SUPERVISED MACHINE LEARNING ALGORITHMS: DEVELOPING AN EFFECTIVE USABILITY OF COMPUTERIZED TOMOGRAPHY DATA IN THE EARLY DETECTION OF LUNG CANCER IN SMALL CELL

Pushkar Garg

Delhi Public School, R.K. Puram, New Delhi

ABSTRACT

Cancer-related medical expenses and labour loss cost annually $10,000 million worldwide. Lung cancer-related deaths exceed 70,000 cases globally every year. Furthermore, 225,000 new cases were detected in the United States in 2016, and 4.3 million new cases in China in 2015. Statistically, most lung cancer-related deaths were due to late-stage detection. Like other types of cancer, early detection of lung cancer could be the best strategy to save lives. In this paper, we propose a novel neural-network based algorithm, which we refer to as entropy degradation method (EDM), to detect small cell lung cancer (SCLC) from computed tomography (CT) images. This research could facilitate early detection of lung cancers. The training data and testing data are high-resolution lung CT scans provided by the National Cancer Institute. We selected 12 lung CT scans from the library, 6 of which are for healthy lungs, and the remaining 6 are scans from patients with SCLC. We randomly take 5 scans from each group to train our model and used the remaining two scans to test. Our algorithms achieve an accuracy of 77.8%.

Keywords *- Image processing; machine learning; computed tomography; small cell lung cancer detection*

INTRODUCTION

According to recent surveys, cancer-related medical expenses and labour loss cost annually 10,000 billion dollars all- over the world [1]–[3]. Lung cancer is a number one killer among all cancer-leading deaths, due to late-stage detection and environmental conditions, such as air pollution, working conditions, life-long smoking habits [4], [5]. For instances, 225,000 new cases were detected in the United States in 2016, and 4.3 million new cases in China in 2015 [2], [6]. Like other types of cancers, early detection is viewed to be the best strategy to save lives.

Computerized Tomography (CT) is 3D imaging modality which has been widely used for lung cancer screening and diagnostics. 3D images are reconstructed from thousands of 2D X-ray transmission projections. Advanced 3D reconstructions [7]–[9] were developed for better image quality and diagnostic accuracy.

Most current machine learning based Computer Aided Diagnostic (CAD) researches are focusing on non-small cell lung cancer (NSCLC) [10]–[12]. CAD systems help to reduce the workload of radiologists significantly [10], [13]. So far there is very few works on small cell lung cancer (SCLC) detection, which is an extremely difficult task for the human observer because the image with SCLC looks almost identical to one without. There are machine learning algorithms that

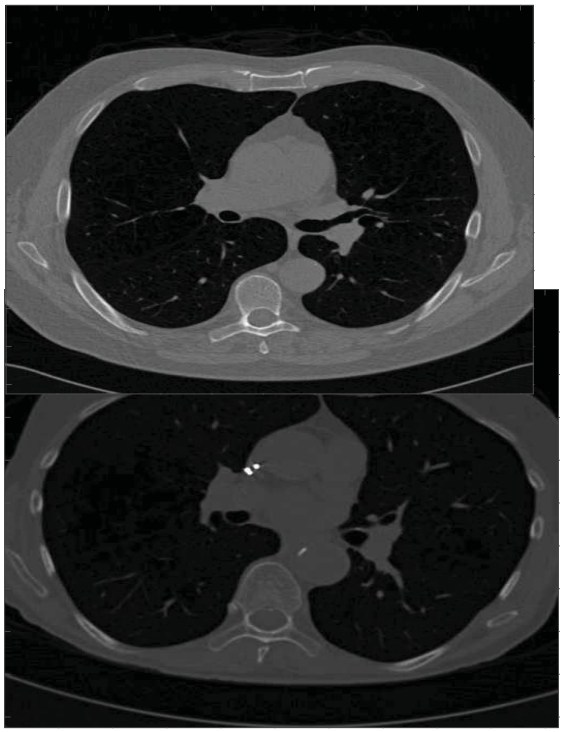


Figure 1. Two random CT lung images: (a) a healthy lung image from Group 0; (b) a cancer-detected lung image from Group 1.

may be candidates for SCLC detection task, such as convolution neural network based deep learning method [14], which starts with building neurons and layers, where a dynamic parameter set is used to calculate forward propagation. During the training process, parameters in each layer are updated by

