Improving the Pulmonary Nodule Classification Based on KPCA-CNN Model

Peichen Jiang*

Department of Physics Fudan University, Shanghai, China

* Corresponding Author Email: 18307110022@fudan.edu.cn

Abstract. Lung cancer mortality, the main cause of cancer-associated death all over the world, can be reduced by screening risky patients with low-dose computed tomography (CT) scans for lung cancer. In CT screening, radiologists will have to examine millions of CT pictures, putting a great load on them. Convolutional neural networks (CNNs) with deep convolutions have the potential to improve screening efficiency. In the examination of lung cancer screening CT images, estimating the chance of a malignant nodule in a specific location on a CT scan is a critical step. Low-dimensional convolutional neural networks and other methods are unable to provide sufficient estimation for this task, even though the most advanced 3-dimensional CNN (3D-CNN) has extremely high computing requirements. This article presents a novel strategy for reducing false positives in automatic pulmonary nodule diagnosis from 3-dimensional CT imaging by merging a kernel Principal Component Analysis (kPCA) approach with a 2-dimensional CNN (2D-CNN). To recreate 3-dimensional CT images, the kPCA method is utilized, with the goal of reducing the dimension of data, minimizing noise from raw sensory data while maintaining neoplastic information. The CNN can diagnose new CT scans with an accuracy of up to 90% when trained with the regenerated data, which is better than existing 2D-CNNs and on par with the best 3D-CNNs. The short duration of training, and certain accuracy shows the potential of the kPCA-CNN to adapt to CT scans with different parameters in practice. The study shows that the kPCA-CNN modeling technique can improve the efficiency of lung cancer diagnosis.

Keywords: deep learning; convolutional networks; kernel PCA; pulmonary nodules

1. Introduction

As one of the most lethal tumors, lung cancer accounts for over a quarter of all cancer-associated fatalities in the United States [1]. When compared to chest radiography, the NLST trial found that using low dose computed tomography (CT) for triennial screening rounds of high-risk patients reduced lung cancer death by 20% after seven years [2]. In the United States, low-dose CT lung cancer screening programs are now in operation. Other nations will almost certainly follow suit as a result of this study and following modeling work. The large volume of CT data which must be evaluated by physicians is one of the key hurdles in putting these screening regimens in place. Researchers have been working on this for the past two decades.

LUNA16 [3], a big database encompassing 888 sets of CT images and annotations from the publicly accessible LIDC-IDRI dataset [4-6], was provided in 2016 as a training and testing environment for automatic CT nodule recognition. Various algorithms were thoroughly evaluated and compared to one another in the field of false positive reduction.

CT scans were previously preprocessed with simple data amplification before being utilized to train 2-dimensional convolutional neural networks (2D-CNNs) and 3-dimensional convolutional neural networks (3D-CNNs). CUMedVis incorporated the predictions of three multi-level contextual 3D-CNNs, each trained on separate receptive field input images [7]. The nodules are augmented with translation and rotation. JackFPR employed a multi-level contextual 3D-CNN architecture and combined the three CNNs with a softmax layer at the end of the network [3]. DIAG CONVNET [8] uses 2D-CNNs with multi-view input. Streams of 2D-CNNs are directly coupled to a softmax layer in the multi-view CNNs. As data augmentation, random rotation and random zooming were applied to candidates. Wide residual networks were exploited by ZNET [9]. Plaques were independently recovered from axial, sagittal, and coronal images for each candidate and evaluated using the network.

The final forecast was calculated by averaging the network's expected output values for these three separate patches. Data augmentation (translating, zooming, rotation and flipping) was employed on both the training and test data sets to improve the test set results. CNNs with several slices were employed by CADIMI [3]. Three patches are obtained from three different coordinate positions for axial, sagittal, and coronal view respectively. The network was made up of two-dimensional convolutional networks with three convolutional layers in a row with max pooling. The last max-pooling layer was coupled to a softmax layer after following a fully connected layer. Data augmentation was used (vertical/horizontal flipping and random cropping). CUMedVis has the greatest Competition Performance Metric (CPM) score of 0.908, while ZNET has the lowest score of 0.758 among the projects listed above.

