Artificial Neural Network Analysis of Volatile Organic Compounds for the detection of lung cancer

John B Butcher1\*[0000-0002-2121-3000], Abigail V Rutter2[0000-0003-2990-1163], Adam J Wootton1,3[0000-0001-6432-0503], Charles R Day1[0000-0001-9707-9189] and Josep Sulé-Suso2,4[0000-0003-0598-737X]

1 School of Computing and Mathematics, Keele University, Staffordshire, ST5 5BG UK

2 Institute for Science & Technology in Medicine, Keele University, Guy Hilton Research Centre, Staffordshire, ST4 7QB UK

3 Foundation year centre, Keele University, Staffordshire, ST5 5BG UK

4 Oncology Department, Royal Stoke University Hospital, University Hospitals of North Midlands, Newcastle Road, Staffordshire ST4 6QG UK  
j.b.butcher@keele.ac.uk

**Abstract.** Lung cancer is a widespread disease and it is well understood that systematic, non-invasive and early detection of this progressive and life-threatening disorder is of vital importance for patient outcomes. In this work we present a convergence of familiar and less familiar artificial neural network techniques to help address this task. Our preliminary results demonstrate that improved, automated, early diagnosis of lung cancer based on the classification of volatile organic compounds detected in the exhaled gases of patients seems possible. Under strictly controlled conditions, using Selected Ion Flow Tube Mass Spectrometry (SIFT-MS), the naturally occurring concentrations of a range of volatile organic compounds in the exhaled gases of 20 lung cancer patients and 20 healthy individuals provided the dataset that has been analysed. We investigated the performance of several artificial neural network architectures, each with complementary pattern recognition properties, from the domains of supervised, unsupervised and recurrent neural networks. The neural networks were trained on a subset of the data, with their performance evaluated using unseen test data and classification accuracies ranging from 56% to 74% were obtained. In addition, there is promise that the topological ordering properties of the unsupervised networks’ clusters will be able to provide further diagnostic insights, for example into patients who may have been heavy smokers but so far have not presented with any lung cancer. With the collection of data from a larger number of subjects across a long time period there is promise that an automated assistive tool in the diagnosis of lung cancer via breath analysis could soon be possible.

**Keywords:** Lung cancer diagnosis, Volatile Organic Compounds, SIFT, Artificial Neural Network analysis

1. Introduction

Lung cancer is associated with a poor prognosis and survival rates of less than 20% at 5 years [1]. One reason for the poor prognosis is that more than 50% of patients have advanced disease at the time of diagnosis [2]. Therefore, urgent work is needed to be able to quickly and cheaply diagnose this disease at an early stage. Importantly any new diagnostic tools should ideally be non-invasive.

There has been some interest over the last few years in the study of trace volatile organic compounds (VOCs) in exhaled breath. Initially, studies were carried out using gas chromatography. However, this is a laborious procedure and cannot easily provide information on absolute concentrations of these compounds. Over the last few years, new techniques have been developed to study trace gases in breath. These include the Selected Ion Flow Tube Mass Spectrometry (SIFT-MS), Proton Transfer Reaction Time-of-Flight Mass Spectrometry, Laser Spectroscopy and Quantum Cascade Laser-based gas sensors amongst others [3]. However, while several studies have linked some VOCs present in exhaled gases of patients with lung cancer (e.g. [4]), this methodology has not yet been applied in clinical practice.

Artificial neural networks are one of the most widely used computational intelligence approaches when tasked with the analysis and investigation of complex and noisy data with the added complication of potential non-linear interactions between subsets of features within the dataset. In this work, neural networks from three broad domains of machine learning have been deployed: the supervised backpropagation multilayer perceptron (MLP); Kohonen’s unsupervised self-organising map (SOM) to carry out some autonomous clustering of the data and expose interrelationships that might be present in the SIFT-MS data; and from the domain of recurrent neural networks, more usually applied to time-series data processing, a ‘clamped’ variant of the echo state network (clamped-ESN) architecture.

The rest of the paper is organized as follows: section 2 outlines the methodology used to collect the data from the patients and control group, in addition to the data preprocessing approaches and artificial neural network architectures, including their configurations, that were used; section 3 presents the results of applying the ANNs to this dataset; section 4 discusses these results in further detail; and finally, section 5 concludes the paper and outlines future work to be conducted.

