A Survey on Early detection of Lung Cancer by gene expression profiles using Data Mining Techniques

Mohammed Mekuriya Adem[1], Aniket Jagtap[2]

[1]Student, Department of Computer Science, SIT and [2]Professor, Department of Computer Science, SIT
[1]Symbiosis Institute of Technology, Pune, India and [2]Symbiosis Institute of Technology, Pune, India

Abstract

Lung Cancer is the most common causes of death for human beings all over the world. Early detection and treatment can increase the survival rate of lung cancer patients. This work proposes a methodology that detects lung cancer using gene expression profiles. The proposed system first extracts significant features from the input patterns by using Information Gain (IG). Then the Genetic Algorithm (GA) is applied for feature reduction. The proposed system is evaluated by considering microarray dataset and compared with the most recent systems. Support Vector Machine (SVM) is applied for classification.

Keywords *Lung Cancer. Genetic Algorithm. Information Gain. Support Vector Machine. Feature Selection*

1. Introduction

According to the up-to-date statistics from the American Society [3], lung cancer is the leading cause of cancer related deaths with over 159,000 deaths estimated for the United States alone in 2013, and overall 5-year survival rate for lung cancer is merely 16%. The survival rate increases to 52% if it is localized, and decreases to 4% if it has metastasized.

According to the World health organization (WHO), "cancer is considered among the leading cause of death over the world, with approximately 14 million cases and 8.2 million cancer related deaths every year" [14]. Cancer arises from a genetic mutation of normal cells. These mutations cause damage to the DNA and affect the life cycle of the cells, causing them to reproduce in an uncontrolled manner, and perhaps resulting in the formation of malignant tumors (cancers) [14]. The diagnosis of complicated genetic disease like cancer is normally based on tumor tissue, irrational characteristics and clinical stages [14].

2. Related Work

In [13], Predicting outcomes of non-small cell lung cancer using CT image features (2014). In this paper, the authors applied radiomics to select a 3D feature from CT images of the lung. The classifiers can be built to predict the survival period. So by comparing the classifiers and feature selection approaches such as support vector machine, Naïve Bayes, Decision Tree, the best accuracy obtained were 77.5% using Decision Tree in leave-one-out cross validation. This result was obtained after selecting 5 features per fold from 219.

In [14], Computer-Aided detection of pulmonary nodules based on SVM in Thoracic CT images (2015). The purpose of this study was to develop a computer aided system to detect pulmonary nodules on CT scan based SVM classifier for the diagnosis of pulmonary nodules. The first step of the developed system was to reduce the volume of the data using data mining techniques. Then, divided by the area of the chest to identify suspicious nodules and eventually nodules are detected and by comparing with the threshold-based methods, SVM classifies the area of the lung accurately. From 147 patients with lung LIPC images the database shows that, per scan 89.9% sensitivity and 3.9 false positives are obtained.

In [15], Lung cancer detection using Genetic Approach (2016). In this work preliminary diagnosis and detection of lung cancer from X-Ray, CT and PET images Genetic Algorithm were applied to optimize the results. Genetic Algorithm allows doctors to identify the presence of nodules in the lung at an early stage. Genetic Algorithm and Naïve Bayes were applied to classify different stages of the cancer images efficiently and accurately. The system results, 80% accuracy in classification.

In [1], Early diagnosis of breast cancer by gene expression profiles (2017). In this paper Intelligent Decision Support System (IDSS) was developed for diagnosis of breast cancer using gene expression profile datasets. The developed system first extracts significant features from the input

pattern by using Information Gain (IG). Then deep genetic algorithm was applied for feature reduction and breast cancer diagnosis. The experimental result shows that the system produces 100% classification accuracy using SVM.

In [16], Lung nodule classification using artificial crawlers, directional texture and vector support machine (2017). The authors developed a methodology that classifies nodules and non-nodules using texture features. Artificial crawlers and Rose diagrams are used for representing patterns over 3D images. Support vector machine was used for classification. At the first stage they used artificial crawlers and rose diagrams to extract directional measurements. Then the hybrid model that combines texture measurements from artificial crawlers and the rose diagram. The authors divided the database into training and testing sets. They used partition for training and testing of 20/80%, 40/60%, 60/40 and 80/20%. The division was repeated 5 times at random. Finally, they reached a mean accuracy of 94.30%, a mean sensitivity of 91.86% and a mean specificity of 94.78% a coefficient of accuracy variance of 1.61% and a mean area under the receiver operating characteristics of 0.922

