ML Classification Based Detection of Cancerous Nodules From Radiology Images

** Senthil Kumar T. K.*
*** Subhabaha Pal*

Abstract

The appearance of pulmonary nodule is the early manifestation of lung cancer and early detection from the lung CT scan images leads to better treatment. Machine Learning can help in the detection of cancerous nodules. The present paper gives an overview of the machine learning-based modelling for detection of cancerous pulmonary nodules from CT scan images. The paper discusses the whole process involving transformation of the CT scan image, extraction of the nodules and conversion of these into quantitative form and ultimate fitting of suitable classification model for detection of cancerous pulmonary nodule after labelling the nodules.

Keywords : Cancerous nodule detection, Lung CT Scan, ML Application in Radiology

I. INTRODUCTION

Lung cancer is one of the most fatal among all the malignant tumors. The pulmonary nodule is the early manifestation of lung cancer, and plays an important role in its discovery, diagnosis, and treatment [1]. There are two types of growths in the lung called pulmonary nodules: benign or malignant [2]. The causes of pulmonary nodules include infections and non-infectious diseases such as sarcoidosis [2]. A pulmonary nodule is a small round or oval-shaped growth in the lung. It may also be called a "spot on the lung" or a "coin lesion" [2]. Pulmonary nodules are generally smaller than three centimetres in diameter. If the growth is larger than that, it is called a pulmonary mass and is more likely to represent a cancer than a nodule. Countless pulmonary nodules are discovered each year during chest X-rays or CT scans. Most nodules are non-cancerous (benign). A solitary pulmonary nodule is found on upto 0.2% of all chest X-ray films. Lung nodules can be found on upto half of all lung CT scans. Risk factors for malignant pulmonary nodules include a history of smoking and older age [2].There are two main types of pulmonary nodules, malignant (cancerous), and benign (noncancerous). Over 90% of pulmonary nodules that are smaller than two centimetres in diameter are benign [2].

The technology of medical imaging has encountered a rapid development in recent years. Thus, the amount of pulmonary nodules that can be discovered are rising, which means that even tiny or minor change in lungs can be recorded by the CT images [1]. As all nodules are not cancerous, early detection of lung cancer depends on effective and accurate identification of cancerous nodules in the early stage. Several active researches are happening in the field of computer-based detection of cancerous pulmonary nodules from the CT scan images. Several research works have been done in this direction. [3, 4, 1, 5, 6, 7] are some of the works in this direction. Machine learning based lung cancer prediction models have been proposed to assist in managing incidental or screen detected indeterminate pulmonary nodules. Such systems may be able to reduce variability in nodule classification, improve decision making, and ultimately reduce the number of benign nodules that are needlessly followed or worked-up [3].

Manuscript received December 15, 2018; revised January 5, 2019; accepted January 10, 2019. Date of publication March 6, 2019.
*T. K. S. Kumar is Senior Faculty with Data Science and Machine Learning, Data Science, MAHE South Bangalore Campus (Manipal ProLearn), 3rd Floor, Salarpuria Symphony, 7, Service Road, Pagathinagar, Electronic City, Bengaluru, Karnataka - 560100.
(e-mail: senthil.kumar@manipalglobal.com)
**S. Pal is Senior Faculty with Data Science and Machine Learning, MAHE South Bangalore Campus (Manipal ProLearn), 3rd Floor, Salarpuria Symphony, 7, Service Road, Pagathinagar, Electronic City, Bengaluru, Karnataka - 560100. (e-mail: subhabaha.pal@manipalglobal.com)
DOI:10.17010/ijcs/2019/v4/i2/144271

[3] provides an overview of the main lung cancer prediction approaches proposed to date and highlight some of their relative strengths and weaknesses. The paper also discusses some of the challenges in the development and validation of such techniques and outline the path to clinical adoption. The paper proposed the use of Convolutional Neural Network models for detecting the cancerous pulmonary nodules from CT Scan images and opined that the models show high amount of accuracy in detection of cancerous nodules. [4] proposed a computer-aided diagnosis (CADx) method for classification between benign nodule, primary lung cancer, and metastatic lung cancer, and evaluated the following: (i) the usefulness of the deep convolutional neural network (DCNN) for CADx of the ternary classification compared with a conventional method (hand-crafted imaging feature plus machine learning), (ii) the effectiveness of transfer learning, and (iii) the effect of image size as the DCNN input.

