 80.6 with 4.7 FP's<br>94 with 15.1 FP's | - |
| Xie et al. (2019) | LUNA 16 | 150414 images<br>150075 - non-nodules<br>339 - nodules | - | - | 86.42 | - |
| Huang et al. (2019) | LUNA 16 | 888 CT scans<br>223 nodules | 91.4 with 1 FP's<br>94.6 with 4 FP's | - | - | - |
| Lakshmanaprabu et al. (2019) | LIDC | 70 images<br>Normal - 27<br>Benign - 21<br>Malignant - 22 | 94.56% | 94.2% | 96.2% | - |
| **Proposed methodology** | LIDC | 1279 images<br>278 - benign<br>437 - malignant<br>569 - non-cancerous | **97.8%** | **97.2%** | **97.1%** | **0.995** |

and testing ratio samples were varied to see if the algorithm results in a consistent behavior as represented in Table 10.

Table 10: Performance evaluation with varying training and testing sets

| Training set | Testing set | No. of Nodules | Accuracy Results |
|---|---|---|---|
| 1279 images from LIDC dataset were used for training | 65% | 831 | 97.8% |
| | 70% | 895 | 97.43% |
| | 75% | 959 | 97.0% |
| | 80% | 1023 | 97.3% |
| | 85% | 1087 | 97.52% |
| | 90% | 1151 | 96.96% |
| | 95% | 1215 | 96.7% |

## 5. Results and Discussions

The proposed methodology evaluated the automated end-to-end learned features from the LIDC dataset for lung cancer diagnosis classifying each nodule according to the malignancy stages. The candidate nodules are extracted from the NROI and are fed to DCNN for training. False positives were observed during classification due to the misinterpretation of human airways, blood vessels, pulmonary trees as nodules. In order to reduce false positives, 10-fold cross-validation and nodule grouping were applied to image datasets
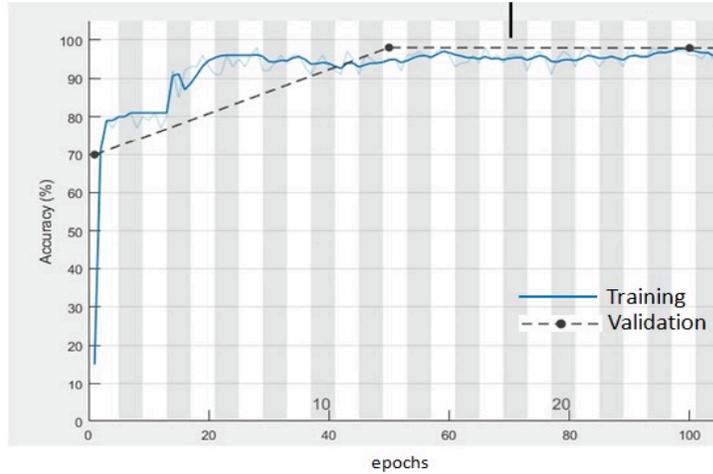
Figure 9: Learning curve of the proposed methodology using 'Adam' optimizer

forming two parts: training and testing folds for evaluation. In the training phase, 'Adam' optimizer was used to learn the weights as shown in Figure 9 with a learning rate of 0.1 for classifying the nodules with a higher degree of variation. A consistent improvement in the performance of the DCNN was observed using nodule grouped dataset and the nodules were classified not just by considering the "size" as a characteristic feature but also considers large variations in image patterns to distinguish malignant from benign. The overall accuracy of 97.8%, the specificity of 97.2%, the sensitivity of 97.1%, and AUC score of 0.9956 was recorded by DCNN method. The corresponding confusion matrix and receiver operating characteristic curve (ROC) curve are depicted in Figure 10 and Figure 11 respectively.