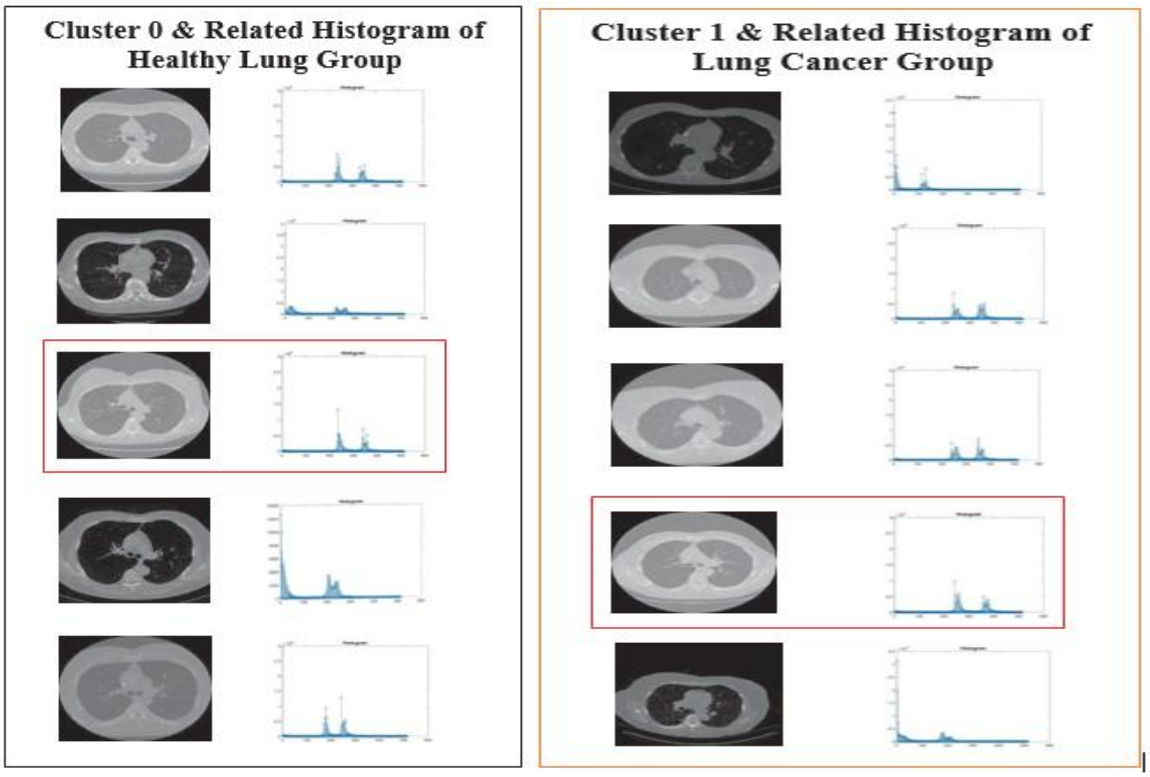


Figure 2. The histogram results for all training sets from Group 0 and Group 1; shown as the red box highlighted two images are the hardest to distinguish by human visual inspection.

back propagation from cost function (i.e. a distance metric between the forward propagation of input data and labels) [14], [15]. However deep learning algorithms usually require an extremely large training dataset, which is not always available we propose a novel neural-network based algorithm, which we refer to as entropy degradation method (EDM), to detect small cell lung cancer (SCLC) from computed tomography (CT) images. This research could facilitate early detection of lung cancers. The training data and testing data are high- resolution lung CT scans provided by the National Cancer Institute. We selected 12 lung CT scans from the library, 6 of which are for healthy lungs, and the remaining 6 are scans from patients with SCLC. We randomly take 5 scans from each group to train our model and used the remaining two scans to test. Our algorithms achieve an accuracy of 77.8%.

METHODS

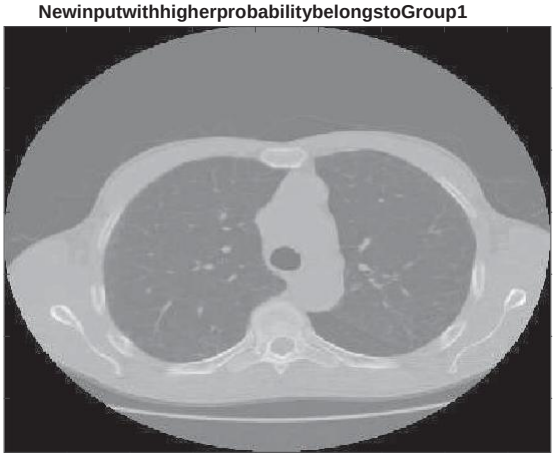
A.Subjects and test data descriptions

This study utilized a data set of CT images with high- resolution scans provided by the National Cancer Institute [16]. It covers hundreds of patient CT scans from various sources and qualities, with ground truth labels given by pathology diagnosis. Each CT images contains multiple axial slices of

the chest cavity, usually varied from 100 to 500 slices depending on scan parameters.
For training data, we chose the CT images of ten patients, five of which have been diagnosed with SCLC, and five others without. The images with SCLS are labelled as cluster 1, and the others are labelled as cluster 0. Because not all CT scans reveal cancer cells for patients with SCLC, we manually selected slices where the lung is present 1. For testing, we chose two additional CT images, one for a patient with SCLC, and one without