[1] proposes a pulmonary nodule computer aided diagnosis (CAD) based on semi-supervised extreme learning machine(SS-ELM). As per [1], the proposed model based on CAD system based on SS-ELM for detection of pulmonary nodule achieves better generalization performance at faster learning speed and higher testing accuracy than ELM (Extreme Learning Model), SVM (Support Vector Machine), PNN (Probabilistic Neural Network), and MLP (Multi-Layer Perceptron).

The present paper gives an overview of the machine learning-based model for detection of cancerous pulmonary nodules. The paper delineates how a CT scan image can be processed to extract the nodules from the image and how these nodules can be quantitatively defined through multiple numerical features which can be used for labelling. The labelled data can be used for fitting a random forest model. The present paper discusses the whole process involving transformation of the CT scan image, conversing it into quantitative form, and ultimate fitting of suitable classification model for detection of cancerous pulmonary nodule.

II. DATASET

The dataset used in this work is downloaded from the cancer imaging archive (TCIA) public database [5]. All the images are in DICOM format. The CT scans were taken using GE Medical and Philips CT scanners with different X-ray tube current and exposure. Each image size is 512 x 512. Totally 4682 CT images are there in this database for 61 patients. CT slice thickness is between 3 and 6 mm.

III. METHODOLOGY

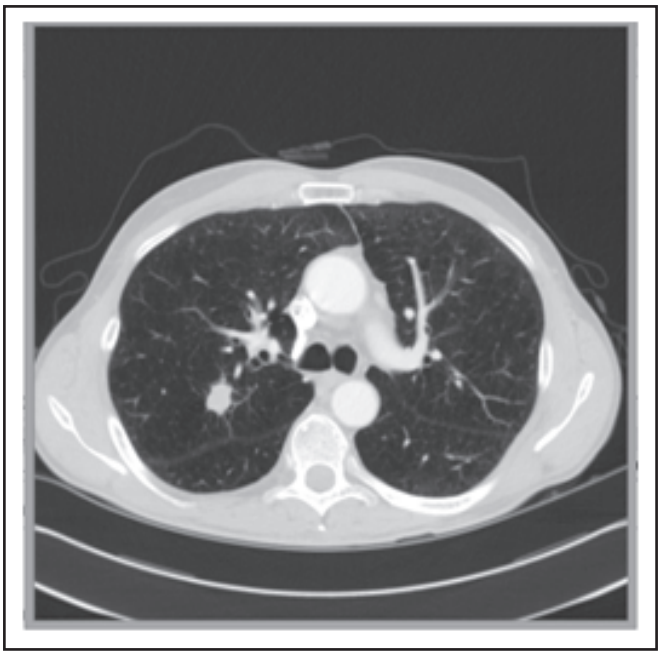

Fig. 1. Radiology Image Converted to JPG Image

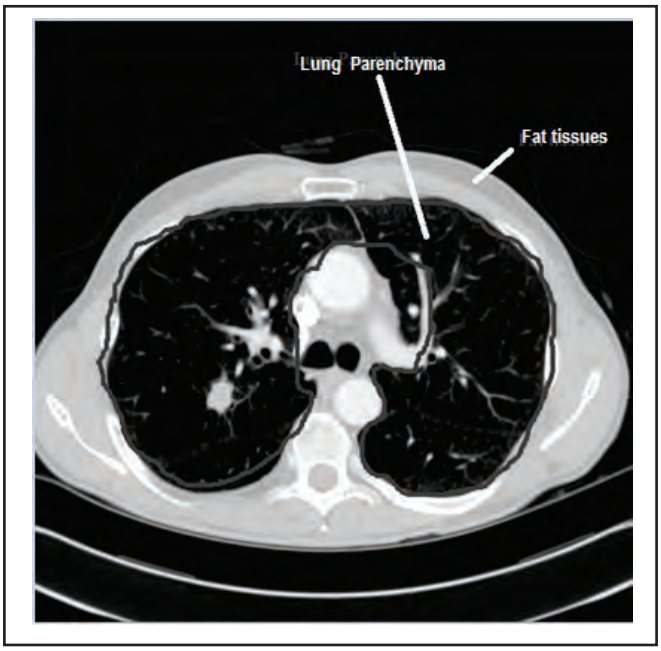

Fig. 2. CT Scan Image With Demarcation of Parenchyma and Fat Tissues

In the CT scan image of lung (Fig. 2), marked internal portion (lung parenchyma) is the region of interest where cancerous nodules are usually found. The major objective is identifying cancerous nodules in the region.

It is needed to make certain transformations in the

image by giving certain threshold value. If the pixel value is less than the threshold value, it becomes white and if it is more that the threshold value, it becomes black (Fig. 3). The image obtained through the threshold-based transformation is inverted (Fig. 4). The image inversion is performed for feature extraction.