In [17], Predicting brain metastases for non-small cell lung cancer based on magnetic resonance imaging (2017). The authors studied the relationship between brain structure and brain metastases occurrences using magnetic resonance images by segmenting into cerebro-spinal fluid, gray matter and white matter using voxel-based morphometry. The automatic anatomic labeling template was used to extract 116 brain regions from the gray matter volume. The system analyzed the elapsed time between magnetic resonance image acquisitions. Brain metastases diagnosed were analyzed by the list absolute shrinkage and by selecting an operator method. Permutation test and Leave-one-out cross validation (LOOCV) were used to validate the model. Twenty patients were used to develop the model. The remaining 69 was used for independent validation and verification of the developed model. As a result, accuracy, sensitivity, and specificity of the model for predicting brain metastases occurrence for the 69 independent validating patients were 70, 75 and 66%, respectively, in 6 months and 72, 82 and 60%, respectively for 1 year.

In [18], Applying high performance genetic data feature selection and classification algorithm for colon cancer diagnosis (2017). In this work a three phase approach was developed. The first and second phase examined the feature selection algorithm and classification algorithms separately. Phase three examined the performance of the combination of the above two phases. From phase one, it was found that the Particle Swarm Optimization (PSO) algorithm performed best with the colon dataset as a feature selection. From phase two, the Support Vector Machine (SVM) algorithm outperformed classification with an accuracy of 86%. In phase three, by combining the PSO and SVM algorithms 94% accuracy and good performance were observed.

Table 1. Comparison of Algorithms used for early detection and classification of cancer diseases

S. No	Title	Author(s)	Year/ Publisher	Techniques	Results
1	Predicting outcomes of non-small cell lung cancer using CT image features	Hawkins, S. H., Korecki, J. N., Balagurunathan, Y., Gu, Y., Kumar, V., Basu, S., ... &Gillies, R. J.	2014/IEEE	✓ support vector machine ✓ Naïve Bayes ✓ Decision Tree	✓ 77.5% accuracy using Decision Tree
2	Computer-Aided detection of pulmonary nodules based on SVM in Thoracic CT images	Eskandarian, P., & Bagherzadeh, J.	2015/IEEE	✓ SVM	✓ 89.9% sensitivity and 3.9 false positives per scan
3	Lung cancer detection using Genetic Approach	Kurkure, M., &Thakare, A.	2016/IEEE	✓ Genetic Algorithm ✓ Naïve Bayes	✓ 80% accuracy
4	Early diagnosis of breast cancer by gene expression profiles	Salem, H., Attiya, G., & El-Fishawy, N.	2017/Springer	✓ Information Gain ✓ Genetic Algorithm ✓ Support Vector Machine	✓ 100% classification accuracy
5	Lung nodule classification using artificial crawlers, directional texture and vector support machine	Froz, B. R., de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M.	2017/Elsevier	✓ Artificial crawlers and Rose diagrams	✓ a mean accuracy of 94.30%, ✓ a mean sensitivity of 91.86% and ✓ a mean specificity of 94.78% ✓ coefficient of accuracy variance of 1.61% and ✓ a mean area under the

					receiver operating characteristics of 0.922
6	Predicting brain metastases for non-small cell lung cancer based on magnetic resonance imaging	Yin, G., Li, C., Chen, H., Luo, Y., Orlandini, L. C., Wang, P., & Lang, J.	2017/Springer	✓ Leave One out Cross Validation(LOOCV)	✓ Accuracy=70% ✓ Sensitivity=75% ✓ Specificity=66for 69 patients for 6 months and ✓ Accuracy =72%, ✓ Sensitivity=82%and ✓ Specificity=60%, respectively for 1 year.
7	Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis	Al-Rajab, M., Lu, J., & Xu, Q.	2017/Elsevier	✓ Particle Swarm Optimization (PSO) algorithm ✓ Support vector machine	✓ Support vector machine =86%. ✓ PSO & SVM=94% accuracy

Conclusions

In this paper, a Genetic Algorithm (GA) and Support Vector Machine (SVM) based classifier is proposed to detect lung cancer from microarray data.Compared to other machine learning classifiers, the proposed classifier can achieve higher classification accuracy and decreases medical errors by minimizing life-threatening events caused by uninformed/ delayed medical decisions. As the result from the review papers Support Vector Machine and Genetic Algorithm has the highest classification accuracy.

References

[1] Salem, H., Attiya, G., & El-Fishawy, N. (2017). Early diagnosis of breast cancer by gene expression profiles. *Pattern Analysis and Applications*, *20*(2), 567-578.

[2] Aličković, E., &Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*, *28*(4), 753-763.

[3] Nguyen, T., Nahavandi, S., Creighton, D., &Khosravi, A. (2015). Mass spectrometry cancer data classification using wavelets and genetic algorithm. *FEBS letters*, *589*(24), 3879-3886.

