Identification of Influential Genes for Lung Cancer by Machine Learning Approaches

***Abstract*—Lung cancer is one of the most severe diseases, and nowadays forecasting it is the most challenging task. As most of the cancer cells are overlapped with each other, it is difficult to detect the cells but also it is crucial to identify the existence of cancer cells in the early stages. Early diagnosis, therapy, and prognosis of lung cancer may be enhanced by understanding the molecular pathways underlying its onset and progression. In the case of treatment for cancer, influential genes (IFGs) identification is necessary. However, it is still insufficient, and more research is needed in this regard. Thus, to further contributions, we have taken the Cancer Genome Atlas (TCGA) dataset to detect the IFGs of LC using the Kruskal-Wallis test and Bonferroni correction. We have successfully identified 14 IFGs from 18784 genes and also differentiated the up and down-regulated genes using fold change values and heatmap plot. The accuracy of our suggested method was predicted using a classifier algorithm such as the Support Vector Machine (SVM), and we discovered an accuracy of 84.54%. Our identified 14 influential genes may be used in further lab-based analysis and to develop therapeutic treatment strategies for LC.**

***Index Terms*—Lung Cancer, Influential Gene, Kruskal-Wallis Test, Bonferroni Correction, Up Regulated gene, Down Regulated gene, Support Vector Machine (SVM)**

I. Introduction

Lung cancer is a type of cancer that is very common around the world and accounts for 25% of cancer deaths and is the most familiar cause of cancer-related death in both men and women. It is the second cancer in both men and women that are examined often. Lung cancer is expected to account for an average of 234,030 new cases in 2018, or 14% of all cancer cases, according to research [1]. For men, the rate has been dropping since the middle of the 1980s; for women, it has only been doing so since the middle of the 2000s. Genuine examples of smoking start-up and cessation during the past few decades are reflected in gender orientation contrasts. In men, frequency rates fell by 2.5% annually, whereas, in women, they fell by 1.2% [1]. In this research, we have used machine learning approaches to identify the IFGs of LC. Firstly, we collected the TCGA dataset of LC from cBioPortal and applied the Kruskal-Wallis test to the data to calculate the p-value. The Kruskal-Wallis test is a nonparametric method that is also known as the one-way ANOVA test through which we got the p-value. On the other hand, Bonferroni correction has also been used with the Kruskal-Wallis test which is the simplest method to repel multiple comparisons problem-solving [2] [3]. For getting the Bonferroni adjusted or corrected p-value, we need to divide the original α-value by the number of analyses on the dependent variable. By using the above-mentioned algorithms, we have found the Influential genes. After that, we used the machine learning approach for the classification of SVM. We have also differentiated the up and down-regulated genes. To build up the training dataset, we have selected some samples erratically and the rest of the samples are testing datasets. In fact, using the training dataset we build up the classification function F(x), and the class of the testing dataset is also determined in response to F(x). Comparing with the actual class label we have measured the accuracy of the classification. We also have done the confusion matrix.

II. Methodology

A. Working principle

We have collected the Cancer Genome Atlas (TCGA) dataset for LC from cBioPortal where control and affected cells were found. Then we preprocessed the dataset and applied two machine learning algorithms such as Kruskal Wallis Test and Bonferroni Correction to find the influential genes. We have differentiated the up and down-regulated genes from the 14 identified IFGs. Finally, we have used a classifier algorithm including SVM for classification and prepared the confusion matrix. The working principle of the study is illustrated in Figure 1.

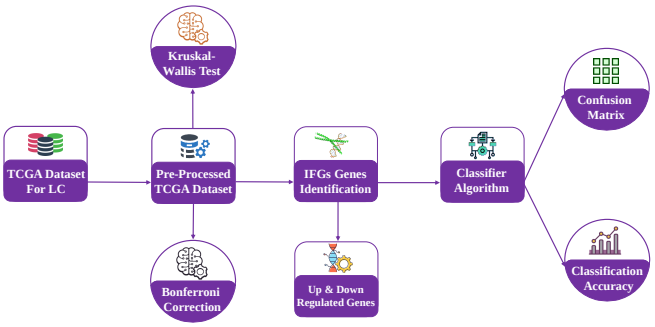

Fig. 1. Flow diagram of the analytical approach used in this study.

