Published online 2023 August 20 **Original Article**

Machine Learning Model-based Detection of Potential Genetic Markers Associated with the Diagnosis of Small-cell Lung Cancer

Mehmet Ediz Sarihan[1*], Zeynep Kucukakcali[2*] and Ibrahim Tekedereli[3]

[1]Department of Emergency Medicine, Inonu University Faculty of Medicine, Malatya, Turkey
[2]Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, Malatya, Turkey
[3]Department of Medical Biology and Genetics, Faculty of Medicine, Inonu University, Malatya, Turkey

*** *Corresponding author:*** Mehmet Ediz Sarihan, Department of Emergency Medicine, Inonu University Faculty of Medicine, Malatya, Turkey. Email: edizsarihan@hotmail.com
Zeynep Kucukakcali, Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, Malatya, Turkey. Email: zeynep.tunc@inonu.edu.tr

Received 2023 February 06; **Revised** 2023 March 27; **Accepted** 2023 July 22.

Abstract

Background: Small-cell lung cancer (SCLC), which is in the category of intractable cancers, has a low survival rate. It is essential to understand the pathophysiological pathways underlying its development to create powerful treatment alternatives for the disease.
Objectives: This study aimed to classify gene expression data from SCLC and normal lung tissue and identify the key genes responsible for SCLC.
Methods: This study used microarray expression data obtained from SCLC tissue and normal lung tissue (adjacent tissue) from 18 patients. An Extreme Gradient Boosting (XGBoost) model was established for the classification by five-fold cross-validation. Accuracy (AC), balanced accuracy (BAC), sensitivity (Sens), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), and F1 scores were utilized for performance assessment.
Results: AC, BAC, Sens, Spec, PPV, NPV, and F1 scores from the XGBoost model were 90%, 90%, 80%, 100%, 100%, 83.3%, and 88.9%, respectively. Based on variable importance values from the XGBoost, the HIST1H1E, C12orf56, DSTNP2, ADAMDEC1, and HMGB2 genes can be considered potential biomarkers for SCLC.
Conclusion: A machine learning-based prediction method discovered genes that potentially serve as biomarkers for SCLC. After clinical confirmation of the acquired genes in the following medical study, their therapeutic use can be established in clinical practice.

Keywords: Classification, Machine learning, Potential biomarkers, Small-cell lung cancer

1. Background

Lung cancer has become the most prevalent of all cancers detected in the past few decades worldwide. In 2018, 2.1 million new lung cancer cases accounted for 12% of the global lung cancer cases (1, 2). Small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) are the two broad histological classifications of lung cancer tumors. NSCLC accounts for 80% to 85% of lung malignancies, with adenocarcinoma accounting for around 40%, squamous cell carcinoma accounting for 25% to 30%, and large cell carcinoma accounting for 10% to 15% (3, 4). SCLC represents approximately 15% of all lung cancers and is known for its high proliferative rate, early metastasis, and poor prognosis. It remains a substantial health conundrum (5). SCLC is strongly associated with tobacco consumption and exposure, and tobacco usage is the primary risk ingredient responsible for the high mutational burden of SCLC (6, 7). Diminishing smoking habits in Western communities has decreased the incidence of SCLC in the last 20 years. The primary obstacles to an early diagnosis of the disease continue to be the absence of particular symptoms in the early stages of tumor development and the lack of screening techniques. Only one-third of patients are diagnosed in the early stages and are potentially suitable for curative treatment (5, 6). The five-year survival rate of the disease is less than 7% (8). Surgery has limited therapeutic benefits for SCLC patients due to the early and fast metastasis that is typical of the illness, and chemotherapy and radiation treatments have short-lived effects. Unlike NSCLC, where genotype-directed therapies have significantly improved treatment outcomes for many patients, SCLC has no approved targeted therapies due to the lack of clear kinase targets (9, 10). In addition, the re-use of drugs in current treatments has not shown significant clinical effects (11). Therefore, there is a critical need to identify new therapeutic targets and strategies for SCLC. A deeper comprehension of the pathophysiological pathways underlying the beginning and development of SCLC is required to create more potent therapy alternatives. According to the Stubborn Cancer Research Act (RCRA) of 2013, which defines SCLC as "recalcitrant cancer", it is imperative to create better preclinical models and therapeutic approaches for this condition. For this reason, interest in studies to determine the genomic infrastructure related to SCLC has increased (12). Health authorities hope that the treatment of SCLC patients will be improved in the coming years by applying target genomic biomarkers and introducing

immunotherapy.

