Journal of Medical Systems (2019) 43: 211
https://doi.org/10.1007/s10916-019-1355-9

PATIENT FACING SYSTEMS

Multi-Class Neural Networks to Predict Lung Cancer

Juliet Rani Rajan [1] · A. Chilambu Chelvan [2] · J. Shiny Duela [3]

Received: 25 March 2019 / Accepted: 20 May 2019 / Published online: 31 May 2019

Abstract
Lung Cancer is the leading cause of death among all the cancers' in today's world. The survival rate of the patients is 85% if the cancer can be diagnosed during Stage 1. Mining of the patient records can help in diagnosing cancer during Stage 1. Using a multi-class neural networks helps to identify the disease during its stage 1 itself. The implementation of multi-class neural networks has yielded an accuracy of 100%. The model created using the neural networks approach helps to identify lung cancer during Stage 1 itself, thus the survival rate of the patients can be increased. This model can serve as pre-diagnosis tool for the practitioners.

Keywords Data mining · Multi-class neural networks

Introduction

Cancer is a broad term used to refer to a condition where the body's cells begin to grow and reproduce in a disorderly way to form lumps or masses of tissues called tumors. These cells can then overrun and devastate healthy tissue and also the organs. Cancer begins in one part of the body before spreading to other parts. There are around 100 different kinds of cancer [1]. Tumors nurture and affect with the all the systems of the body such as the nervous, circulatory and digestive, and discharge hormones that modify functions of the body. Normal cells in the body usually have a systematic path of growth, detachment, and death. When the apoptosis process breaks down, cancer initiates to form. Cancer cells do not have foreseen death but endures to mature and replicate which leads to a huge volume of atypical cells that propagates out of control. Damages in DNA mutations can outcome uncontrolled growth and therefore, impairment to the genes are intricate in cell division. The gene types that are accountable for the cell division course are the tumor suppressor genes, oncogenes, DNA-repair genes and suicide genes. Cancer transpires when the cell gene mutation make the cell incapable to correct the DNA damage. A carcinogen is any substance that is an agent directly act as a source of cancer cause. Carcinogens do not cause cancer in every case and all the time. When human bodies are subjected to carcinogens, free radicals are made that tries to snip electrons from further molecules in the body. These free radicals harm cells thus distressing their capability to serve as normal. There are also likelihoods to have positive genetic mutations during birth or an imperfection in gene that could likely develop cancer during the later stages in life. The amount of probable cancer-causing mutations in our DNA surges as the age increases. This marks age an imperative risk factor for cancer.

Initial diagnoses of cancer can significantly improve the probabilities of successful treatment and survival. Medical practitioners use evidence from signs and several other procedures to spot cancer. Certain imaging procedures such as CT scans, X-rays are used commonly in order to discover where a tumor is sited and what organs could conceivably been affected. Medical practitioners sometimes make use of a lean tube with a light and camera at the one end of the tube, to look out

This article is part of the Topical Collection on *Patient Facing Systems*

✉ Juliet Rani Rajan
julietrajan@gmail.com

A. Chilambu Chelvan
chill97@gmail.com

J. Shiny Duela
shinyduela@jerusalemengg.ac.in

[1] Sathyabama Institute of Science and Technology, Chennai, India

[2] Department of Electronics and Instrumentation, RMD Engineering College, Chennai, India

[3] Department of Computer Science and Engineering, Jerusalem College of Engineering, Chennai, India

for deformities inside the body. The unconditional way to diagnose cancer is removing cancer cells and observing them under a microscope which is refer as biopsy. Further types of molecular diagnostic tests are often engaged as well. Practitioners will analyze the carbohydrates, proteins and fats and also the DNA in the human body at the molecular level. Molecular diagnostics, imaging techniques and biopsies are all used collectively to diagnose cancer.

Cancer is one of the foremost grounds of death in the United States. Lung cancer is the prominent cause of cancer death in most of the adults; leukemia is the common cancer that is found in young children. The cancers that has been identified early, before metastasis, have the greatest cure rates. Nowadays, there are large number of screening tools to facilitate early identification and treatment [2]. Nevertheless, cancer is diagnosed only at a very later stage when the cancer has been aggravated to other parts of the body. All the existing methods can diagnose cancer only after the disease has crosses Stage 1. The stage defines which selections will be suitable for treatment and enlighten prognoses. The commonly used cancer staging method is the TNM system where the T (1-4) value specifies the size and also the direct degree of the primary tumor. The N (0-3) shows the grade to which the disease has extended and M (0-1) refers whether the cancer has spread to other organs in the body. Though most of the Stage 1 tumors are curable, the Stage 4 tumors are neither untreatable nor inoperable. However, early analysis can amplify the 5 year survival rate of the patients.