B. Dataset description and preprocessing

In this study, we examined the Cancer Genome Atlas (TCGA) dataset named Lung Adenocarcinoma (OncoSG, Nat Genet 2020) [4] from cBioPortal which is free and available to work. The LC dataset is made up of microarray data from 305 lung cancer patients. There are 302 samples with 181 attributes in the clinical dataset of LC (Lung Adenocarcinoma, OncoSG, Nat Genet 2020) including 302 cases. After identifying the dataset, we preprocess the dataset (LC_data_mRNA_median_Zscores.csv) e.g. remove the null

Purpose: The Bonferroni correction adjusts probability (p) values because of the increased risk of a type I error (A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population) when making multiple statistical tests.

value of the dataset, replace the gene name and attribute with the nominal value to make it easier to import the dataset into RStudio. After importing the dataset, we started to conduct the dataset as per our requirements. The narration of the dataset is given in Table I.

TABLE I
THE NARRATION OF THE USED MICROARRAY DATASET

Disease Name	Datasets Name in the cBioPortal	Number of Samples		
		Patients	CNA	RNA-Seq
Lung Cancer (LC)	Lung Adenocarcinoma (OncoSG, Nat Genet 2020) [4]	305	302	181

C. Methods for influential genes identification

1) Kruskal-Wallis Test: Kruskal-Wallis Test is a non-parametric approach that is accustomed to compare various independent arbitrary samples in place of one-way ANNOVA. We assume that the population samples are randomly picked and reciprocally independent. Each and every one of the samples is independent and is computed on an ordinal scale [5]. If no samples are egalitarian, the Kruskal-Wallis test can be defined as follows:

$$H = \frac{12}{N(N+1)}[\sum_{i=1}^{c}\frac{R_i}{n_i} - 3(N+1)] \quad (1)$$

Where C is the number of groups or samples, Ri indicates the sum of the rankings in the I^th sample, Ni represents the number of inspections in the I^th sample, and N indicates the entire number of Ni, which refers to the entire number of inspections in all integrated samples.

If the H value is large, the test result is abandoned by the null hypothesis. The hypothesis HO estimates that all groupings condescend from the same population. As it is an authentic dataset, no tie can be endorsed. For this type of real-world dataset, the tie is a congenital fit. If ties arise, every particular investigation is given the average of the ranks for which it is bounded. The calculation of the H then finds out through the above equation and divided by,

$$1 - \frac{\sum T}{N^2 - N} \quad (2)$$

Where T = (t - 1)t(t + 1) = t3 -t for every particular group of interrelations. The number of tied inspections in the group is specified by t. The entire is estimated by all the tie groupings. As a consequence, the final equation of H is as follows:

$$H = \frac{\frac{12}{N(N+1)}[\sum_{i=1}^{c}\frac{R_i}{n_i} - 3(N+1)]}{1 - \frac{\sum T}{N^2-N}} \quad (3)$$

The differentiation between the H of the first equation and the H of the third equation can be quite diminutive at times. At the time of handling the ties using mean ranks, H can be assorted as

$$X^2(c-1) \quad (4)$$

2) Bonferroni correction: A Bonferroni correction is based on Olive Jean Dunn's theory of Bonferroni inequality for determining confidence intervals. It is a trouble-free solution to prevent multiple comparisons. BC provides a greater chance of failure to reject a false null hypothesis rather than other approaches, as the potentially beneficial information such as the dispensation of p-values across all inspections is ignored by it [3]. Let there be k means, and they are as follows μ_i,....,μ_k. Based on their estimation, they are $\mu_1^{'}$,....,$\mu_k^{'}$. As a result of variance $a_{ii}\sigma^2$ in which i=1,....,k and With covariance $a_{ij}\sigma^2$ interim $\mu_i^{'}\mu_j^{'}$ contrariwise $i \neq j$. Additionally, let $\sigma^{'2}$ be an estimation of σ^2 with degrees of freedom $\mu_1^{'}$,....,$\mu_k^{'}$ such that $\frac{v\sigma^{'2}}{\mu^2}$ follows a chysquare distribution where degrees of freedom is v [5]. Let m be the linear combination of means, and let m be estimated as follows:

$$\theta_s = c_{1s}\mu_1 + ... + c_{ks}\mu_k; s = 1, ..., m \quad (5)$$

In this case, linear contrast refers to

$$\sum_{i=1}^{k} c_{is} = 0 \quad (6)$$

The unbiased estimates for θ_1,.....,θ_k are as follows:

$$\theta_s^{'} = c_{1s}\mu_1^{'} + ... + c_{ks}\mu_k^{'}; s = 1, ..., m(7)$$

As a result $\theta_s^{'}$ is the variance of ${b_s}^2\sigma^2$ in which

$${b_s}^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}c_{is}c_{js} \quad (8)$$

This makes finding confidence intervals easier. Based on t distributions with v degrees of freedom, the variates $t_1, ..., t_m$ will be

$$t_s = \frac{\theta_s^{'} - \theta_s}{b_s\sigma^{'}}(9)$$

By using the Bonferroni inequality, we can calculate the lower probability limit where all $t|_2$ lie between -c and +c. Therefore

$$P[-c < t_i < +c; i = 1, 2, ..., m] \geq 1 - 2m\int_c^{\infty} f^{(v)}(t)dt \quad (10)$$

The frequency of Student t variable is represented by $f^{(v)}(t)$. In this case, v represents the degree of freedom. The right side of equation 9 equals 1- α if we select c accordingly. Thereafter

$$P[-c < \frac{\theta_s^{'} - \theta_s}{b_s\sigma^{'}} < +c; s = 1, 2, ..., m] \geq 1 - \alpha(11)$$

1-α represents the overall level of confidence here. Defining c as

$$\int_c^{\infty} f^{(v)}(t)dt = \frac{\alpha}{2m} \quad (12)$$

Assuming that the sample means of $y_1^{'}$,....,$y_k^{'}$ are $\mu_1^{'}$,....,$\mu_k^{'}$ and $y_1^{'}$,....,$y_k^{'}$ have a statistically independent correlation.Therefore, the $C_{1s}\mu_1$ + ... + $C_{ks}\mu_k$ confidence interval would be

$$c_{1s}y_1^{'} + ... + c_{ks}y_k^{'} \pm c\sqrt{\frac{\sum_{i=1}^{k}{c_{is}}^2}{n_i\sigma^{'}}}; s = 1, 2, ..., m(13)$$

To pick out IFGs from LC dataset, the Kruskal-Wallis H test and Bonferroni correction methods are used. The chosen experimental steps are as follows [5]:

- Preprocess the TCGA datasets for implementing tests.
- Pick out the 'alpha' value as the significant level.
- For every particular existing gene
  - Make vectors of TCGA dataset for every particular class
  - The vectors are being scaled
  - Find P-value through Kruskal-Wallis H-test
- Set adjusted P-value' through BC.
- The IFGs are selected through which the adjusted assurance level is in expectation.

D. Methods for Classification

Utilizing the identified influential genes, the samples of the dataset have been classified. Classification algorithm Support Vector Machine (SVM) have been used for calculating classification accuracy.

1) Working Procedure for Classification: Our classification procedure using the identified influential genes is as followed [5]:

- Generate new dataset including the influential genes
- Divide the dataset into two part; one for training dataset and other of testing
- Prepare the SVM classifier with proper arguments, training dataset additionally choose the best model
- Classify the testing samples with response to the training classifier
- Evaluate the classification accuracy

2) Support Vector Machine (SVM): A superintend machine learning algorithm is SVM. Through which an optimal hyperplane can be found by isolating the two classes with the highest possible margin. In this research, the SVM has been used with "C-Classification" [6]. Assume that, there are training vectors $x_i \in R^n; i = 1, 2, ...l$ in which there are two classes and an indicator vector $y \in R^l$ that is alike $y_i \in -1, 1$. It discover a solution to the restricted optimization problem which is shown below

$$min_\omega, b, \in \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l} \in_i \tag{14}$$