Further processing of the image is done by removal of the border region from the transformed image (Fig. 5)

Fig. 3. Threshold Based Transformed Image

Fig. 4. Inverted Image for Feature Extraction

Fig. 5. Removal of Border Region

Fig. 6. Creation of Lung-Mask Template

The holes in the parenchyma region are filled to create the lung-mask template (Fig. 6).

The unwanted white-cluster in the bottom is removed to get lung-lobe portion alone (Fig. 7). The lung-lobe mask is ANDed with the original lung image (Fig. 8).

The white clusters inside the lung-lobe/parenchyma are the regions of interest for the study. Those white clusters have been segmented out from the image

Fig. 7. Modified Lung-Mask Template

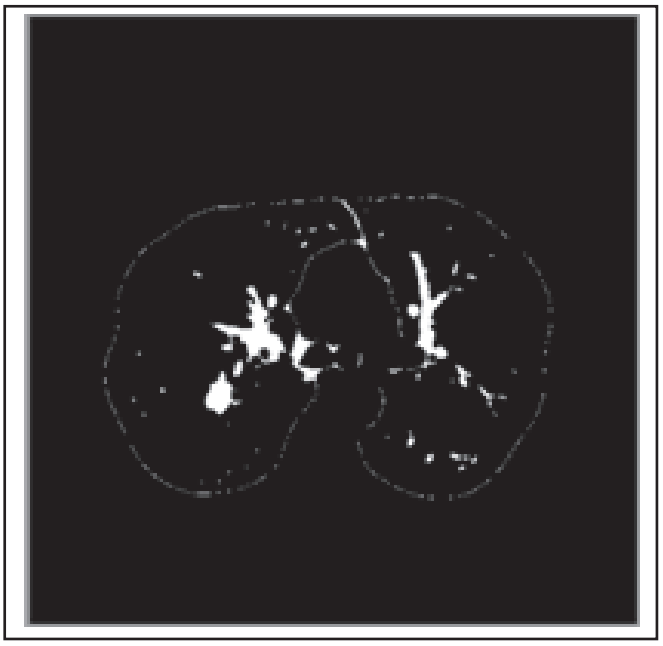

Fig. 9. Image Showing White Clusters

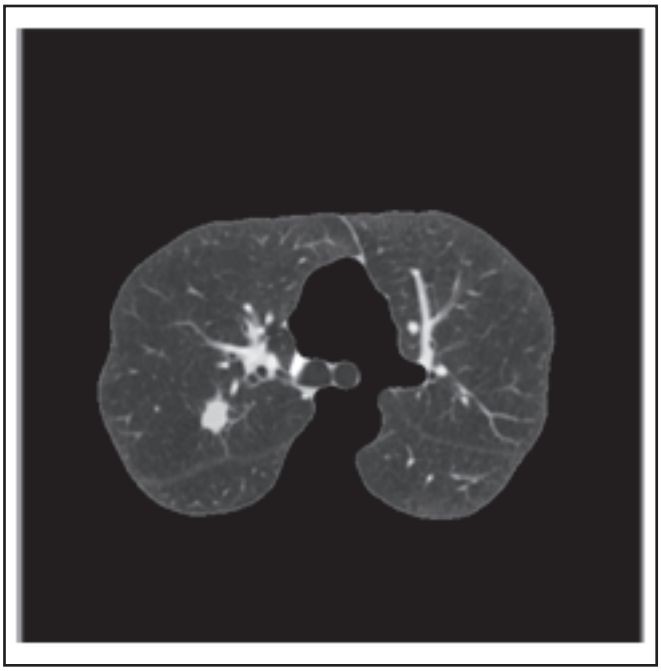

Fig 8. ANDed Lung Image

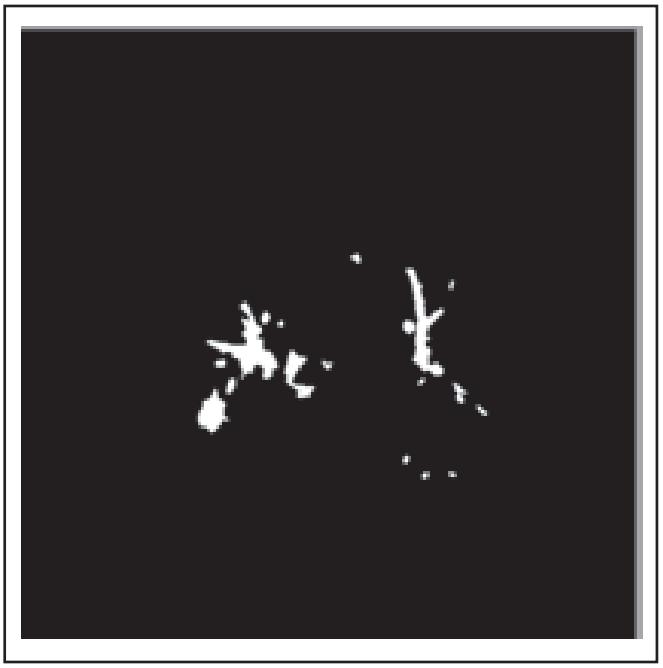

Fig. 10. Final Clusters

(Fig. 9). There are many small white clusters which don't have any significant effect in diagnosing the malignancy of the cluster. The small clusters are removed keeping only the clusters which are useful for detection of the malignancy. Fig. 10 presents the final nodules obtained from the lung image.

Now, these discrete clusters are defined quantitatively in the following table in terms of nine parameters (Aspect Ratio, Area, Extent, Hull_area, Solidity, Equi-diameter, Angle, Centroid_X, and Centroid_Y).

Altogether 18 nodules have been detected in the current CT scan image. Out of these 18 nodules, nodule 14 has been detected as cancerous by the medical expert. Hence, the corresponding nodule has been labelled as '1',

TABLE I.
DEFINITION OF THE NODULES IN QUANTITATIVE FORM