Projects that used 3D-CNNs outperformed projects that used 2D-CNNs in terms of performance. 3D-CNN, on the other hand, has very high processing requirements. In most cases, the high criteria are not met in practice. It is critical to develop a new algorithm that is comparable to 3D-CNN in terms of processing requirements and accuracy.

This study aims to propose a unique way for minimizing false positives in automatic pulmonary nodule identification from 3-dimensional computed tomography by merging a kernel Principal Component Analysis (kPCA) approach with a 2D-CNN.The algorithm's performance is reported and compared to prior results. This paper will use kPCA technique and a 2D-CNN to process these candidates produces excellent results.

2. Methodology

2.1. LUNA16 Dataset

A web-based contest performed in 2016 yielded the LUNA16 dataset. 888 CT scans in the open - source Lung Image Database Consortium (LIDC) collection were filtered out for the task [4]. The volumes had a resolution of 512512 in the transverse plane, an element spacing of 0.740.74 mm2, and a variable slice thickness of 2.5 mm. Team of highly experienced thoracic radiologists used a double manual labeling technique to collect the annotations of lung lesions. Each radiologist identified the lesions as non-nodules, nodules with diameters of 3 mm, or nodules of which diameters were all larger than 3 mm in the process. The reference criteria for the challenge were 1186 nodules larger than 3mm, upon which at least 3 radiologists concurred (i.e., ground truth). Non-nodules, nodules of which diameters were smaller than 3 mm, and nodules detected by less than 3 radiologists were all identified as irrelevant. The organizers gave a list of pre-screened submissions to competitors participating in false positive reduction track of the competition. The candidates were discovered using three current candidate identification algorithms [10-12], with 551, 065 candidates detecting 1120 out of 1186 ground truth negative nodules.

2.2. Method of Data Preprocessing

Each nodule has a positive and negative label, which may be found in the supplemental material provided by LUNA16 [3] for candidates. Patches in size of 50*50*11 were extracted centered on the candidate locations. Fig. 1 depicts the products of the preceding operation. Data augmentation for the positions of ground truth negative samples were performed to address the substantial imbalanced data between the ground truth negative samples and the false positive samples (1: 490). Methods of data augmentation include rotating the centroid coordinates 90, 180, and 270 degrees within the transverse plane, flipping the image according to each axis, and translating the coordinates by one voxel along each axis. To train the networks, 30 thousand data in total were collected. The intensities were standardized to a range of (0, 1).

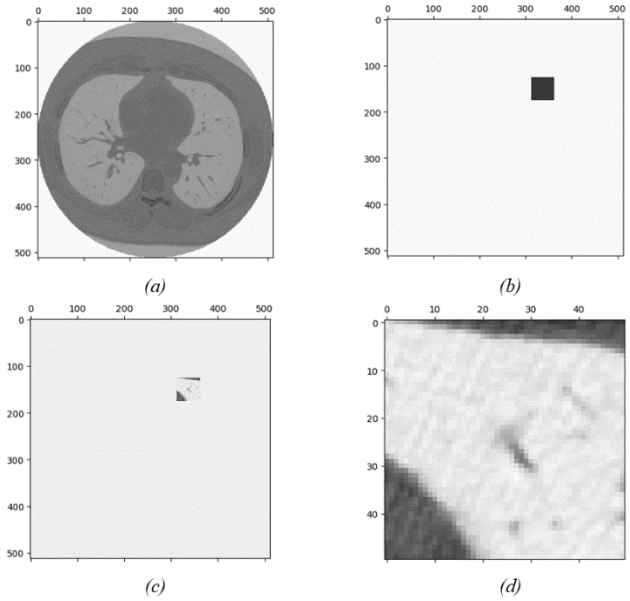


Figure 1. (a) A slice of the original CT scan; (b) A cutting mask based on tumor coordinate information.; (c) The image cut with mask; (d) The image ready for data augmentation

2.3. Kernel PCA Based Input Abstraction

For linear dimension reduction and feature derivation, principal component analysis (PCA) is often utilized [13]. The main correlation variables are transformed into the main orthogonal variables called principal components (PCs) through the diagonal of the correlation matrix. Then PCs are evaluated as a linear combination of input parameters.

