International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Survey on Lung Cancer Identification using Data Mining Techniques

[1]Mrs. V. Prasanna, [2]Dr. M. Thangamani
[1]Assistant Professor, KTVR Knowledge Park for Engineering and Technology, Coimbatore, vp.kpet@gmail.com
[2]Assistant Professor, Kongu Engineering College, Perundurai-638052, manithangamani2@gmail.com

Abstract: *According to statistics, lung cancer is the leading cause of cancer related deaths compared to any other type of cancer in the world. Lung cancer is contributing about 1.3 million deaths per year globally. Further, these reports indicate that the survival rate of lung cancer is only 14 percentages but still, if defective nodules are detected at an early stage, the survival rate can be increased up to 50 percentages. Thus the early detection of lung nodules is important in the treatment of lung cancer. These research papers contribute survey of Lung cancer identification in various aspects.*

Keywords: *Data Mining, Image mining, Image Retrieval, Image classification and Lung cancer detection*

1. Introduction

Lung cancer is a disease characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung in a process called metastasis into nearby tissue and eventually, into other parts of the body. Most cancers that start in lung, known as primary lung cancers, are carcinomas that derive from epithelial cells. The most common cause of lung cancer is long-term exposure to tobacco smoke, which causes 80-90% of lung cancers. Worldwide, lung cancer is the most common cause of cancer-related death in men and women, and is responsible for 1.38 million deaths annually, as of 2008. In lung cancer research, one of the most sensitive methods for detecting pulmonary nodules is Computed Tomography (CT), in which a nodule is defined as a rounded and irregular opaque figure on a CT scan, with a diameter up to 30mm. Each scan contains hundreds of images that must be evaluated by a radiologist, which is a difficult process. So for this reason, the use of a Computer-Aided Detection (CAD) system can provide an effective solution by assisting radiologists in increasing the scanning efficiency and potentially improving nodule detection.

2. Related work

Wook-Jin Choi et al. [1] proposed an automated pulmonary nodule detection system based on a genetic programming (GP)-based classifier. Thresholding and 3D-connected component labeling are used to segment the lung volume. Optimal multiple thresholding and rule-based pruning are applied to detect and segment nodule candidates. A GP-based classifier (GPC) is trained and used to classify nodules and non-nodules. This method reduces the number of false positives in the nodule candidates, achieving 94.1 % sensitivity at 5.45 false positives per scan. Shanhui Sun et al. [2] proposed a fully automated approach for segmentation of lungs with high-density pathologies. A novel robust active shape model (RASM) matching method is used to roughly segment the outline of the lungs. Then, an optimal surface finding approach is utilized to adapt the initial segmentation result to the lung. An evaluation on 30 data sets with 40 abnormal and 20 normal lungs resulted in an average Dice coefficient of 0.975$\pm$ 0.006 and a mean absolute surface distance error of 0.84$\pm$ 0.23 mm, respectively. This method shows better segmentation results compared to two commercially available lung segmentation approaches.

Tao Xu et al. [3] proposed an automatic lung field segmentation technique to address the inadequacy of ASM in lung field extraction. In this paper, an automatic global edge and region force field guided method with non-linear exponential point evolution for lung field segmentation, by introducing global edge and force field information together with a new point evolution technique is proposed. Experimental results demonstrated that the proposed method is time efficient and improves the accuracy, sensitivity, specificity and robustness of the segmentation results, compared to the typical ASM and hybrid LSSP. The experimental results using both normal and abnormal chest radiographs show that the proposed technique provides better performance and can achieve 3-6% improvement

on accuracy, sensitivity and specificity compared to traditional ASM techniques.

Temesguen Messay et al. [4] have presented a novel computer aided detection (CAD) system for the detection of pulmonary nodules in thoracic CT images. The proposed CAD system combines intensity thresholding with morphological processing to detect and segment nodule candidates simultaneously. A set of 245 features is computed for each segmented nodule candidate. A 7-fold cross-validation performance analysis using the LIDC database only shows CAD sensitivity of 82.66% with an average of 3 FPs per CT scan/case. Eva M.van Rikxoort et al. [5] proposed a method for automatic segmentation of pulmonary lobes from computed tomography (CT) scans is presented that is robust against incomplete fissures. The method is based on a multiatlas approach in which existing lobar segmentations are deformed to test scans in which the fissures, the lungs, and the bronchial tree have been automatically segmented. The method is evaluated on two test sets of 120 scans in total. The results show that the lobe segmentation closely follows the fissures when they are present. When the fissures are incomplete, an observer study shows agreement of the automatically determined lobe borders with a radiologist for 81% of the lobe borders on average.