Machine learning (ML) is a branch of artificial intelligence (AI) that makes predictions based on data. AI/ML algorithms have been widely employed in illness diagnosis and clinical decision support systems in recent years. It has a wide range of applications in the health sector for areas such as early detection of genetic diseases and cancer. With the availability of massive datasets and increased processing power over the past decade, ML approaches have attained excellent performance in a variety of contexts (13, 14).

2. Objectives

This study aimed to classify gene expression data from patients with SCLC tissue and normal lung tissue (adjacent tissue) using the Extreme Gradient Boosting (XGBoost) approach and identify significant genes that may contribute to the development of SCLC.

3. Methods

3.1. Data Collection and Variables

In the present study, which is a retrospective case-control study, one of the ML methods, XGBoost, was used to classify microarray expression data obtained from open-access SCLC tissue and normal lung tissue and identify new candidate genes that could be biomarkers for SCLC. The dataset was obtained by taking SCLC tissue and normal lung tissue (adjacent tissue) samples from the lungs of 18 patients. Samples were obtained by surgical resection, and gene expression profiling was performed by the microarray method. In the dataset used in the study, of the 18 patients whose tissue samples were taken, 13 were male, and 5 were female. In addition, 12 were smokers, while three were not. For three patients, this information was not available (15). According to the results of the experimental (post-hoc) power analysis and the findings obtained from the study, the achieved power for the analyses performed by taking SCLC tissue samples and adjacent tissue samples from 18 patients was calculated to be nearly 100%.

3.2. Feature Selection

Choosing which variables to include in a model is a crucial part of any predictive modeling process, and data selection is an integral part of any statistical modeling process. Determining the most valuable elements of the dataset to be utilized in the study before dealing with massive datasets and models with high computing costs will lead to significant efficiency in terms of outcomes. Finding which aspects of a dataset affect the dependent variable is the goal of feature selection. There is a risk of overlearning the data and producing biased findings if there are too many explanatory factors and the computation time needed to process them is too great. Moreover, it is challenging to understand models that contain a large number of variables. Important influencing factors should be chosen before statistical modeling (16). Large datasets can overwhelm the effectiveness of most ML and data mining techniques, leading to poor outcomes. As a result, reducing the dimensionality yields better outcomes using these approaches (17).

Gene expression datasets are massive. Modeling analyses require a long time due to massive gene expression datasets, and these datasets may lead to computational inefficiencies in the studies performed. Because of the high dimensionality, the model's performance may suffer. A classification method may overfit the training instances and undergeneralize novel samples in gene expression datasets with many genes. In this study, Lasso, a feature selection technique, was utilized to overcome these challenges. The Lasso approach demands that the sum of the absolute values of the model parameters be smaller than a specified value (upper limit). The approach accomplishes this by penalizing the regression variable coefficients, leading some of them to decrease to zero, and is particularly useful when the dataset contains a large number of variables but few observations. Lasso also enhances model interpretability and removes the problem of over-learning by deleting extraneous variables that are unrelated to the response variable (18).

3.3. XGBoost

In ML, Gradient Boost is a potent tool for regression and classification issues where ensemble versions of decision trees are typically the result of poor predictive models. The boosting-based Gradient Boost technique aims to build numerous sequentially weak learners and merge them into an elaborate model (19).

One of the most powerful supervised learning techniques is gradient boosting machines, and one of its applications is XGBoost. It is based on gradient boosting and decision tree algorithms, which form its basic structure. Its speed and efficiency are far beyond those of competing algorithms. In addition to its strong predictiveness, XGBoost is 10 times quicker than competing algorithms and has many regularizations that boost the overall performance while mitigating overfitting or overlearning. To produce a robust classifier, gradient boosting uses a collection of weak classifiers and the boosting technique to combine them. The powerful learner is educated in an iterative process, commencing with a primary learner. XGBoost works on the same fundamentals as gradient boosting. The main distinction lies in the specifics of their use. It is possible to improve the performance of XGBoost by

using a variety of regularization approaches to the trees' complexity (20, 21).