Contingent on $y_i(\omega^T\varphi(x_i)+b) \geq 1- \in_i, \in_i \geq 0; i = 1, 2, ...l$ in which, C is bigger than 0 and works as the formalize parameter $\varphi(x_i)$ is function that delineates to a bigger spatial space. Because of the high spatial behavior, we find out the solution of the problems

$$min_\alpha \frac{1}{2}Q^\alpha + e^T\alpha \tag{15}$$

Contingent on $y^T\alpha = 0, 0 \leq \alpha_i \leq C, i = 1, 2, ...l$ in which $e = [1, 1, ..., 1]^T$. Q is 1 by 1 Such as semidefinite affirmative matrix $Q_ij = y_iy_j(x_i, x_j)$ K is known as a kernel. After getting the solution of the eq. ω gratifies

$$\omega = \sum_{i=0}^{l} y_i\alpha_i\varphi(x_i) \tag{16}$$

Then the final conclusion function will be

$$sgn(\omega^T\varphi(x) + b) = sgn(\sum_{i=0}^{l} y_i\alpha_iK_(x_i, x) + b) \tag{17}$$

E. Up and down-regulated genes identification

We differentiated the up and down-regulated genes from the influential genes using their fold change values. We used a proportional fold change value due to a zero average LC dataset [4]. If the value of the fold change is larger than 0.5 then the gene is over-expressed in the X category specimen and is therefore up-regulated. Otherwise, the value of the proportional fold change represents the down-regulated gene.

III. Result Analysis

A. Influential genes identification

We have explored the TCGA dataset of LC from cBioPortal for identifying the IFGs. We have pre-processed the original dataset in which all the samples are in different classes. If the class label is the same then it will be put on one cluster but if the class label is different, it will be put on another cluster. There are 1 and 2 class labels in our dataset in which the 60 samples are in class label 1 and another 110 samples are in class label 2. After preprocessing the range of the label 1 dataset contains 1- to 60th columns in the matrix and the range of the label 2 dataset contains 61- to 170th column. We got a total of 18784 genes after completing preprocessing. To identify the IFGs, we applied two algorithms such as Kruskal's Wills H test and Bonferroni correction for finding the p-values and adjusted p-values. Besides, we had to keep the adjusted p-value less than or equal to 0.05 for selecting the appropriate IFGs. At last, we have recognized 14 IFGs from 18784 genes of LC successfully. These 14 IFGs are either up-regulated or down-regulated genes. We have established an IFGs regulatory network, which is shown in figure 2.

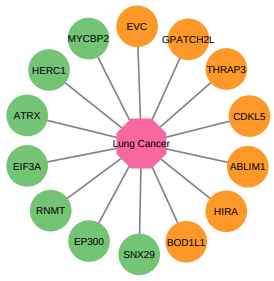

Fig. 2. IFGs regulatory network for LC. Here dark pink color centered octagon shape represents lung cancer, light orange color circular nodes/shapes represent upregulated genes and green color circular shapes represent down-regulated genes.

B. Classification of the LC dataset

The dataset contains 18784 genes among those 14 are identified as influential genes which have their own class label. After that, we divided the datasets into 35% data randomly using RStudio which has different random functions and other samples are used for testing. We have used SVM

classifier model which have different arguments as type = 'C-classification', kernel = 'radial', cost = 32, gamma = 0.1428571.The classifier is then trained with a training dataset. Testing datasets are tested with trained models. Under the circumstances, the classification accuracy is 84.54% of it. A confusion matrix can be used to assess the performance of our classifier when presented with new data. Using the prediction, we can compute the confusion matrix and see the accuracy score. As SVM provides the highest accuracy among all the classification algorithms for our study, so we generate a confusion matrix that shows (Table II).

TABLE II
CONFUSION MATRIX WITH PREDICTION AND REFERENCE AND FOR THE DATASET OF LC.

