ORIGINAL RESEARCH PAPER

Engineering

MACHINE LEARNING TECHNIQUE TO PREDICT THE MODEL FOR NON-SMALL CELL LUNG CANCER-BINARY LOGISTIC REGRESSION MODEL APPROACH

KEY WORDS: Lung cancer, Non-small cell lung cancer (NSCLC), Genes, Regression analysis, Geo-datasets.

K. Vaishnavi Devi* Sri Ramachandra Faculty of Engineering and Technolgy Sri Ramachandra Institute of Higher Education and Research (DU) Porur, Tamilnadu, India. *Corresponding Author

S. Venkatesan Sri Ramachandra Faculty of Engineering and Technolgy Sri Ramachandra Institute of Higher Education and Research (DU) Porur, Tamilnadu, India.

ABSTRACT

Lung cancer is one of the common types and deadly cancer in both men and women. This lung cancer accounts for high mortality and morbidity throughout the world. Detection of lung cancer has been made through surgery, chemotherapy, biopsy and microarray studies. Gene expression plays an important role in molecular fluctuations and disease prophecy of a disease. The aim of the study is to design a statistical model and to find the genes influencing the cause of lung cancer. Microarray gene expression data was collected from Gene Expression Omnibus datasets (GEO-DATASET)-an open source database. The dataset contains a total of 161 samples which has 89 lung cancer samples and 72 normal samples. From this the upregulated and influenced genes were identified and determined by using logfc from the GPL file. Wide use of statistical models leads to exploring machine learning methods to find a better model. These study methods implement the performance of regression analysis using multilayer perceptron. By using the regression analysis method, the overall accuracy is found to be 91.3%. By this, the gene expression data analysis reveals that the regression analysis is one of the best models to show the accuracy in implementation of genes influencing the NSCLC.

INTRODUCTION:

Lung cancer (LC) is one of the common types of cancer in both men and women which accounts for 13% of all new cancer cases and 19% of cancer related deaths worldwide. In India, LC accounts for 6.9% of all new cancer cases and 9.3% of all cancer related deaths in both the sexes[1]. Lung cancer has a high fatality rate compared with other cancers[2]. Lung cancer begins in the lungs and may even spread to lymph nodes or also to other parts of the body. Lung cancer is mostly caused by cigarette smoking[3]. There are different types of lung cancer namely-Small cell lung cancer (SCLC), Non-small cell lung cancer (NSCLC), Adenocarcinoma, Squamous cell carcinoma, large cell carcinoma and Bronchial carcinoid [4]. Lung cancer is majorly caused due to smoking, asbestos, radon, passive smoking, heredity and air pollution[5]. Lung cancer can be diagnosed by imaging tests like CT/X-rays, sputum cytology, bronchoscopy, transthoracic needle aspiration (TTNA), biopsies etc[6]. Treatments majorly include surgeries, radiotherapy, chemotherapy [7].

A dataset has been studied and taken for the analysis of non-small cell lung cancer by using regression analysis. Regression analysis is a statistical method which is used to look into the relationship between the two or more variables. It helps in determining the influence of dependent and independent variables. An independent variable is the variable which is changed or controlled by the experimenter to check for the effects of independent variables. Whereas, a dependent variable which is being tested and measured in an experiment[8]. By using this study, it helps in determining the top upreguated and influenced genes which helps in obtaining the accuracy rate. So it is easy to conclude that binary logistic regression analysis is one of the better models for the prediction of non-small cell lung cancer using datasets analysis.

OBJECTIVES:

The primary objective of the study is to build the binary logistic regression model for the prediction of non-small cell lung cancer by comprehensive searching of datasets on microarray gene expression profiling.

METHODOLOGY:

For the analysis of lung cancer, A gene expression profiling data (microarray data)-GSE18385 (Smoking-induced-upregulation of AKR1B10 expression in the airway epithelium of healthy individuals) from gene expression omnibus (GEO) from NCBI. A dataset contains a total of 161samples which has 89 smokers with lung cancer cases and 71 non-smokers with no lung cancer groups are defined for the samples as smokers (case) and non-smokers(control)[9].Samples are assigned to each group. Bonferroni method is selected to adjust the p-values for multiple comparisons. Bonferroni method is a statistical method which compares and counteracts multiple problems. Finally click "top250"to perform the calculation. The results are presented in a separate table which has top 250 values with its unique ID, gene symbol, p-value, logFC, logFC value is copied in a separate excel sheet to calculate the top 10 influencing genes in smokers. A platform file of this dataset is downloaded and the samples file for each sample has been downloaded to extract a value for each top 8 genes in both smokers and non-smokers respectively.