As the deep learning algorithm performs an end-end learning procedure, the only input passed is the re-sampled ROI images. However, Pre-processing the image datasets plays a vital role in our study enhancing the nodules for early detection of lung cancers. The pre-requisites for the deep structured scheme includes input data of the same size and a feasible procedure for pre-processing all the images in the datasets. The candidate nodules segmented differs in sizes having information concerning to nodule's shape and its surroundings. Using the deep learning scheme, the information around the nodules structure along with its shapes and sizes were extracted efficiently at the same time. Thus the algorithm detects different irregular structures nodules which include solitary nodules, pleural and juxta-pleural nodules and vascular nodules with high accuracy and are applicable to huge datasets. The entire neural network was trained on a CUDA enabled graphics card: Nvidia GeForce GTX 960, with Matlab 2018b version on a desktop machine with the memory of 8GB, 12(4C and 8G) core AMD A10 processor with a training time of 28 seconds and overall time complexity of 1.33 min on GPU mode for the classification results.

## 6. Conclusion

The proposed methodology implemented a DCNN for automatically learning the features extracted for lung nodule classification with reduced false positives. The study exhibited the classification accuracy of

Figure 10: Confusion matrix of the proposed methodology



Figure 11: ROC curve of the proposed methodology
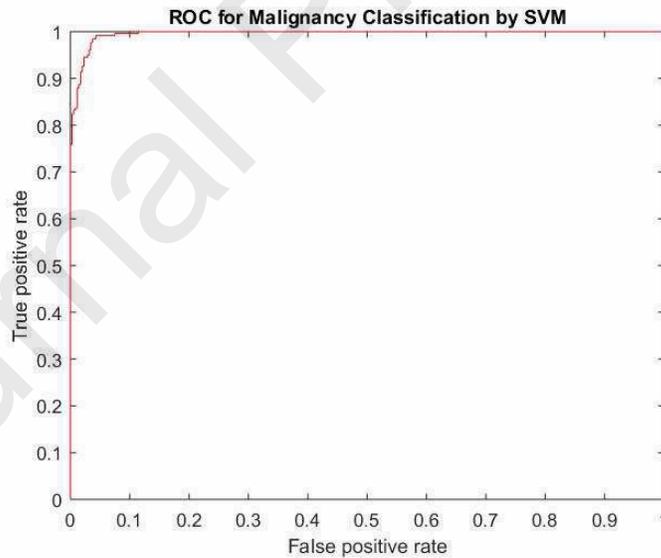
97.8%, the specificity of 97.2%, the sensitivity of 97.1%, and the AUC score of 0.9956. The results showed the DCNN with self-learned features achieves promising results with traditional hand-crafted methods and state of art methods. With a limited depth of layers, DCNN demonstrated potential feature learning and is robust to variable sizes of training and testing datasets.

17

The following observations were made as to future works

1. Although the results of the preliminary study are encouraging, we tested the images only to a limited number of deep learning layers. Increasing the number of layers in the CNN methodology can improve the performance of the diagnosis as the methodology is analogous to the human brain structure.

2. The optimal size of the input patch for deep learning algorithms are to be further investigated.

3. Also features from 3D input data are to be extracted to train the DCNN even though it could incur more network complexity.

4. Advancement in extracting context-based features (currently unable to extract) from input patches can increase the efficiency of the traditional hand-crafted methods.

## References

Armato III, M. G. B. L. M.-G. M. F. M. C. R. R. A. P. . C. L. P., Samuel G. (2015). The cancer imaging archive. data from lidc-idri. URL: `http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX`.

de Carvalho Filho, A. O., de Sampaio, W. B., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M. (2014). Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial intelligence in medicine*, *60*, 165–177.

de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M. (2017). Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm. *Journal of digital imaging*, *30*, 812–822.

de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M. (2018). Classification of patterns of benignity and malignancy based on ct using topology-based phylogenetic diversity index and convolutional neural network. *Pattern Recognition*, *81*, 200–212.

Huang, X., Sun, W., Tseng, T.-L. B., Li, C., & Qian, W. (2019). Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic ct scans using deep convolutional neural networks. *Computerized Medical Imaging and Graphics*, .

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*, 411–430.

Jacobs, C., van Rikxoort, E. M., Murphy, K., Prokop, M., Schaefer-Prokop, C. M., & van Ginneken, B. (2016). Computer-aided detection of pulmonary nodules: a comparative study using the public lidc/idri database. *European radiology*, *26*, 2139–2147.