1. Methods

Patients with the histological diagnosis of lung cancer treated at the Oncology Department, Royal Stoke University Hospital were included in this study. Local ethical and R&D approval was obtained. Twenty patients with the histological diagnosis of lung cancer and 20 control cases with either basal cell carcinoma (BCC) or squamous cell carcinoma (SCC) of the skin, but with no diagnosis of lung cancer, attending the Oncology department for treatment were also included. The two types of skin tumors don't usually spread and tend to remain localized to the skin. Furthermore, they also present in an elderly population, as is the case for lung cancer, making them a good case control for this study. Data on smoking patterns, alcohol consumption, and the presence of chest infection were also recorded. In order to avoid contamination by other metabolites in their breath, both patients and control subjects with diabetes, renal failure or other metabolic diseases were excluded from the study. Patients (prior to starting treatment) and the control subjects all gave a morning breath sample. They did not eat food or drink alcohol nor smoke for the previous 12 hours and were not allowed to clean their teeth. They rinsed their mouth with water prior to giving a breath sample. They were asked to take a deep inhalation and exhale into a specially designed bag. As soon as the patient had finished the exhalation, the bag was sealed and taken for analysis to the laboratory facilities at the Institute of Science and Technology in Medicine, Guy Hilton Research Centre. The bags were kept at 37 °C and the breath gas from the bag was passed directly into the SIFT-MS instrument. Using this method, partial pressures of the trace gases down to about 10 parts per billion can be measured. Furthermore, the time response of the instrument is 20 ms, allowing the time profiles of the trace gas concentration in the breath to be obtained during a normal breathing cycle. (see [5] for more details of the SIFT-MS). However, breath can contain 850 different VOCs, with the precise significance and origin of many of them only poorly understood [6]. Thus, not only is more knowledge needed on the chemical pathways followed by VOCs between being released and finally expelled through breath [7], but also, improved data analysis methods and tools such as artificial neural networks (ANNs) to identify important biomarker(s) in this myriad of VOCs.

The analytical mass spectrometer in the SIFT-MS instrument is scanned over a pre-determined mass range of ions (10 to 180 m/z) to determine which breath metabolites are present and at what concentration. These data can readily be obtained within 1 minute.

Prior to presenting any data to the neural networks, 15 VOCs were selected from the SIFT-MS data as potential indicators of lung cancer. In no particular order these were acetone, acetaldehyde, ethanol, pentanol, hexanol, butyric acid, pentene, putresciene, terpenes, xylene, propanol, acetic acid, benzene, toluene and butanol. The data was then normalized between the range -1 and +1. The relatively modest number of subjects, just forty in total, recorded in the dataset meant that measures to try to avoid overfitting any very small training datasets had to be taken. Accordingly, the available dataset was subjected to a randomized, five-fold cross-validation decomposition that yielded five pairs of training and testing datasets. Each fold delivered 32 subjects and 8 subjects to be used as disjoint sets of training and testing data, respectively.

Artificial neural networks are one of the most widely used computational intelligence approaches for the analysis of complex and noisy data. One of the most established supervised neural network architectures is the multi-layer perceptron (MLP) using the backpropagation of error learning algorithm [8]. The MLP’s supervised training process involves evaluating a network’s performance by comparing its actual output to the target output associated with an input pattern in order to calculate the error of the network. The error is then backpropagated through the network to change the weights on each connection in order to progressively reduce the network’s overall error rate. MLPs have been successfully used in many studies across a wide range of domains for static pattern recognition, hence were considered to be a suitable architecture in this study. Examples in the medical domain include the detection of breast cancer [9], lung and oral cancer [10], early stage lung cancer using chest computed tomography images [11]. More recently, Adetiba et el. [12] demonstrated that an MLP could achieve 96% accuracy, outperforming a support vector machine and a Naïve Bayes classifier, when classifying VOCs from the Catalogue of Somatic Mutations in Cancer (COSMIC) database which contains lung cancer biomarker genes.