As the extent of data is growing proportionally with the magnifying population, there is a great need to excavate the knowledge from the data. Pattern extraction, in other words, the data mining makes it contribution much towards this and uncovers its application in various diverse fields including the healthcare industry. The data mining is the method of analyzing through historical data providing an insight into the patterns from hefty dataset. Researchers are signifying that the application of data mining techniques in identifying pre-diagnosis of the disease can improve practitioner performance. Lung cancer, a disease being highly dependent on historical data can employ data mining for its early detection.

This paper proposes a model for assessing if applying data mining techniques to lung cancer dataset can provide trustworthy performance in the detection of lung cancer at the very early stage. The paper is organized as follows: Section 2 provides a literature survey on using data mining techniques on gene expression data which helps in the early diagnosis of lung cancer. Section 3 deals with the data collection and the method used. Section 4 discusses on the equality test and the experimental evaluation of the classification. Section 5 covers the implementation and conclusion.

Literature survey

Several data mining techniques have been projected during the last few decades for the diagnoses of lung cancer.

Oncologists say that the mutation in cancer is not only present in the tumor affected area but also in the entire body if the gene has been inherited [3]. If the gene is an inherited cancerous gene, the nodule can be found at an early stage. Such gene can be found only by means of extracting the pattern from the previously available gene data as the type of oncogene and the tumor suppressor gene varies from person to person. In this case, the person need not be a smoker. There are around 20,000 deaths every year because of lung cancer among the people who have never smoked.

Gazdar et al. have specified that pinpointing the information on gene and information on the mutation gives researchers a clear target of the therapy that could be undertaken [4]. Data mining can serve as one such technique in pinpointing the gene and its mutation as the experts in the field of bioinformatics and biostatics need to analyze loads of data points from the specimen given to them by the clinicians.

Roslan et al. proposed a method to predict the survival of patients with NSLC unsupervised hierarchical clustering and Pearson correlation [5]. It has been found that there are some distinct gene profile that significantly predict the survival rate. These significant gene could be useful to predict early diagnoses of NSLC.

In 2012, Wang et al. came up with a seed-based approach to identify risk disease sub-networks in human lung cancer [6]. In this paper, the Gene Ontology is incorporated into the microarray gene expression to pinpoint the differentially expressed gene using augmenting fuzzy measure similarity. This method indicates that the sub networks based on disease specific could provide specific information for the development of precision prediction and therapies for lung cancer. This method also uncovers the relationships among genes/proteins and how one gene can affect other gene.

The tumor suppressor genes have been predicted by performing clustering techniques on the profile of developmental stage gene expression by Nitin et al. [7]. The work mainly focuses on identifying the differential feature that serves potentially to differentiate between drivers and passengers. Here, the main challenge was to reproduce the result at a higher level of significance using the k-means classification approach.

Radha et al. have worked on data mining techniques to identify the gene expression profile of breast cancer patients [8]. This technique recognizes a set prognostic biomarkers that are responsible for causing breast cancer. The classification accuracy of k-means is dependent on the initial cluster centroid selection which is a major difficulty faced with the k-means algorithm. Halder et al. presented a semi-supervised classification technique, i.e., fuzzy k-nearest neighbor, which

makes use of both unlabeled and labelled samples to improve the accuracy of prediction for cancer classification [9]. The technique is being applied on various gene expression datasets of cancer. The fuzzy classifiers used gives better accuracy in class prediction of cancer classification problem because in gene expression data, the classes have overlying nature. The class is predicted to be taken whose fuzzy membership is more. Since it makes use of fuzzy algorithm, a clear cut boundary among the classes is not established.

Wei et al. aims to perform a hierarchical clustering of lung cancer related genes. Here, 367 genes downloaded from the Gen Bank have been clustered and the correlativity between the clusters and the Gene Ontology function classification have been studied [10].

In 2004, Dettling previously came up with the bag boosting technique for tumor classification for the gene expression data [11]. The major goal of this approach was to make the class prediction of cancerous malignancies at an early stage. The Bag Boosting consistently lesser the misclassification error rate and is also found to do better than random forest and support vector machine but this technique is time consuming as it take longer time for fitting in complex prediction tools. Whatever the case may be, as the size of the dataset rises the prediction accuracy falls.