Xujiong Ye et al. [6] investigated a new CT lung nodule Computer Aided Detection (CAD) method for detecting both solid nodules and Ground-Glass Opacity (GGO) nodules. The methodology uses Antigeometric diffusion, which diffuses across the image edges, is used as a pre-processing step; adaptive thresholding technique is used to segment the potential nodule objects. The proposed methodology uses a classification technique called Support Vector Machine (SVM), to reduce the number of false positive (FP) objects. The experimental results give an average detection rate of 90.2%, with approximately 8.2 FP/scan. The two difficulties are the detection of nodules that are adjacent to vessels and which are non spherical in shape. In this approach produce shape-based genetic algorithm template-matching (GATM) method for the detection of nodules with spherical elements. A fitness function is defined by combining a 3D geometric shape feature and global nodule intensity distribution. This method has been validated on a clinical dataset of 70 thoracic CT scans that contains 178 nodules. A total of 160 nodules were correctly detected by the proposed method and resulted in a detection rate of about 90%, with the number of false positives at approximately 14.6/scan.

Feature subset selection method based on genetic algorithms to improve the performance of false positive reduction in lung nodule computer-aided detection [7]. It is coupled with a classifier based on support vector machines. This approach find the automatically the optimal size of the feature set, and chooses the most relevant features from a feature pool. Its performance was tested using a lung nodule database acquired by multislice CT scans. Han, H. et. al [8] explored CAD system based on a hierarchical vector quantization scheme. Compared with the commonly-used simple thresholding approach, the high-level Vector quantization yields a more accurate segmentation of the lungs from the chest volume. In identifying initial nodule candidates (INCs) within the lungs, the low-level Vector quantization proves to be effective for INCs detection and segmentation.

Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. The multi layered method combining clustering and decision tree techniques to build a cancer risk prediction system is proposed [9] which predicts lung, breast, oral, cervix, stomach and blood cancers and is also user friendly, time and cost saving. This research uses data mining technology such as classification, clustering and prediction to identify potential cancer patients.

Learning method based on unsupervised learning which can be used in building a predictive model for early detection of lung cancer [10]. In this research, ANN can be used to predict the disease even with the occurrence of new symptoms. Then the disease can be further analyzed by extracting the resultant weight vector after the training process.

Monali Dey et al. [11] survey facilitate the most common data mining algorithms, implemented in modern Healthcare Decision Support Systems, and evaluate their performance on several medical datasets. Three algorithms were chosen: C4.5, Multilayer Perceptron and Naïve Bayes and different disease database are taken.

Medical images mining is a promising area of computational intelligence applied to automatically analyzing patient's records aiming at the discovery of new knowledge potentially useful for medical decision making. Zakaria suliman et. al [12] used data mining, techniques such as neural networks and association rule mining techniques, for detection and classification Lung Cancer in X-Ray chest films.

Support vector machine (SVM) is a popular method for classification, but there are few methods that utilize SVM for survival analysis in the literature because of the computational complexity. Zhenqiu Liu [13] proposed method can simultaneously identify survival-associated prognostic factors and predict survival outcomes.

Cloud computing is one of the major Information Technology (IT) trends that adopt IT maximum utility. It aids to analyze larger datasets for the hiding information. Existing methods may have a good performance, but it takes a lot of time to analyze microarray data. Ming-Tai et al. [14] investigate the

Genetic algorithm (GA)-Fuzzy-based voting mechanism combined with the Hadoop to find the critical genes that affect the symptom and increase the speed to voting mechanism adopted the Hadoop technique. Lung cancer is a cancer that starts in the lungs. Smoking is the biggest risk factor of lung cancer. The more years and larger number of cigarettes smoked the greater the risk of developing lung cancer. The average age of some one Diagnosed with lung cancer is 65 to 70 years old, but people who are younger can develop lung cancer. Young adults who have never smoked also can develop lung cancer [15].

3. Conclusion and Further Direction

This survey illustrates the Lung cancer detection through data mining approach. Further author may be investigating the clinical decision support system for mining Lung Cancer.