For this study, MLPs with 15 input units and a single output unit were used with high valued activations (i.e. activations > 0.5) at the output unit interpreted as cancer patients and low valued activations (i.e. activations < 0.5) at the output unit assigned to the class of non-cancer patients. The very small SIFT dataset size and the attendant risk of overfitting due to MLP over-parameterization meant that only modestly sized MLP hidden layers were ranged over (i.e. between 2 to 10 hidden units) in order to find the best performing hidden layer size: which was found to be around 6 hidden units. The MLPs were all trained with backpropagation learning rate and momentum values of 0.2 and 0.5, respectively. To avoid overfitting during training MLP performance on the training and the test data was assessed at 200 epoch intervals and training regimes lasting around 4,000 epochs were found to work best overall.

Echo state networks (ESNs) are a recurrent neural network (RNN) and have traditionally been applied to time-series data processing [13]. ESNs are part of the reservoir computing (RC) family of RNNs and rather than modifying all weighted connections in the network, only the ESN’s reservoir to output layer weights are modified during the training process. This reduces the complexity of training, making RC approaches well suited to time-series data analysis where excellent performance has been obtained [14-17]. However, *clamped-ESNs* have also recently been shown to be well-suited to static pattern recognition tasks (i.e. involving non-time-series data, e.g. [18-19]). For a clamped-ESN approach, each input pattern is clamped to the input units for a succession of presentations until the output node(s) settle to a stable state. Unlike conventional static pattern recognizers (such as MLPs that produce exactly the same response when presented with a static input pattern regardless of how many times the pattern is presented and regardless of what might have been presented in the recent past) the recurrent clamped-ESN, after its initial response, follows a trajectory towards an attractor in the high dimensional reservoir space that might not have been reached by a single static pattern presentation. This clamped-recurrent processing allows the weights connecting the ESN’s reservoir units to the output units to better classify an input pattern [18-19].

The ESNs used in this work consisted of 15 input units, 55 reservoir units and one output unit. The reservoir neurons had a 90% connectivity factor, spectral radius of 0.181, leak rate of 0.3820 and a *tanh* activation function, while the input scaling was set to 1. Each data sample was presented to the network 24 times before the output activation was harvested. The network was trained to give an output of +1 when presented with data from a patient with cancer, and -1 when presented with data from a patient without cancer. 20 different ESNs were trained on each set of training samples and tested on the corresponding set of testing samples, giving a total of 100 different trained ESNs. This was done to account for the variability of an ESN's weights that are generated randomly at initialization. After the testing data were presented, a final output value greater than zero was taken as a positive (i.e. a patient with cancer), while any output value less than zero was taken as a negative (i.e. a patient without cancer).

1. Classification Results

The lung cancer patients and controls both consisted of 12 males and 8 females. Regarding age, the control cases included a more elderly population. 55% of patients presented with metastatic disease (stage IV) while the remaining patients presented with locally advanced disease. Finally, there were more smokers and ex-smokers in the lung cancer patients group.

The performance of each ANN architecture was evaluated over each of the 5 folds used for cross validation. In addition, an overall average performance across all of the folds was determined and all of the results are shown in Table 1, below.

**Table 1.** The test performance of each architecture for each fold of cross validation and the overall average performance (standard deviation shown in brackets)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Cross validation performance (% correct) for each fold | | | | | Overall(% correct) |
| Architecture | 1 | 2 | 3 | 4 | 5 |  |
| MLP | 75  (0) | 88  (0) | 75  (0) | 69 (6.9282)  () | 63 (14.4337) | 73.9 (3.202) |
| Clamped-ESN | 40 (13.204) | 73.125 (4.5793) | 51.25 (10.653) | 50.625 (2.7951) | 65 (6.5394) | 56  (14.3768) |

As shown in Table 1, the MLP offers the best performance with 73.9%. Table 2 shows sensitivity analysis of the performance of the two architectures detailing the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) averaged across all 5 folds. The Matthews correlation coefficient is also shown where a score of +1 indicates perfect classification, a score of 0 random classification, and -1 a complete disagreement between the ANN output and the ground truth.