New input with higher probability belongs to Group 0

(a)



(b)

Figure 3. Lung Cancer detected using entropy degradation method (EDM) algorithm: (a) One new input is detected to Group 0 (i.e. health patient) (b) Another new input is detected to Group 1 (i.e. lung cancer detected).

B.Data Analysis Method

We treat SCLC detection as a binomial problem, which indicates that either the input image belongs to Group 0 of healthy lung patients or Group 1 of cancer developed or developing lung patients. From the training sets, 512 features in the histogram are extracted from the chosen slices which are shown as in Figure 2 as an example. One can barely tell the differences between two highlighted images from the two groups. There is no obvious nodule or symptoms presented in the anatomy of Group 1 image. In the training process, the vectorized histogram from five healthy lungs, and five cancer labelled lungs are fed into our algorithm. Details will be addressed in part IV.

EDM is designed with the concept of shallow neural network, which transforms the vectorized histogram of each training set into a score. The score is further transformed into probability through a logistic function, a cost function where the difference between the transformation and label is calculated and later be fedback by back-propagation stage [14], [15], [17]–[22]. It's called score-probability policy. In the forward process, a score is calculated from the vectorized histogram. The score is transformed into probability through a logistic function. In the back propagation process, a lost function is used to update the parameters. For the testing part, a new input without the label is fed into the well-trained network, where its probabilities associated with those scores are calculated as outputs. Results indicate which group the testing data belongs to (i.e. more likely to a lung cancer patient or to a healthy patient). In details, we defined ML QW() function, inside which we initialized the points from the outputs of function ICA − QW().

PERFORMANCE EVALUATION OF EDM

To evaluate the performance of the proposed EDM algorithm, we chose 12 datasets from the database (6 from healthy patients and 6 from SCLC patients). We then randomly selected 10 out of 12, where 5 from 6 healthy patients and 5 from 6 SCLC patients, as the training set and extract the vectorized histogram as training input (i.e. 512 x 10 as input). The remaining two samples are used as the testing set (i.e. 512 x 2 as input). These combinations give us a total of 36 tests. The result is shown in terms of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), accuracy (TP+TN)/(TP+FP+FN+TN).

Item	EDM
True Positive (TP)	30
True Negative (TN)	26
False Positive (FP)	10
False Negative (FN)	6
Accuracy	77.8%

RESULTS

Figure 3 shows one randomly selected testing set, where (a) is from Group 0 (healthy lung) and (b) is from Group 1 (small- cell lung cancer case), and EDM gives the right prediction. It is noticeable that none of the images presents visible nodule and there is no pathological difference observable from the human observer.

Table I presents the result of our experiment. As can be seen, our algorithms makes 10 false positive predictions (among

36 tests), meaning 10 healthy patients are labelled to have SCLC by mistake. Similarly, our algorithm misses 6 cases when the patients actually have SCLC. This shows that there is a large room for improvement. We plan to integrate EDM with Adaboost where EDM is treated as a weak classifier for better prediction.

CONCLUSIONS AND DISCUSSIONS

In this study, we proposed an EDM machine learning algorithm with vectorized histogram features to detect SCLC for early malicious cancer prediction. While we show that EDM has reasonably good prediction accuracy, there is a large room for improvement before our algorithm can be used in the clinical setting. The ultimate goal of this study is to develop a clinical decision-making system for radiologists to better predict a malicious lung cancer from SCLC with computed tomography (CT) imaging. For the future work, we would train the proposed method with the larger training set and deeper network and combine it with convolution neural network, which has been used in CT imaging for different applications [17], [23].

REFERENCES

[1]L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, andA. Jemal, "Global cancer statistics, 2012," CA: a cancer journal for clinicians, vol. 65, no. 2, pp. 87–108, 2015.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," CA: a cancer journal for clinicians, vol. 66, no. 1, pp. 7–30, 2016.

[3] I. Hwang, D. W. Shin, K. H. Kang, H. K. Yang, S. Y. Kim, and J.-H. Park, "Medical costs and health care utilization among cancer decedents in the last year of life in 2009," Cancer research and treatment: official journal of Korean Cancer Association, vol. 48, no. 1, p. 365, 2016.