Nodules	Aspect Ratio	Area	Extent	Hull area	Solidity	Equi-diameter	Angle	Centroid_X	Centroid_Y
Nodule 1	1	30.5	0.376543	35.5	0.859155	6.23168	150.249	279.659	189.854
Nodule 2	0.348315	788.5	0.285792	1845	0.427371	31.6852	159.649	335.058	248.16
Nodule 3	0.5	11	0.34375	13.5	0.814815	3.74241	13.3688	358.421	212.526
Nodule 4	0.9375	1041.5	0.271224	2109.5	0.493719	36.4154	141.711	190.555	264.887
Nodule 5	0.7	42	0.6	43.5	0.965517	7.31273	13.31	203.741	239.519
Nodule 6	0.833333	11.5	0.383333	12	0.958333	3.82652	139.736	216.5	244.611
Nodule 7	0.833333	71.5	0.595833	74	0.966216	9.54131	156.121	323.244	247.081
Nodule 8	0.625	326.5	0.3265	614	0.531759	20.389	159.57	229.711	286.905
Nodule 9	1.5	29.5	0.546296	31	0.951613	6.12867	81.095	166.575	277.8
Nodule 10	1.28571	22	0.349206	29	0.758621	5.29257	119.901	255.094	278.969
Nodule 11	0.5	52	0.530612	55.5	0.936937	8.13686	18.0453	175.104	296.403
Nodule 12	1	11.5	0.319444	11.5	1	3.82652	45	333.722	292.722
Nodule 13	0.666667	65.5	0.303241	99	0.661616	9.13221	162.512	365.172	302.483
Nodule 14	0.702703	484.5	0.503638	576.5	0.840416	24.8372	9.18256	159.629	318.956
Nodule 15	1	21	0.259259	24	0.875	5.17088	136.903	384.129	317.387
Nodule 16	0.714286	18	0.514286	18.5	0.972973	4.78731	20.5219	320.923	357.962
Nodule 17	0.833333	14	0.466667	14.5	0.965517	4.22201	28.568	336.905	371.476
Nodule 18	1.75	11	0.392857	12	0.916667	3.74241	110.04	359	370.5

while the other nodules have been labelled as '0'.

The nodules are extracted in similar manner from the CT Scan images of 6 patients and each nodule is labelled as '1' (cancerous) or '0' (non-cancerous) under the supervision of medical expert. The Random Forest Model is used to build a model for automatic detection of cancerous nodules.