**Table 2.** Sensitivity analysis of the two architectures showing the true positives (TP), false positives (FP), true negatives (TN), false negatives (FN) scores, as well as their Mathews correlation coefficient (MCC) (standard deviation shown in brackets).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Architecture | TP | FP | TN | FN | MCC |
| MLP | 3.05 (1.8235) | 1.15 (0.8587) | 2.85 (1.4534) | 0.95 (1.3038) | 0.6397 |
| Clamped-ESN | 2.22 (0.2727) | 1.74 (0.4491) | 2.26 (0.4491) | 1.78 (0.2727) | 0.2018 |

Inspection of Table 2 shows that the MLP has more true positives and true negatives, and low false positives and false negatives when compared to the clamped-ESN. A higher MCC score further corroborates the superior performance of the MLP.

1. Discussion

One well known risk factor associated with lung cancer is smoking. In this study, 4 patients were still active smokers over period when the breath sample was taken and 16 were ex-smokers (stopped smoking at least 1 month before obtaining a breath sample). For the 20 control cases, 10 subjects were ex-smokers and the remaining 10 had not smoked before.

Early investigations using the unsupervised clustering properties of SOM neural networks suggest that one of the source of error in the, already very good classification scores, is the extent to which a non-lung-cancer patient has previously smoked. An important VOC constituent of cigarette smoke [20] and possible marker for active/recent smoking is acetonitrile [21] and so far, this has not been used as one of the inputs to the classification process reported above.

Even if the inclusion of additional VOCs such as acetonitrile do nothing to improve the overall classification accuracies, the two-dimensional topological ordering of clusters of patients on the trained SOM’s detector grid provides an excellent two-dimensional data visualization tool that will be useful in a clinical setting. This applies in particular when difficult decisions need to be made regarding high-dimensional breath-samples that might be at the margins of the classifications otherwise determined by the MLPs and clamped-ESNs, making the SOM a well-suited architecture when a breath sample falls between two or more classes.

The inferior performance provided by the clamped-ESN is surprising given previous performance on other static datasets. Further investigations are required to ascertain why this is the case, although the small size of the dataset analysed in this study could be a factor.

1. Conclusion

The early detection of lung cancer is of vital importance in the treatment and prognosis of the disease. Here we have presented preliminary results detailing the automated analysis of VOCs found in exhaled breath samples using several artificial neural networks. Despite the small size of the dataset (n=40), MLPs and clamped-ESNs were able to offer very good test classification performance: the clamped-ESN achieving an overall test average of 96%. Future work will involve further analysis of the characteristics of the whole VOC dataset and the collection of more SIFT data from a larger, longitudinal study, with the overall aim of creating an automated tool to assist in the early detection of lung cancer. The use of other architectures that have shown promise in this and wider domains (e.g. deep belief networks [22]) could also be the focus of future work.

Smoking is also a major risk factor for most chronic lung and pleural diseases. Confirming a patient’s smoking status has many treatment implications, for example, in cases where a potential lung transplant is required it is prudent that patients should have stopped smoking for at least 6 months. Smoking status is presently assessed by measuring levels of cotinine, the main breakdown product of nicotine, in the patient’s urine. Therefore, SIFT-MS in combination with the supervised and unsupervised neural network techniques used here hold the prospect of cheaper, on-line, identification of patient compliance and smoking abstinence regimes as well being able to make accurate lung cancer determinations.

### Acknowledgments

We gratefully acknowledge funding from the Slater & Gordon Health Projects & Research Fund/14/15 Round 1/A34489.