The total range of parameters is equal to the sum of the eigenvalues, and the eigenvalues of the PCs are suggestive of correlated variances. PCA can only lower the dimension in a linear way. On the other hand, CT scan data has a more complex structure that is harder to explain in a linear subspace. As a nonlinear version of PCA, kernel PCA (kPCA) [14] is employed in this paper to accomplish nonlinear dimension reduction for 3-D CT images. The kernel is a higher-dimensional data translation, of which products can be utilized for linear PCA.

The diagonalization of an l-sample estimation of the covariance matrix can be used to describe the process of PCA in feature space F[15]. The diagonalization can be represented by (1):

$$\hat{C} = \frac{1}{l}\sum_{i=1}^{l} \Psi(x_i)\,\Psi(x_l)^T \tag{1}$$

Where $\Psi(x_i)$ are central nonlinear mappings of original variables $x_i \epsilon \mathbb{R}^n$. The following eigenvalue problem needs to be solved:

$$\lambda S = \hat{C}S, S\epsilon F, \lambda \geq 0 \tag{2}$$

As all the solutions S with $\lambda \geq 0$ are among $\Psi(x_1), \Psi(x_2), \cdots, \Psi(x_l)$. The problem is correspondingly defined by (3):

$$n\lambda\alpha = M\alpha \tag{3}$$

Where αexpressed the column vector that meets (4):

$$S = \sum_{i=1}^{l} \alpha_i \Psi(x_i) \tag{4}$$

And M is a kernel matrix that meets (5):

$$\iint M(x,y)f(x)f(y)dxdy > 0, \int f^2(x)dx < \infty \tag{5}$$

Where $M(x,y) = \sum_{i=1}^{\infty} \alpha_i \psi(x)\psi(y), \alpha \geq 0$. Then, as $\Psi(x)$ the projection of x onto the eigenvector S^k,the k^{th} nonlinear main component of may be determined.

$$\beta(x)_k = S^k\Psi(x) = \sum_{i=1}^{l} \alpha_i^k M(x_i, x) \quad (6)$$

The initial $p < l$ nonlinear components with the desired proportion of data variation are then picked. The original data series' complexity can be considerably reduced by doing so. The preprocessed data with size of 50*50*11 were transformed into data with size of 50*50*1 using the kPCA algorithm.

2.4. Construction of 2-D Convolutional Neural Network

2D convolution network consists of 2D convolutional layer, 2D max-pooling pool and fully connected layer, and SoftMax layer for final feature classification to extract multi-level data features. Each layer has a varied amount and type of connections between neurons.

2D Convolutional Layer: To make a 2D convolutional layer, the model starts by making a bunch of little 2D feature extractors (also known as kernels) that can scan their input and retrieve a bunch of elevated representations.

2D Max-pooling Layer: To sub-sample the 2D feature volume and provide local translation invariance in 2D space, 2D max-pooling layers are regularly placed between subsequent 2D convolutional layers.

Fully connected Layer: Compared with convolutional layer neurons, there are more dense connections between fully connected layer neurons. Each neuron is connected to neurons in the surrounding layers. This is not the case with the convolutional layers' local connection technique. These dense connections can help the extracted representations have a better representation capability.

Softmax Layer: The completely connected layer is fully connected to the softmax layer, and the classification result is produced. The number of neurons in softmax layer is consistent with that in classification

A 2-dimensional CNN was built to accept and utilize the kPCA-processed data. The overall architecture of CNN was designed to be relatively simple in order to research the function of KPCA and increase the program's running performance. The network's framework map is depicted by Fig. 2.