Artificial neural networks (ANN) provide a dominant technique to help doctors to analyze, come up with a model and make logics out of composite clinical data through a comprehensive collection of medical applications [12]. The neural network takes a broad view from the encoded input data to patterns that are built-in the data, and utilize these patterns to make classifications of the data. It is basically a model created mathematically on the basis of biological neurons of human brain.

As per the survey done by Juliet et al. about the diagnoses of lung cancer, it suggests that Self Organizing Map could give a better classification compared to other neural network approaches as the network has the capability of adapting to human brain [13, 14].

Material and methods

The data used here is the secondary data which has already been collected from the https://data.world/cancerdatahp/lung-cancer-data.

Experimental evaluation

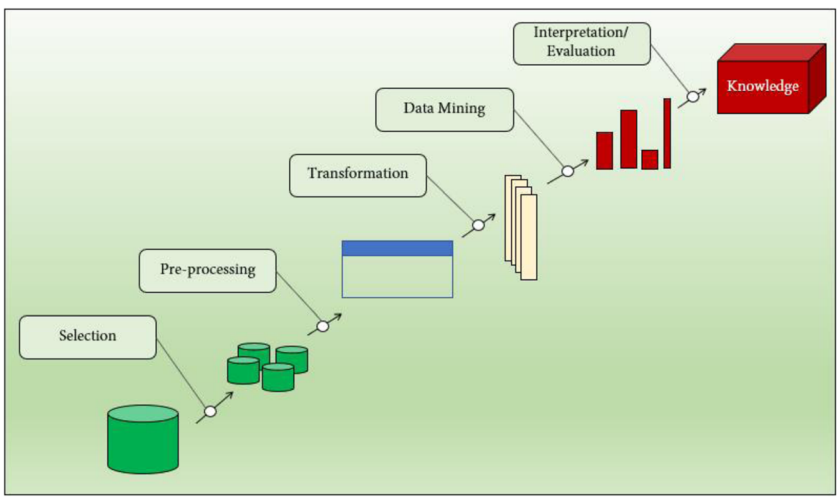
Steps in mining process

Data preprocessing The patient report collected must be processed before applying classification techniques as there might be missing data or irrelevant data. This step is important to get accurate results.

Data classification Machine learning algorithm is applied on the records to perform classification. Here, we use multi-class neural networks to perform the classification.

Pattern extraction On successful classification, the underlying pattern has to be extracted to identify the split attributes. The attributes/parameters that contribute much towards the diagnosis is identified here and put in chronological order.

Pattern representation The identified pattern has to represented in a form so that the medical practitioner can use this information/knowledge for further analysis of the disease.



Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level	Scored Probabilities for Class "High"	Scored Probabilities for Class "Low"	Scored Probabilities for Class "Medium"	Scored Labels
3	4	1	5	2	6	2	Low	0	0.999435	0.002281	Low
2	1	4	7	2	1	6	Medium	0.000379	0.000011	0.996782	Medium
2	3	4	2	1	1	1	Low	0	1	0	Low
2	4	6	5	4	2	5	Medium	0	0.000324	0.999875	Medium
2	1	2	4	2	3	2	Low	0	1	0	Low
7	6	7	8	7	6	2	High	1	0	0.000005	High
4	2	4	2	4	3	1	Low	0	0.998632	0.000001	Low
6	5	1	9	3	4	2	Medium	0.000079	0.000192	0.99973	Medium
2	4	6	5	4	2	5	Medium	0	0.000491	0.999431	Medium

Fig. 1 Snapshot of the classification

Proposed method

The proposed system involves multi-class neural networks algorithm. This follows the supervised learning. The goal of this is to discover some underlying structure of the data. This algorithm requires a tagged dataset. The lung cancer dataset obtained as input from the https://data.world/cancerdatahp/lung-cancer-data database for lung cancer is divided into two sets, as the training data and the testing data. The algorithm performs the classification based on the knowledge learnt from the training data during the learning process.