3.4. Bioinformatics Analysis

Gene expression patterns were analyzed for samples of SCLC tissue and normal lung tissue using differential expression analysis performed via the Limma package of the R programming language (22). Differential expression analysis is the statistical examination of normalized read count data to discover quantifiable variations in expression activity between treatment arms. A pipeline was built for the critical analysis using the R software environment. The output includes a table describing the relative importance of the genes and a graph showing the genes with various expression levels. The most reliable genes are those with lower P-values in the table of results, which also includes corrected P and log2-fold change (log2FC) values. Genes with a log2FC of >1 were considered up-regulated, whereas those with a log2FC of -1 were considered down-regulated (23). We used a volcano plot to visually emphasize readily noticeable high values concerning the key genes.

3.5. Biostatistical Analysis

The normal distribution of values was determined by the Shapiro-Wilk test. The independent samples t-test was employed to compare the output variable and input variables, which consisted of normal lung tissue and small cell cancer tissue categories. Statistical significance was assumed at a P-value of less than 0.05. The study was conducted using IBM's SPSS Statistics (version 25.0).

3.6. Modeling Process

One of the ML techniques utilized in the modeling was XGBoost. In the modeling, the dataset was used as 70:30 by dividing the training and test datasets. The n-fold cross-validation strategy was used for the analyses. The n-fold cross-validation technique divides the data into n subsets and then applies the model to each subset. The n-part dataset is divided as follows: one part is utilized for testing, while the remaining n-1 parts are used for model training. The cross-validation approach is assessed by looking at the median of the results. The modeling in this research used five-fold cross-validation. The employed performance metrics were accuracy (AC), balanced accuracy (BAC), sensitivity (Sens), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), and F1 scores. Moreover, variable importance scores were determined, which revealed how much each input variable contributed to the overall explanation of the outcome.

4. Results

A total of 36 tissues were obtained from 18 patients in the study. Eighteen of the tissues were SCLC tissues, and 18 were adjacent normal tissues. The mean age of the patients was 56.5±9.85 years. The dataset comprised 20,425 expressions. According to the findings of the bioinformatics analysis, Table 1 contains a summary of the first 10 results concerning the minimum adjusted P-values. Based on the statistics from Table 1, all genes were up-regulated. According to Table 1, Log2FC values for the UBE2T, NUF2, EXO1, HEPACAM2, ZWINT, ORC6, GINS1, TPX2, TOP2A, and TTK genes were 3.69, 4.22, 3.15, 5.99, 3.21, 3.17, 3.46, 3.90, 4.30, and 4.03, respectively.

Table 1. Results of the bioinformatics analysis

ID of Gene	Adjusted P-Value*	P-Value*	t	B	Log2FC	Gene Name	Differential Expression
29089_at	<0.001	<0.001	13.086	24.823	3.69	UBE2T	UP
83540_at	<0.001	<0.001	12.589	23.701	4.22	NUF2	UP
9156_at	<0.001	<0.001	12.390	23.243	3.15	EXO1	UP
253012_at	<0.001	<0.001	12.374	23.205	5.99	HEPACAM2	UP
11130_at	<0.001	<0.001	12.330	23.104	3.21	ZWINT	UP
23594_at	<0.001	<0.001	12.315	23.069	3.17	ORC6	UP
9837_at	<0.001	<0.001	12.257	22.933	3.46	GINS1	UP
22974_at	<0.001	<0.001	12.055	22.458	3.90	TPX2	UP
7153_at	<0.001	<0.001	11.974	22.268	4.30	TOP2A	UP
7272_at	<0.001	<0.001	11.932	22.168	4.03	TTK	UP

*: P<0.001

Figure 1 represents the volcano plot displaying the differentially expressed genes. The volcano graph compares significance against fold-change in log2 based on the y- and x-axes, respectively, to determine rapid genes with significant expression differences.