References

1. Lewis, D.R., Chen, H.S., Feurer, E.J., et al., SEER Cancer Statistics Review, 1975-2008. Bethesda, MD National Cancer Institute (2010).
2. Comella, P., Frasci, G., Panza, N. *et al*., Randomised trial comparing cisplatin, gemcitabine and vinorelbine with either cisplatin and gemcitabine or cisplatin and vinorelbine in advanced non small cell lung cancer: interim analysis of a Phase III trial of the Southern Italy Cooperative Oncology Group. J. Clin Oncol. 18: 1451-1457 (2000).
3. Amann, A. and Smith, D., eds. Breath analysis for clinical diagnosis and therapeutic monitoring. World Scientific Publishing Co. Singapore, (2005).
4. Buszewski, B., Ulanowska, A., Kowalkowski, T., Cieliski, K., Investigation of lung cancer biomarkers by hyphenated separation techniques and chemometrics. Clin. Chem. Lab. Med., 50, 573-581 (2012).
5. Rutter, A. V., Chippendale, T. W. E., Yang, Y., Španĕl, P., Smith, D. and Sulé-Suso, J., Quantification by SIFT-MS of acetaldehyde released by lung cells in a 3D model, Analyst 138, 91-95 (2013).
6. Amann, A., de Lacy Costello, B., Miekisch, W., Schubert, J., Buszewski, B., Pleil, J., Risby, T., The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. Journal of breath research, 8(3), 034001 (2014).
7. Hakim, M., Broza, Y.Y., Barash, O., et al. Volatile Organic Compounds of Lung Cancer and Possible Biochemical Pathways. Chem Reviews 112, 5949-5966 (2012).
8. Rumelhart, D.E., Hinton, G.E. and Williams, R.J., Learning internal representations by error propagation Parallel Distributed Processing, Explorations in the Microstructure of Cognition, 1, (1986).
9. Delen, D., Walker, G. and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127 (2005).
10. Choudhury, T., Kumar, V., Nigam, D. and Mandal, B., Intelligent Classification of Lung & Oral Cancer through Diverse Data Mining Algorithms. In *Micro-Electronics and Telecommunication Engineering (ICMETE), 2016 International Conference on* (pp. 133-138). IEEE (2016).
11. Emaminejad, N., Qian, W., Guan, Y., Tan, M., Qiu, Y., Liu, H. and Zheng, B., Fusion of quantitative image and genomic biomarkers to improve prognosis assessment of early stage lung cancer patients. *IEEE Transactions on Biomedical Engineering*, *63*(5), 1034-1043 (2016).
12. Adetiba E., Adebiyi M.O., Thakur S., Breathogenomics: A Computational Architecture for Screening, Early Diagnosis and Genotyping of Lung Cancer. In: Rojas I., Ortuño F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2017. Lecture Notes in Computer Science, vol 10209. Springer, Cham (2017).
13. Jaeger, H., The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, *148*(34), 13 (2001).
14. Butcher, J. B., Day, C. R., Austin, J. C., Haycock, P. W., Verstraeten, D., & Schrauwen, B. Defect detection in reinforced concrete using random neural architectures. Computer‐Aided Civil and Infrastructure Engineering, 29(3), 191-207 (2013).
15. Verstraeten, D., Schrauwen, B., & Stroobandt, D., Reservoir-based techniques for speech recognition. In *IEEE International Joint Conference on* *Neural Networks (IJCNN'06)*, 1050-1053 (2006).
16. Butcher, J.B., Verstraeten, D., Schrauwen, B., Day, C.R. and Haycock, P.W., Reservoir computing and extreme learning machines for non-linear time-series data analysis. Neural networks, 38, 76-89 (2013).
17. Scardapane, S., Butcher, J.B., Bianchi, F.M. and Malik, Z.K., Advances in Biologically Inspired Reservoir Computing. Cognitive Computation, 9(3), 295-296 (2017).
18. Wootton, A.J., Taylor, S.L., Day, C.R. and Haycock, P.W., Optimizing Echo State Networks for Static Pattern Recognition, Cognitive Computation, 9(3), 391-399 (2017).
19. Emmerich, C., Reinhart, R.F. and Steil, J.J. Recurrence Enhances the Spatial Encoding of Static Inputs in Reservoir Networks, in Artificial Neural Networks - ICANN 2010, vol. 6353, K. Diamantaras, W. Duch, and L. Iliadis, Eds. Springer Berlin Heidelberg, 148–153 (2010).
20. Campbell J. K., Rhoades J. W., and Gross A. l., Acetonitrile as a Constituent of Cigarette Smoke. Nature 198, 991-992 (1963)
21. Kushch, I., et al. Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using PTR-MS. Journal of breath research, 2(2), 026002 (2008).
22. Hinton, G. E, Osindero, S., and Teh, Y. W., A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527-1554, (2006).