References

[1] Wook-Jin Choi, Tae-Sun Choi, 'Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images', Information Sciences 212 57–78, 2012.

[2] Shanhui Sun, Christian Bauer, and Reinhard Beichel, 'Automated 3-D Segmentation of Lungs With Lung Cancer in CT Data Using a Novel Robust Active Shape Model Approach', IEEE transactions on medical imaging, vol. 31, no. 2.

[3] Tao Xu, Mrindal Mandal, Richard Long, Irene Cheng and Anup Basu, 'An edge-region force guided active shape approach for automatic lung field detection in chest radiographs', Computerized Medical Imaging and Graphics, 2012.

[4] Temesguen Messay, Russell C. Hardie, Steven K. Rogers, 'A new computationally efficient CAD system for pulmonary nodule detection in CT imagery', Medical Image Analysis 14 390–406, 2010

[5] Eva M. van Rikxoort, Mathias Prokop, Bartjan de Hoop, Max A. Viergever, Josien P. W. Pluim, and Bram van Ginneken, 'Automatic Segmentation of Pulmonary Lobes Robust Against Incomplete Fissures', IEEE transactions on medical imaging, vol. 29, no. 6, 2010.

[6] Xujiong Ye, Xinyu Lin, Jamshid Dehmeshki, Greg Slabaugh, Gareth Beddoe 'Shape-Based Computer-Aided Detection of Lung Nodules in Thoracic CT Images' , IEEE transactions on biomedical engineering, vol. 56, 2009.

[7] Boroczky, L, Luyin Zhao, Lee, K.P, "Feature Subset Selection for Improving the Performance of False Positive Reduction in Lung Nodule CAD", IEEE Transaction on Information Technology in Biomedicine, Vol.10, issue 2, 2006.

[8] Han, H. Li, L. Han, F. ;Song, B. Moore, W. Liang, Z.Fast and Adaptive Detection of Pulmonary Nodules in Thoracic CT Images Using a Hierarchical Vector Quantization Scheme IEE Journal of Biomedical and Health Informatics, Vol. 9, Issue 2 , pp. 648 - 659 , 2014.

[9] P.Ramachandran, N.Girija, T.Bhuvaneswari, Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications, Vol. 97, Issue 13, pp. 7622-7626, 2014

[10] Juliet R Rajan1, Jefrin J Prakash2, Early Diagnosis of Lung Cancer using a Mining Tool, International Journal of Emerging Trends in Computer Science, Special issue, 2013

[11] Monali Dey, Siddharth Swarup Rautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies, Vol. 5 , issue 1 , pp. 470-477, 2014.

[12] Zakaria suliman zubi, Rema Asheibani Saad, Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer, Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, pp-32-37, 2014

[13] Zhenqiu Liu, Dechang Chen, Guoliang Tian, Man-Lai Tang, Ming Tan, and Li Sheng, Chapter 2, Efficient Support Vector Machine Method for Survival Prediction with SEER Data, University of Maryland at Baltimore, Advances in Computational Biology, pp-11-18, 2010.

[14] Ming-Tai Wu1 Jain-Shing Wu1 Chung-Nan Lee, Ming-Cheng Chen1, A Genetic Algorithm-Fuzzy-Based Voting Mechanism Combined with Hadoop Map-Reduce Technique for Microarray Data Classification, pp. 41-48, 2013

[15] Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, 2014

Author Profile

Mrs. V. Prasanna completed her Master's degree in Software Engineering from Anna University of Technology, Coimbatore, Tamil Nadu, India. Currently, she is working as Assistant Professor in the Department of Information Technology, KTVR Knowledge Park for Engineering and Technology, Tamil Nadu, India. She has nearly 7 years of academic experience and one year industry experience. She has presented papers in four National International conferences. Her research interests include Software Engineering, Data mining, Image retrieval, Big data and Database Management System.

Dr. M. Thangamani is nearly 20 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers data mining,

machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published 50 articles in International journals and presented over 67 papers in national and international conferences in above field. She has delivered more than 35 Guest Lectures in reputed engineering colleges on various topics. She has organized many self supporting and sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She is on the editorial board and reviewing committee of leading research journals, and on the program committee of top international data mining and soft computing conferences in various countries. She also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and abroad. She is currently working as Assistant Professor in Kongu Engineering College.