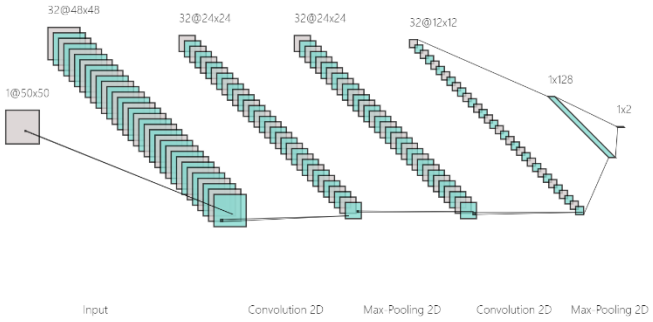


Figure 2.*The* framework *map of the 2D CNN.*

2.5. Workflow of kPCA-CNN Model

Fig. 3 depicts the kPCA-CNN model's approach for reducing false positives in lung nodules. This workflow involves two crucial steps: reducing the dimension of data while maintaining key information with kPCA and predicting new data with the CNN model.

To begin, the KPCA approach is applied to the preprocessed CT image to reduce the artifacts in the original picture by reducing the data dimension and deleting the less significant major components. As a result, the leftover 50*50 matrixes are used to create fresh inputs for the prediction model.

Whether the tumor is malignant or not in the new data can be detected after the CNN model has been completely trained and tested. As shown in the figure, the adoption of the kPCA approach distinguishes the kPCA-CNN model from other existing lung nodule recognition models. Image noise reduction, conserving the most critical information, and reducing calculation time for false positive reduction projects can all benefit from this phase.

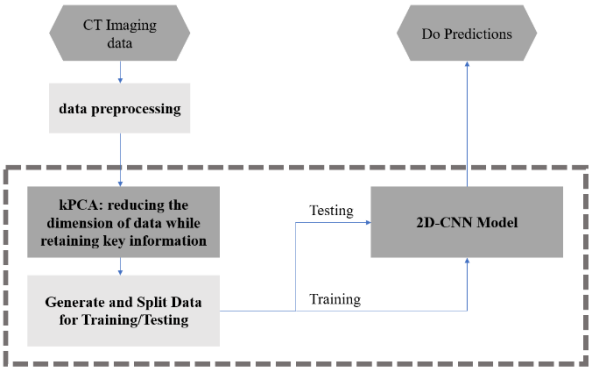


Figure 3. Workflow for reducing false positives in pulmonary nodules using the kPCA-CNN model. The important components of this proposed workflow are highlighted by the dotted box.

2.6. Training Process

A mini batch of training data was used to determine the parameters of kPCA. The number of epochs was set to 30 to ensure that the model is fully trained and avoid over fitting. The CNN part was created utilizing deep learning package provided by TensorFlow based on Python. Using an NVIDIA GeForce GTX 1050ti GPU, the whole network took roughly 0.5 hours to train.

3. Results

3.1. Evaluation metrics

3.1.1. Receiver Operator Characteristic Curve

The diagnostic capability of a binary classifier is graphically depicted by a Receiver Operator Characteristic (ROC) curve. Its roots are in signal detection theory, but it's now used in industries as diverse as medical science, tomography, weather events, and machine learning.

The connection between true positive rate (TPR) and false positive rate (FPR) is plotted to form the ROC curve (FPR). The FPR is the frequency of occurrence of negative samples wrongly projected as positive.

$$FPR = FP / (TN + FP) \tag{7}$$

Where FP represents the measurement of all the false positive samples, and TN represents the measurement of all the true negative samples. Similarly, the TPR is the percentage of all positive findings that were actually positive.

$$TPR = TP/(TP + FN) \tag{8}$$

Where FP represents the measurement of all the false positive samples, and TN represents the measurement of all the true negative samples.

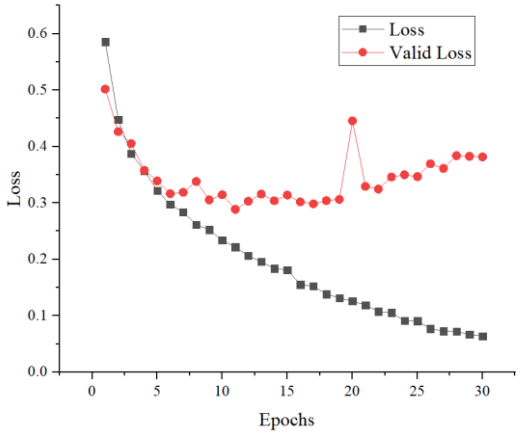
To compare different classifiers, a single metric of AUC can be used to describe each classifier's performance. The AUC stands for the area under the curve of ROC. AUC works well as a broad measure of prediction accuracy because it is similar to the odds that a positive example picked at random has a higher rank than a negative case chosen at random.