Jacobs, C., van Rikxoort, E. M., Twellmann, T., Scholten, E. T., de Jong, P. A., Kuhnigk, J.-M., Oudkerk, M., de Koning, H. J., Prokop, M., Schaefer-Prokop, C. et al. (2014). Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical image analysis*, *18*, 374–384.

Jiang, H., Ma, H., Qian, W., Gao, M., & Li, Y. (2018). An automatic detection system of lung nodule based on multigroup patch-based deep learning network. *IEEE journal of biomedical and health informatics*, *22*, 1227–1237.

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

Lakshmanaprabu, S., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on ct images. *Future Generation Computer Systems*, *92*, 374–382.

Lampert, T. A., Stumpf, A., & Gançarski, P. (2016). An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, *25*, 2557–2572.

Lu, L., Yapeng, L., & Hongyuan, Z. (2018). Benign and malignant solitary pulmonary nodules classification based on cnn and svm. In *Proceedings of the International Conference on Machine Vision and Applications* (pp. 46–50). ACM.

Orozco, H. M., Villegas, O. O. V., Sánchez, V. G. C., Domínguez, H. d. J. O., & Alfaro, M. d. J. N. (2015). Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *Biomedical engineering online*, *14*, 9.

Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision* (pp. 525–542). Springer.

Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., & Summers, R. M. (2016). Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE transactions on medical imaging*, *35*, 1170–1181.

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, *69*, 7–34. URL: https://doi.org/10.3322/caac.21551. doi:10.3322/caac.21551.

da Silva, G. L., da Silva Neto, O. P., Silva, A. C., de Paiva, A. C., & Gattass, M. (2017). Lung nodules diagnosis based on evolutionary convolutional neural network. *Multimedia Tools and Applications*, *76*, 19039–19055.

Tajbakhsh, N., & Suzuki, K. (2017). Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: Mtanns vs. cnns. *Pattern recognition*, *63*, 476–486.

19

Tan, J., Huo, Y., Liang, Z., & Li, L. (2017). Apply convolutional neural network to lung nodule detection: Recent progress and challenges. In *International Conference on Smart Health* (pp. 214–222). Springer.

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., & Tian, J. (2017a). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical image analysis*, *40*, 172–183.

Wang, S.-H., Muhammad, K., Hong, J., Sangaiah, A. K., & Zhang, Y.-D. (2018). Alcoholism identification via convolutional neural network based on parametric relu, dropout, and batch normalization. *Neural Computing and Applications*, (pp. 1–16).

Wang, Y., Qiu, Y., Thai, T., Moore, K., Liu, H., & Zheng, B. (2017b). A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on ct images. *Computer methods and programs in biomedicine*, *144*, 97–104.

Wikipedia contributors (2018). Deep learning — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=875207371`. [Online; accessed 31-December-2018].

Wikipedia contributors (2019). Stochastic gradient descent — Wikipedia, the free encyclopedia. URL: `https://en.wikipedia.org/w/index.php?title=Stochastic_gradient_descent&oldid=919828819` [Online; accessed 27-October-2019].

Wu, H., Sun, T., Wang, J., Li, X., Wang, W., Huo, D., Lv, P., He, W., Wang, K., & Guo, X. (2013). Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography. *Journal of digital imaging*, *26*, 797–802.

Xie, H., Yang, D., Sun, N., Chen, Z., & Zhang, Y. (2019). Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, *85*, 109–119.

Xie, Y., Zhang, J., Xia, Y., Fulham, M., & Zhang, Y. (2018). Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct. *Information Fusion*, *42*, 102–110.

Xiuhua, G., Tao, S., Zhigang, L. et al. (2011). Prediction models for malignant pulmonary nodules based-on texture features of ct image. In *Theory and Applications of CT Imaging and Analysis*. InTech.

Yuan, X., Xie, L., & Abouelenien, M. (2018). A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. *Pattern Recognition*, *77*, 160–172.

Zhang, Y.-D., Dong, Z., Chen, X., Jia, W., Du, S., Muhammad, K., & Wang, S.-H. (2019). Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimedia Tools and Applications*, *78*, 3613–3632.