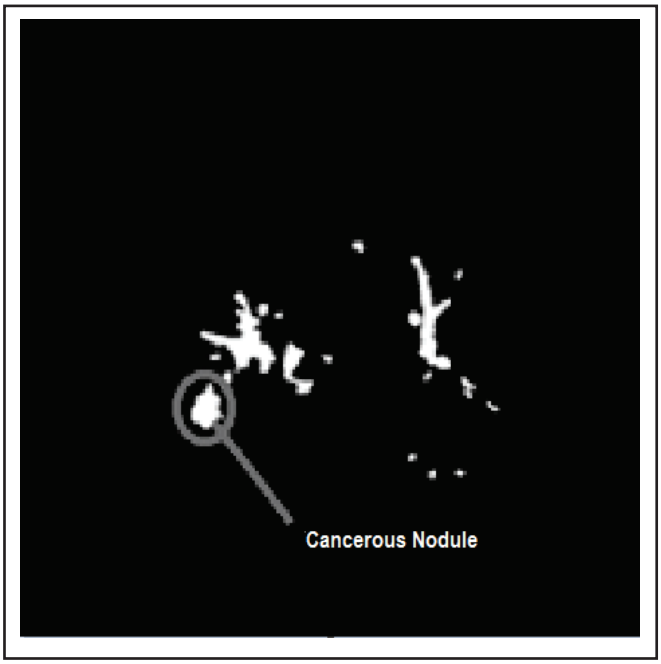

Fig. 11. Cancerous Nodule Detection

IV. RESULTS

The Random Forest model gives the accuracy of 0.96 on the test data-set extracted from the whole data-set in the 70:30 ratio. The model can be further improved including the CT scan images of more patients and using improved Neural Network models like Multi-Layer Perceptron Network (Feed Forward Network) Models. This machine learning model will help in automatic detection of cancerous nodules from the CT Scan images of the patients.

V. CONCLUSION

This paper gives an overview on the Machine Learning technique used for the detection of the cancerous nodules from the CT scan images of the lung. The use of ML techniques in medical diagnostics is in its nascent stage. Inclusion of more trained data (more

labelled CT scan images) and use of neural network models can enhance performance of the machine learning models and will be able to provide more accurate results.The use of ML technique in cancerous nodule detection from radiology images is a promising field, and shows huge potential for practical application in the time to come.

REFERENCES

[1] T. Kadir and F. Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques," *Translational Lung Cancer Res.*, 2018 June, vol. *7*, no. *3*, pp. 304-312. doi: 10.21037/tlcr.2018.05.15.[

[2] M. Nishio, O. Sugiyama, M. Yakami, S. Ueno, T. Kubo, T. Kuroda, and K. Togashi, "Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning," *PLoS One,* vol. *13*, no. *7*, 2018. doi: 10.1371/journal.pone.0200721.

[3] Z. Wang, J. Xin, P. Sun, Z. Lin, Y. Yao and X. Gao, "Improved lung nodule diagnosis accuracy using lung CT images with uncertain class," *Computational Methods Programs Biomed*, vol. *162*, 2018, pp. 197-209. doi: 10.1016/j.cmpb.2018.05.028.

[4] Cleveland Clinic, "Pulmonary Nodules." [Online]. Available: https://my.clevelandclinic.org/health/diseases/14799-pulmonary-nodules

[5] K. Clark, B. Vendt, K. Smith, K., J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. of Digital Imaging*, vol. *26*, no. *6*, pp.1045-1057, 2013. doi: 10.1007/s10278-013-9622-7

[6] S. Krishnamurthy, G. Narasimhan and U. Rengasamy,"Three-dimensional lung nodule segmentation and shape variance analysis to detect lung cancer with reduced false positives," *Proc. of the Institution of Mech. Engineers*, Part H: *J. of Eng. in Medicine*, vol. *230*, no. *1*, pp. 58-70, 2016. doi: 10.1177/0954411915619951

[7] S. Krishnamurthy, G. Narasimhan and U. Rengasamy, "Early and accurate model of malignant lung nodule detection system with less false positives," *Brazilian Archives of Biology and Technol.*, vol. *61*, pp. 1-12, 2018. doi: http://dx.doi.org/10.1590/1678-4324-2018160536

[8] S. Krishnamurthy, G. Narasimhan and U. Rengasamy, "Lung nodule growth measurement and prediction using auto cluster seed K-means morphological segmentation and shape variance analysis," *Int. J. of Biomedical Eng. and Technol.,* vol. *24*, no. *1*, pp. 53-71, 2017. doi: 10.1504/IJBET.2017.083818

About the Authors

Senthil Kumar is Senior Faculty, Data Science and ML at Manipal Group. He is on the verge of completion of Ph.D. in medical image classification from Anna University. He has published more than 14 research papers. His research areas include Deep Learning applications in solving real life business problems.

Dr. Subhabaha Pal has been honoured by Analytics India Magazine as among the '20 Most Prominent Data Science and Machine Learning Academicians in India – 2018'. He has also been awarded 'Best University Mentor' by Data Science Society, Bulgaria. He completed his Ph.D. in Non-life Insurance Analytics from the University of Calcutta. He is a versatile researcher who has authored 36 research papers and 3 books. These have been published. His research interests include applications of Machine Learning in solving real life business problems.