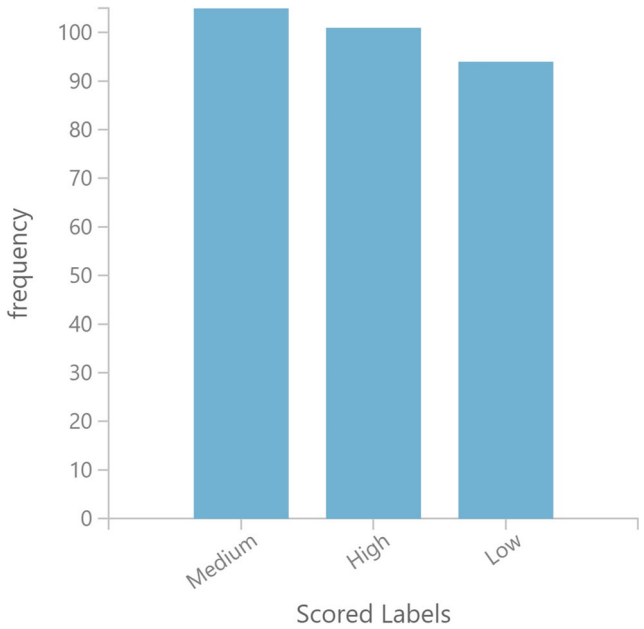


Fig. 2 Classification of the patient record represented in histogram

Experimental evaluation

Multi-class neural networks

A neural network is a set of layers that are interconnected. The first layer is the inputs which are connected to an output layer by an acyclic graph which is comprised of weighted edges and nodes. Multiple hidden layers are present between the input and output layers. The relationship between inputs and outputs is obtained training input data of the neural network. All the nodes in a layer are associated by the weighted edges to nodes in the successive layer.

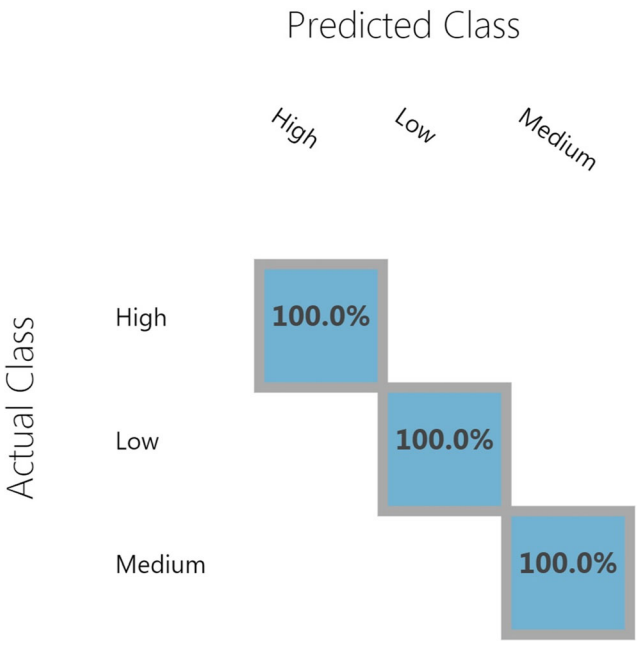


Fig. 3 Confusion Matrix

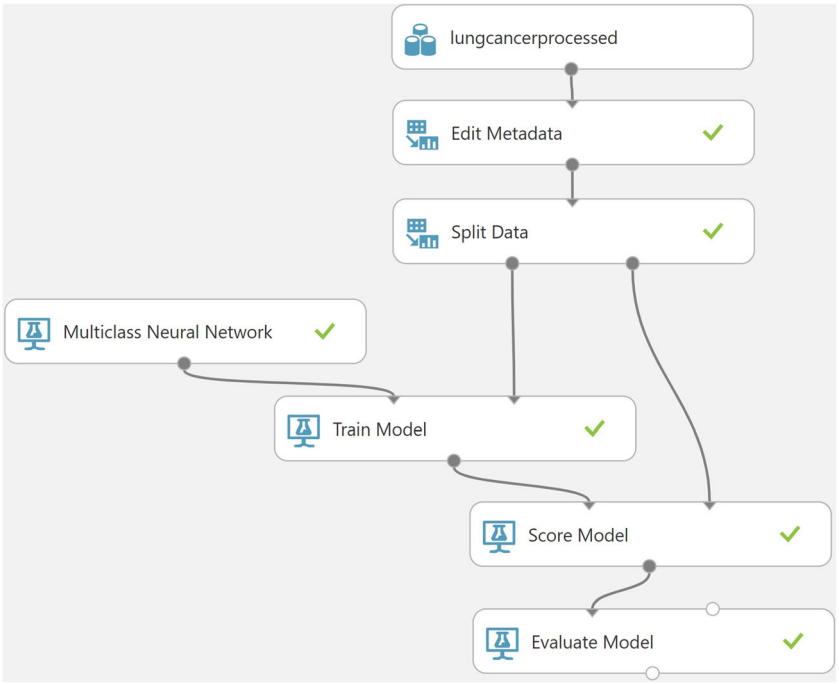


Fig. 4 Classification using Multi-Class Neural Networks

In order to calculate the output of the model for an input, value is calculated at each node in the hidden layers and in the output layer. An activation function is then applied to that weighted sum.