[4] A. A. Farag, H. E. A. El Munim, J. H. Graham, and A. A. Farag, "A novel approach for lung nodules segmentation in chest ct using level sets," IEEE Transactions on Image Processing, vol. 22, no. 12, pp. 5202– 5213, 2013.

[6] W. Chen, R. Zheng, P. D. Baade, S. Zhang, H. Zeng, F. Bray, A. Jemal,

X. Q. Yu, and J. He, "Cancer statistics in China, 2015," CA: a cancer journal for clinicians, vol. 66,

no. 2, pp. 115–132, 2016.

[5] J. Lortet-Tieulent, I. Soerjomataram, J. Ferlay, M. Rutherford, E. Wei- derpass, and F. Bray, "International trends in lung cancer incidence by histological subtype: adenocarcinoma stabilizing in men but still increasing in women," Lung cancer, vol. 84, no. 1, pp. 13–22, 2014.

[7] J.-B. Thibault, K. D. Sauer, C. A. Bouman, and J. Hsieh, "A three- dimensional statistical approach to improved image quality for multislice helical ct," Medical Physics, vol. 34, no. 11, pp. 4526–4544, 2007.

[8] S. Xu, A. Uneri, A. J. Khanna, J. Siewerdsen, and J. Stayman, "Polian- energetic known-component ct reconstruction with unknown material compositions and unknown x-ray spectra," Physics in Medicine and Biology, vol. 62, no. 8, p. 3352, 2017.

[9] S. Xu, J. Lu, O. Zhou, and Y. Chen, "Statistical iterative reconstruction to improve image quality for digital breast tomosynthesis," Medical Physics, vol. 42, no. 9, pp. 5377–5390, 2015.

[10] K. Suzuki, "Pixel-based machine learning in computer-aided diagnosis of lung and colon cancer," in Machine Learning in Healthcare Informatics. Springer, 2014, pp. 81–112.

[11] M. N. Gurcan, B. Sahiner, N. Petrick, H.-P. Chan, E. A. Kazerooni,
P. N. Cascade, and L. Hadjiiski, "Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system," Medical Physics, vol. 29, no. 11, pp. 2552– 2558, 2002.

[12] M. C. Lee, L. Boroczky, K. Sungur-Stasik, A. D. Cann, A. C. Bor- czuk, S. M. Kawut, and C. A. Powell, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," Artificial intelligence in medicine, vol. 50, no. 1, pp. 43–53, 2010.

[13] H. Dang, J. Stayman, A. Sisniega, J. Xu, W. Zbijewski, J. Yorkston,
N. Aygun, V. Koliatsos, and J. Siewerdsen, "Cone-beam ct of traumatic brain injury using statistical reconstruction with a post-artefact-correction noise model," in Proceedings of SPIE–the International Society for Optical Engineering, vol. 9412. NIH Public Access, 2015.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for Matlab," in Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015, pp. 689–692.

[16] "Data Science Bowl 2017 data description," https://www.kaggle.com/c/ data-science-bowl-2017, accessed:2017-01-13.

[17] G. Wang, "A perspective on deep imaging," IEEE Access, vol. 4, pp. 8914–8924, 2016.

[18] X. Luo, Y. Xu, W. Wang, M. Yuan, X. Ban, Y. Zhu, and W. Zhao, "Towards enhancing stacked extreme learning machine with sparse autoencoder by correntropy," Journal of the Franklin Institute, 2017.

[19] X. Luo, J. Deng, J. Liu, W. Wang, X. Ban, and J.-H. Wang, "A quantized kernel least mean square scheme with entropy-guided learning for intelligent data analysis," China Communications, vol. 14, no. 7, pp. 1–10, 2017.

[20] X. Luo, J. Deng, W. Wang, J.-H. Wang, and W. Zhao, "A quantized kernel learning algorithm using a minimum kernel risk-sensitive loss criterion and bilateral gradient technique," Entropy, vol. 19, no. 7, p. 365, 2017.

[21] X. Luo, Y. Lv, M. Zhou, W. Wang, and W. Zhao, "A Laguerre neural network-based ADP learning scheme with its application to tracking control in the internet of things," Personal and Ubiquitous Computing, vol. 20, no. 3, pp. 361–372, 2016.

[22] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," Future Generation Computer Systems, vol. 61, pp. 85–96, 2016.

[23] S. Xu, P. Prinsen, J. Wiegert, and R. Manjeshwar, "Deep residual learning in ct physics: scatter correction for spectral ct," arXiv preprint
arXiv:1708.04151, 2017.