3.1.2. Free-response Receiver Operating Characteristic curve

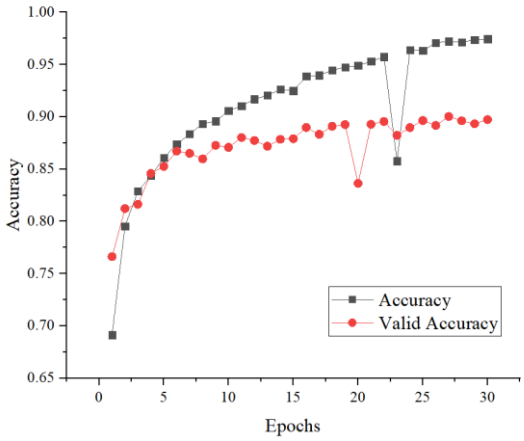
The ROC curve can be transformed into the free-response receiver operating characteristic (FROC) curve by modifying coordinate variables. Instead of FPR, the mean measurement of false positives occurred in each scan, or the false positives rate, is plotted on the x-axis. This graph is especially useful for unbalanced detection issues in which the number of positives P is much less than N. This would cause all relevant data to be pushed to the left of a ROC curve graphic, making interpretation difficult.

3.2. Accuracy and Loss

As shown in Fig. 4, during the 30 epochs of training, the training accuracy rate continues to rise, whereas the growth rate drops as the training degree increases. The training loss dropped over time, and the rate of reduction decreased as the training degree increased. The training accuracy rate reaches 0.98 and the valid accuracy rate hits 0.90 after the last epoch is completed. The model was fully trained during the training procedure, with no overfitting.



(a)



(b)

Figure 4. (a)Changes in loss during training of the kPCA-CNN. (b) Changes in accuracy during training of the kPCA-CNN. The training process contains 30 epochs in total.

3.3. Result of ROC

Fig. 5 shows the ROC curve for the prediction provided by the kPCA-CNN model. When the FPR exceeds 0.21, the sensitivity, or TPR, is greater than 0.9. The sensitivity is larger than 0.8 when the FPR exceeds 0.07. The AUC value is 0.94.

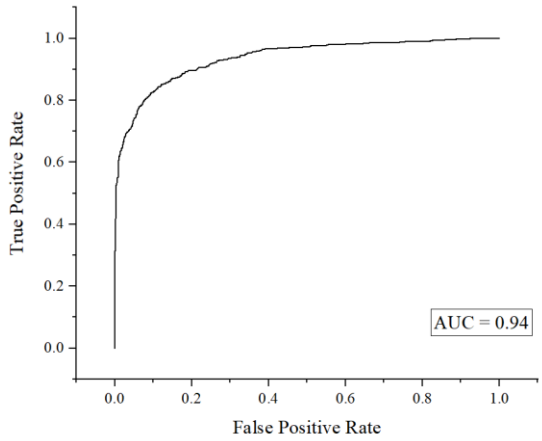


Figure 5. The ROC of the kPCA-CNN model. ROC: Receiver Operating Characteristic.

3.4. Result of FROC

Fig. 6 depicts the FROC curve for the forecast provided by kPCA. The FROC curve of a prediction generated by a 2D-CNN model, which is trained from raw data without being processed by kPCA and uses the same CNN, is also shown in the figure. When the false positives rate is 8.0, kPCA-CNN has a sensitivity of 0.97, while kPCA-CNN has a sensitivity of 0.84. When the false positives rate is 4.0, kPCA-CNN has a sensitivity of 0.90, while kPCA-CNN has a sensitivity of 0.73. The sensitivity of kPCA-CNN is more than 0.83 when the false positives rate is greater than 0.07.