	Reference	
Prediction	**1**	**2**
1	27	12
2	5	66

C. Identified Up and down-regulated genes

We have distinguished the up and down-regulated genes through the fold change value, if the fold change value of a gene is bigger than 0.5 then it is considered as up-regulated otherwise it is down-regulated gene. We have found 7 up-regulated and 7 down-regulated genes among our identified 14 influence genes that are presented in Figure 2, Table III We specify biomarker genes for lung cancer based on logarithm fold change and heatmap. We have determined 14 influential genes from 18784. Up-regulated or down-regulated values can specify the biomarker gene for (type 1) lung cancer in its expression. e.g. EVC, ATRX, GPATCH2L, THRAP3, CDKL5, ABLIM1, HIRA, BOD1L1 are biomarker genes [5]. The adjusted p-values and proportion FC values for each gene are shown in Table III. Figure 3 shows that ATRX, GPATCH2L, THRAP3, CDKL5, ABLIM1, HIRA, and BOD1L1 genes form a different cluster from others. Based on the heatmap results, it can be concluded that the fold change value of the lung cancer dataset can be used in order to identify biomarker genes.

IV. DISCUSSION

Every year more than 10 million people die due to cancer and LC is the most agitating one among those cancers. About 56% LC-attacked patients cannot survive. In this study, we have created a model to detect the IFGs that may be responsible for occurring LC. At first, we collected the TCGA dataset of LC from cBioPortal. We have then preprocessed the dataset and applied the Kruskal-Wallis test and Bonferroni correction to detect IFGs [2] [3]. As a result, we have successfully identified 14 IFGs from 18784 genes. Also, differentiated 7 up and 7 down-regulated genes. We have applied the SVM classifier algorithm for classification and found it 84.54% accuracy for our dataset. Then, we prepared the confusion matrix for SVM that presents their predictions with reference. Finally, we can recommend that our identified 14 influential genes may be used in further lab-based analysis, therapeutic treatment strategies, drug designing, and survival predictions for LC which would be beneficial for other researchers, additionally save their time and potentially save a valuable life.

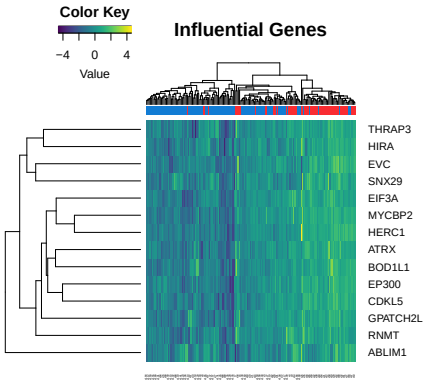

Fig. 3. Heatmap plot of the influential genes for lung cancer. Here 7 genes are downregulated and the rest 7 genes are upregulated.

TABLE III
THE UP-REGULATED AND DOWN-REGULATED IFGS OF LC AND WITH ADJUSTED P-VALUE AND FC VALUE

Gene Name	p-adjusted value	Proportion FC value
CDKL5	0.01045239	1.48710929831832
BOD1L1	0.004349733	3.552919731
THRAP3	0.003560394	0.875718597
EVC	0.023125837	0.613435118
GPATCH2L	0.036689184	1.320884525
ABLIM1	0.015582896	0.761097544
HIRA	0.017531098	2.671130546
MYCBP2	0.014440747	-11.21569018
SNX29	0.015891384	-4.657832766
ATRX	0.028459763	-11.6941285
EP300	0.003189949	-0.10807445
RNMT	0.02278102	-4.610186988
HERC1	0.00697347	0.485509327
EIF3A	0.000894213	-0.575757576

REFERENCES

[1] F. T. Johora, M. H. Jony, M. S. Hossain, and H. K. Rana, "Lung cancer detection using marker controlled watershed with svm," *GUB Journal of Science and Engineering*, vol. 5, no. 1, pp. 24–30, 2018.

[2] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[3] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[4] J. Chen, H. Yang, A. S. M. Teo, L. B. Amer, F. G. Sherbaf, C. Q. Tan, J. J. S. Alvarez, B. Lu, J. Q. Lim, A. Takano *et al.*, "Genomic landscape of lung adenocarcinoma in east asians," *Nature genetics*, vol. 52, no. 2, pp. 177–186, 2020.

[5] U. Das, M. A. M. Hasan, and J. Rahman, "Influential gene identification for cancer classification," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2019, pp. 1–6.

[6] C. Chih-Chung, "Libsvm: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27–1, 2011.