Implementation and result

After performing the classification using multi-class neural network algorithm, the classification happens as given in Fig. 1.

The graphical representation of the multi-class classification is given in Fig. 2.

The prediction accuracy for each of the classes is as per Fig. 3.

Figure 4 shows the overall implementation of the model.

Conclusion

To assist medical practitioners in the diagnoses of lung cancer, recent research has looked into the development of computer based tools [15]. As discussed in [16], X-ray, biopsy, sputum cytology are not suitable for patients with other pathologies. In this paper, a brief discussion is carried out on the various prevailing techniques existing in cancer diagnoses and ultimately come up with a conclusion that the multi-class neural networks could yield a better performance compared to other machine learning technique. The algorithm is implemented using Azure Machine Learning Studio. It has been found that the supervised method did give a very good accuracy. The method can also be used for the diagnoses of cancers like the breast cancer. This model can help assist medical practitioners in the pre-diagnoses process for early detection of cancer thereby increasing the 5 year survival rate of the patients.

Compliance with Ethical Standards

Conflict of Interest This paper has not communicated anywhere till this moment, now only it is communicated to your esteemed journal for the publication with the knowledge of all co-authors.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Cancer: Facts, Causes, Symptoms and Research, http://www.medicalnewstoday.com/info/cancer-oncology; 2015 [accessed August 11, 2016].
2. Cancer: Prevention and Detection, Columbia Electron. Encycl., http://www.infoplease.com/encyclopedia/science/cancer-medicine-prevention-detection.html; 2012 [accessed August 11, 2016].
3. Singh, N. K., Vidyasagar, M., White, M. A., Predicting tumor-suppressing genes in cancer via clustering the developmental stage gene expression profile. Proceeding of the 2011 IEEE/NIH Life Science Systems and Applications Workshop. p 116–20, 2011.
4. Gazdar, A., Robinson, L., Oliver, D., Xing, C., Travis, W. D., Soh, J. et al., Hereditary lung cancer syndrome targets never smokers with germline EGFR gene T790M mutations. J Thorac Oncol. 9: 456–463, 2014.
5. Harun, R., Hadi, J., Mhazir, N. S., Chyang, P. J., Rose, I., Manap, R. A., et al. Gene expression profiles predict survival of patients with advanced non-small cell lung cancers. Proc. Fourth Int. Conf. Model. Simul. Appl. Optim. p 1–4, 2011.

6. Wang, Y. B., Cheng, Y. M., Zhang, S. W., Chen, W., A seed-based approach to identify risk disease sub-networks in human lung cancer. Proceedings of the 2012 IEEE 6th International Conference on Systems Biology. p 135–41, 2012.
7. Inherited Risk Mutation for Lung Cancer? Researchers Launch INHERIT EGFR Registry to Investigate, https://www.lungcancerfoundation.org/2013/05/inherited-risk-mutation-for-lung-cancer-researchers-launch-inherit-egfr-registry-to-investigate/;2013 (accessed August 14, 2016).
8. Radha, R., Rajendiran, P. Using K-means clustering technique to study of breast cancer. Proceedings of the 2014 World Congress on Computing and Communication Technologies. p 211–4, 2014.
9. Halder, A., Misra, S., Semi-supervised fuzzy K-NN for cancer classification from microarray gene expression data. Proceedings of the International Conference on Engineering Technology and Technopreneush. p 156–60, 2014.
10. Wei, Y., Huajia, Z., Kuanheng, W., Qiangqian, L., Miao, H., Hierarchical Clustering of Lung Cancer Related Genes. Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering. p 63–5, 2008.
11. Dettling, M., BagBoosting for tumor classification with gene expression data. Bioinformatics. p. 3583–93, 2004.
12. Al-Shayea, Q. K., Artificial neural networks in medical diagnosis. Int J Comput Sci. 8:150–154, 2011.
13. Rajan, J. R., Chelvan, C. C., A survey on mining techniques for early lung cancer diagnoses. Proceedings of the 2013 International Conference on Green Computing, Communication and Conservation of Energy. p. 918–22, 2013.
14. Burges, C. J. C., A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov 2:121–167, 1998.
15. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 16:906–914, 2000.
16. Wang, L., Screening and biosensor-based approaches for lung cancer detection. Sensors. 17:2420, 2017.