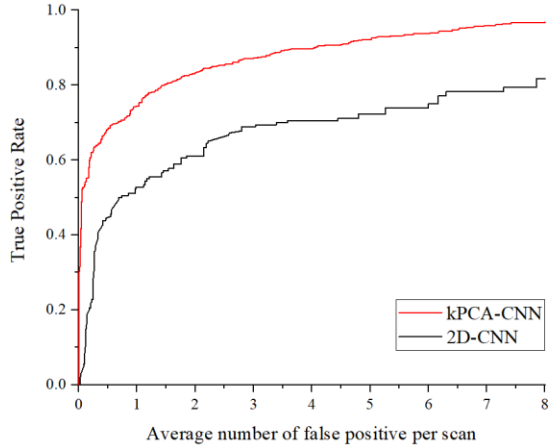


Figure 6. The FROC of the kPCA-CNN model and the FROC of a 2D-CNN model without kPCA part. FROC: Free-response Receiver Operating Characteristic.

3.5. Comparison with 2D-CNN model

The AUC of kPCA-CNN may reach 0.94, which is a good prediction performance for such a simple model, according to the computations. It means that a positive example chosen at random has a 94% chance of being ranked higher by the system than a negative case chosen at random. By comparison, the AUC of the 2D-CNN is 0.84.

The FROC of kPCA-CNN was significantly greater than that of 2D-CNN. The FROC of kPCA-CNN is 28 percent, 22 percent, and 15 percent greater than the FROC of 2D-CNN when the false positives rate is 2.0, 4.0, and 8.0, respectively. The 2D-CNN was built using discrete 2D slices in areas where volumetric contextual information could not be fully investigated. The prediction result demonstrated that kPCA-CNN captured volumetric contextual information that 2D-CNN missed.

3.6. Comparison with 3D Models

Because kPCA-CNN uses a single CNN for both training and judgment, three 3D-CNNs used in the model [7] with the highest score in the luna16 competition were chosen to compare with kPCA-CNN.

The detection sensitivity of various network designs at various false positives rates determined by the competition performance metric (CPM) score are listed in Table. The data are extracted from FROC of these models. The average detection sensitivity for seven predetermined false positives rates was calculated as a CPM: 1/8, 1/4, 1/2, 1, 2, 4, and 8 average false positives per scan [16].

As demonstrated in Table. I, the CPM of kPCA-CNN is greater than that of CUMedVis-Archi1 and can be compared to those of CUMedVis-Archi2 and CUMedVis-Archi3. When the different false positives rate is 0.125, 4 and 8 respectively, kPCA-CNN has the highest sensitivities of the models, with a value of 0.55, 0.90 and 0.97. This demonstrates that, with the help of KPCA technology, the prediction accuracy of 2D-CNN has surpassed that of 3D-CNN.

Table 1. The Sensitivities of KPCA-CNN and 3D-CNNs at False Positives Rates Assigned by The Competition Performance Metric (CPM) Score

Model	CNN	False Positives Rates							CPM Score
		0.125	0.25	0.50	1.00	2.00	4.00	8.00	
CUMedVis-Archi1 [7]	3D	0.46	0.58	0.68	0.78	0.85	0.88	0.91	0.73
CUMedVis-Archi2 [7]	3D	0.48	0.61	0.74	0.84	0.88	0.90	0.92	0.77
CUMedVis-Archi3 [7]	3D	0.55	0.65	0.74	0.82	0.86	0.90	0.92	0.78
kPCA-CNN	2D	0.55	0.63	0.69	0.75	0.85	0.90	0.97	0.76

4. Discussions

4.1. Deficiency of KPCA-CNN

Because kPCA-CNN only has one CNN for training and judgment, it has the same issues as other models with only one CNN. KPCA-CNN has a poor sensitivity ranging from 0.55 to 0.69 under circumstances of exceptionally low false positives rate. The low sensitivity of kPCA-CNN will limit its use in circumstances when there are few false positives.

4.2. Future Applications

It has been demonstrated that using a system that integrates numerous CNNs can successfully enhance the sensitivity of identifying pulmonary modules when the false positives rate is exceptionally low, which is one of the key limitations of using CNNs in medical picture interpretation (Dou and coworkers, 2016).