[4] Kashyap, A., Gunjan, V. K., Kumar, A., Shaik, F., & Rao, A. A. (2016). Computational and Clinical Approach in Lung Cancer Detection and Analysis. *Procedia Computer Science*, *89*, 528-533.

[5] Wu, W. J., Lin, S. W., & Moon, W. K. (2012). Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Computerized Medical Imaging and Graphics*, *36* (8), 627-633.

[6] Fernandez-Cuesta, L., Perdomo, S., Avogbe, P. H., Leblay, N., Delhomme, T. M., Gaborieau, V., &Mukeria, A. (2016). Identification of circulating tumor DNA for the early detection of small-cell lung cancer. *EBioMedicine*, *10*, 117-123.

[7]Sudheesh, R. K., Rajan, J., Veena, V. S., &Sujathan, K. (2016, September). Study of malignancy associated changes in sputum images as an indicator of lung cancer. In *Technology Symposium (TechSym), 2016 IEEE Students'* (pp. 102-105). IEEE.

[8]Yin, Y., Sedlaczek, O., Muller, B., Warth, A., Gonzalez-Vallinas, M., Grabe, N., ...&Drasdo, D. (2017). Tumor cell load and heterogeneity estimation from diffusion-weighted MRI calibrated with histological data: an example from lung cancer. *IEEE Transactions on Medical Imaging*.

[9] Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A., & Huang, J. (2016, April). Lung cancer survival prediction from pathological images and genetic data—An integration study. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 1173-1176). IEEE.

[10] Sun, B., Yue, S., Hao, Z., Cui, Z., Wang, H., & Zhang, W. (2017, May). Early lung cancer identification based on ERT measurements. In *Instrumentation and Measurement Technology Conference (I2MTC), 2017 IEEE International* (pp. 1-5). IEEE.

[11]Deshmukh, S., &Shinde, S. (2016, September). Diagnosis of Lung Cancer using Pruned Fuzzy Min-Max Neural Network. In *Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on* (pp. 398-402). IEEE.

[12] Chauhan, D., & Jaiswal, V. (2016, October). An efficient data mining classification approach for detecting lung cancer disease. In *Communication and Electronics Systems (ICCES), International Conference on* (pp. 1-8). IEEE.

[13] Hawkins, S. H., Korecki, J. N., Balagurunathan, Y., Gu, Y., Kumar, V., Basu, S., ... &Gillies, R. J. (2014). Predicting outcomes of nonsmall cell lung cancer using CT image features. *IEEE Access*, *2*, 1418-1426.

[14] Eskandarian, P., & Bagherzadeh, J. (2015, May). Computer-aided detection of Pulmonary Nodules based on SVM in thoracic CT images. In *Information and Knowledge Technology (IKT), 2015 7th Conference on* (pp. 1-6). IEEE.

[15] Kurkure, M., &Thakare, A. (2016, August). Lung cancer detection using Genetic approach. In *Computing, Communication Control and automation (ICCUBEA), 2016 International Conference on* (pp. 1-5). IEEE.

[16] Froz, B. R., de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M. (2017). Lung nodule classification using artificial crawlers, directional texture and support vector machine. *Expert Systems with Applications*, *69*, 176-188.

[17] Yin, G., Li, C., Chen, H., Luo, Y., Orlandini, L. C., Wang, P., & Lang, J. (2017). Predicting brain metastases for non-small cell lung cancer based on magnetic resonance imaging. *Clinical & experimental metastasis*, *34*(2), 115-124.

[18] Al-Rajab, M., Lu, J., & Xu, Q. (2017). Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Computer Methods and Programs in Biomedicine*, *146*, 11-24.

[19] Saini, A. K., Bhadauria, H. S., & Singh, A. (2016, February). A Survey of Noise Removal Methodologies for Lung Cancer Diagnosis. In *Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on* (pp. 673-678). IEEE.

[20] Kureshi, N., Abidi, S. S. R., &Blouin, C. (2016). A predictive model for personalized therapeutic interventions in non-small cell lung cancer. *IEEE journal of biomedical and health informatics*, *20* (1), 424-431.

[21]Pengo, T., Muñoz-Barrutía, A., & Ortiz-de-Solorzano, C. (2014). A Novel Automated Microscopy Platform for Multiresolution Multispectral Early Detection of Lung Cancer Cells in Bronchoalveolar Lavage Samples. *IEEE Systems Journal*, *8* (3), 985-994.

[22] Melissa, C.S.:'Lungcancer'(Medicine.net, 2011)

[23] Azzawi, H., Hou, J., Xiang, Y., &Alanni, R. (2016). Lung cancer prediction from microarray data by gene expression programming. *IET systems biology*, *10*(5), 168-178.