A complete set of data for all the top 8 genes were studied and it is studied in IBM SPSS software for regression analysis to find its accuracy. In SPSS, under the "analyze" option, select regression and select logistic regression. Select the dependent and independent variables as smokers and non-smokers and top 8 influencing genes respectively. And, the method of selection is the backward Wald method. Finally, there appears a complete set of data which has case processing data summary, classification table for smokers and non-smokers with predicted and observed groups and percentage correct, variables in equation and variables not in equation, backward stepwise bald methods which has 4 steps including chi-square, model summary with -2log likelihood, cox-Snell R square, Nagelkerke R square, classification table of smokers and non-smokers with observed and predicted correct percentage with variables in equation and with no variables in equation. A set of data containing a top 8 upregulated influencing genes in both smokers and non-smokers have been run in an SPSS software to obtain data on binary logistic regression analysis to find its accuracy rate.

RESULT:

Table 1: Top 8 influencing and upregulated genes in case of non-small cell lung cancer based on log fold change (logFC value):

ID	logFC	Gene symbol
240699_at	1.808	SEC14L3
204041_at	1.162	MAOB
202018_s_at	1.69	LTF
212750_at	1.322	PPP1R16B
222871_at	1.369	KLHDC8A

219049_at	1.303	CSGALNACT1
223597_at	2.055	ITLN1
229158_at	1.253	WNK4

The dataset GSE18385 was retrieved from GEO-DATASET was pasted in an excel sheet to find the top differentially expressed genes. Out of top 8 upregulated genes, only 1 gene was identified as best influenced upregulated gene in non-small cell lung cancer.

Table 2: Classification of smokers and non-smokers from a linear binary regression for recognition of non-small cell lung cancer.

This table depicts the values and accuracy which were analyzed using SPSS software to build a binary logistic regression model for non-small cell lung cancer. After performing binary logistic regression analysis, the net overall accuracy percentage is found to be 91.1%.Sensitivity, specificity, positive predictive value and negative predictive value.

	Smokers	Non-smokers
Smokers	79	9
Non-smokers	5	67

Sensitivity 94.65%,Specificity 88.16%.PPV 89.77%.NPV 93.06%. and the overall accuracy is 91.1%.

CONCLUSION:

A gene expression profiling by microarray dataset (GSE18385) on non-small cell lung cancer was searched using GEO-dataset in NCBI.A regression analysis for this dataset has been built using SPSS software. The accuracy obtained by this analysis is 91.1%.So, it is concluded that binary logistic regression analysis is one of the best methods to predict the non-small cell lung cancer based on gene expression data.

REFERENCE:

1. Raina V, Malik P. Lung cancer: Prevalent trends & emerging concepts. Indian Journal of Medical Research. 2015. p. 5. doi: 10.4103/0971-5916.154479
2. Website. [cited 5 Mar 2020]. Available: John Stoddard Cancer Center. Top Five Most Dangerous Cancers in Men and Women. [cited 22 Feb 2020]. Available: https://www.unitypoint.org/homecare/services-cancer-article.aspx?id=c9f17977-9947-4b66-9c0f-15076e987a5d
3. Website. [cited 5 Mar 2020]. Available: Lung cancer - Symptoms and causes. In: Mayo Clinic [Internet]. 13 Aug 2019 [cited 22 Feb 2020]. Available: https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620
4. Website. [cited 5 Mar 2020]. Available: Types of lung cancer | Cancer Research UK. [cited 22 Feb 2020]. Available: https:// www. cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/ types
5. Website. [cited 5 Mar 2020]. Available: What Causes Lung Cancer. In: American Lung Association [Internet]. [cited 25 Feb 2020]. Available: https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/learn-about-lung-cancer/what-is-lung-cancer/what-causes-lung-cancer.html
6. Website. [cited 5 Mar 2020]. Available: Wexler A. Stages of lung cancer: Stages, symptoms, and diagnosis. In: Medical News Today [Internet]. 6 Sep 2019 [cited 2 Mar 2020]. Available: https://www. medicalnewstoday. com/articles/316198
7. Website. [cited 5 Mar 2020]. Available: Lung cancer - Diagnosis and treatment- Mayo Clinic. 13 Aug 2019 [cited 2 Mar 2020]. Available: https://www. mayoclinic.org/ diseases-conditions/lung-cancer/diagnosis-treatment/ drc-20374627
8. Dependent and independent variables review (article) | Khan Academy. In: Khan Academy [Internet]. [cited 5 Mar 2020]. Available: https:// www.khanacademy.org/math/pre-algebra/pre-algebra-equations-expressions/pre-algebra-dependent-independent/a/dependent-and-independent-variables-review
9. GEO Accession viewer. [cited 5 Mar 2020]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18385