KPCA CNN uses a unique feature extraction strategy that complements other CNN models, as compared to pure CNN. kPCA-CNN has a higher sensitivity than other single CNN models when the false positives rate exceeds a particular threshold. Using correct weighting functions or full connection layers to combine the prediction results of kPCA-CNN and other deep learning neural networks into a consistent model might result in greater prediction accuracy.

5. Conclusions

This study offers a novel strategy for reducing false positives in automated pulmonary nodule diagnosis from volumetric CT images by merging a kernel PCA algorithm with a 2-dimensional convolutional neural network (2D-CNN). The kPCA approach is used to recreate 3-dimensional CT scans, with the goal of decreasing the dimension of data, reducing noise from raw data while preserving neoplastic information. The CNN can diagnose new CT scans with an accuracy of up to

90% when trained with the regenerated data. The AUC is up to 0.94, and the CPM score is 0.76 which is better than 2D-CNNs and on par with the best 3D-CNNs. The short duration of training, and certain accuracy shows the potential of the kPCA-CNN to adapt to CT scans with different parameters in practice. On the basis of retaining high diagnosis accuracy, the research reveals that kPCA-CNN modeling technology may considerably minimize the hardware requirements and time consumption of deep learning modeling.

References

[1] American Cancer Society, Cancer facts and figures 2016, 2016.

[2] National Lung Screening Trial Research Team, D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, J. D. Sicks, "Reduced lung-cancer mortality with low-dose computed tomographic screening," N Engl J Med. vol. 365(5), pp. 395-409, 4 Aug 2011.

[3] A. A. A. Setio, A. Traverso, T. de Bel, et al, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," Med Image Anal, vol. 42, pp. 1-13, Dec 2017.

[4] S. G. Armato III, G. McLennan, L. Bidaut, 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(2), pp. 915-31, Feb 2011.

[5] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," Journal of Digital Imaging," vol. 26, pp. 1045–1057, 2013.

[6] S. G. Armato III, G. McLennan, L. Bidaut, et al, "Data from LIDC-IDRI," 2015.

[7] Q. Dou, H. Chen, L. Yu, J, Qin, P. A. Heng, "Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection," IEEE Trans Biomed Eng, vol. 64(7), pp. 1558-1567, Jul 2017.

[8] A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. Wille, M. Naqibullah, C. I. Sanchez, B. van Ginneken., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," IEEE Transactions on Medical Imaging, 2016.

[9] S. Zagoruyko, N. Komodakis, "Wide residual networks," arXiv preprint arXiv:arXiv:1605.07146, 2016.

[10] C. Jacobs, E. M. van Rikxoort, T. Twellmann, E.T. Scholten, P.A. de Jong, J.M. Kuhnigk, M. Oudkerk, H.J. de Koning, M. Prokop, C. Schaefer-Prokop, B. van Ginneken, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," Med Image Anal, vol. 18(2), pp. 374-84, Feb 2014.

[11] A. A. Setio, C. Jacobs, J. Gelderblom, B. van Ginneken, "Automatic detection of large pulmonary solid nodules in thoracic CT images," Med Phys, vol. 42(10), pp. 5642-53, Oct 2015.

[12] K. Murphy, B. van Ginneken, A. M. R. Schilham, B. J. de Hoop, H. A. Gietema, M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," Medical Image Analysis, vol. 13, no. 5, pp. 757-770, 2009.

[13] S. Wold, K. Esbensen, P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems 2, pp. 37–52, 1987.

[14] B. Schölkopf, A. Smola, K. R. Müller, "Kernel Principal Component Analysis," In: Gerstner, W., Germond, A., Hasler, M. and Nicoud, J.D., Eds., "Artificial Neural Networks—ICANN97, Springer, Berlin, pp. 583-588, 1997.

[15] H. Ince, T. Trafalis, "Kernel principal component analysis and support vector machines for stock price prediction," IIE Trans, vol. 39, pp. 629–637, 2007.

[16] M. Niemeijer, M. Loog, M. D. Abràmoff, M. A. Viergever, M. Prokop and B. van Ginneken, "On Combining Computer-Aided Detection Systems," in IEEE Transactions on Medical Imaging, vol. 30, no. 2, pp. 215-223, Feb. 2011.