Eighteen expression results were acquired by implementing the Lasso method on 20,425 expression results. Table 2 depicts descriptive statistics for the chosen genes in terms of the categories. Table 2 shows that statistically significant differences were found across groups for all genes (P<0.05). The results from the XGBoost model performance metrics are detailed in Table 3.

AC, BAC, Sens, Spec, PPV, NPV, and F1 scores from the XGBoost model were 90%, 90%, 80%, 100%, 100%, 83.3%, and 88.9%, respectively. The performance criteria values are plotted for the XGBoost model in Figure 2.

Table 4 shows the variable importance values of the input variables that best explain the output variable as a result of the modeling. Figure 3 shows the graph of variable importance values obtained

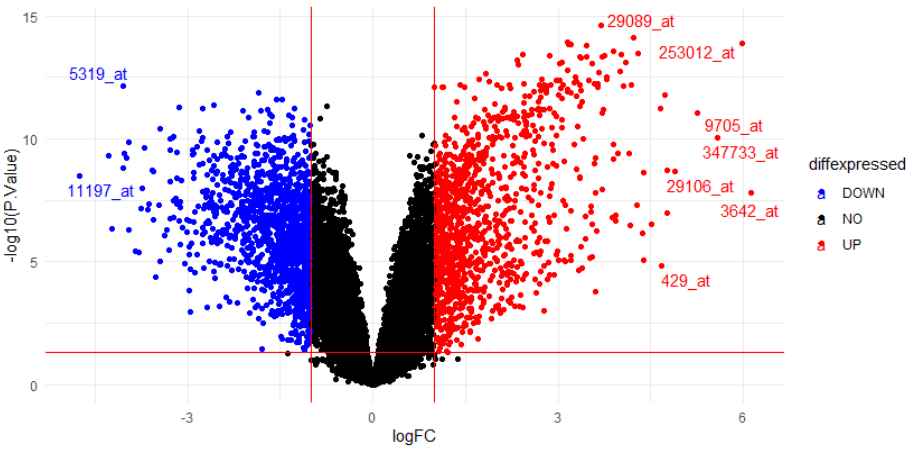

Figure 1. Volcano plot

Table 2. Descriptive statistics for input variables

ID of Gene	Gene Name	Group		P*
		Cancer	Control	
		Mean±SD	Mean±SD	
115749_at	C12orf56	5.583±1.276	3.426±0.138	**<0.001**
171220_at	DSTNP2	5.141±0.238	5.678±0.274	**<0.001**
27299_at	ADAMDEC1	8.96±1.425	5.603±1.27	**<0.001**
3008_at	HIST1H1E	4.703±0.534	3.683±0.219	**<0.001**
3148_at	HMGB2	12.268±0.657	10.38±0.365	**<0.001**
400360_at	C15orf54	2.688±0.16	2.855±0.136	**0.002**
5319_at	PIK3C3	4.823±1.486	8.863±0.701	**<0.001**
5372_at	PMM1	7.059±0.434	7.931±0.362	**<0.001**
5454_at	POU3F2	6.296±1.636	3.583±0.192	**<0.001**
55282_at	LRRC36	6.16±0.478	8.004±0.592	**<0.001**
55766_at	H2AFJ	5.742±0.604	6.671±0.163	**<0.001**
55775_at	TDP1	7.096±0.504	5.912±0.25	**<0.001**
7126_at	TNFAIP1	7.981±0.242	8.588±0.206	**<0.001**
7161_at	TP73	4.837±0.352	4.249±0.179	**<0.001**
8354_at	HIST1H3I	3.488±0.446	2.835±0.149	**<0.001**
90141_at	EFCAB11	6.116±0.222	5.751±0.172	**<0.001**
9108_at	MTMR7	3.978±0.859	2.776±0.182	**<0.001**
97_at	ACYP1	10.118±1.313	7.523±0.184	**<0.001**

SD: Standard deviation, *: Independent samples t-test

Table 3. Performance metrics of the XGBoost model

Performance Metrics	Value (%)
Accuracy	90.0
Balanced Accuracy	90.0
Sensitivity	80.0
Specificity	100
Positive predictive value	100
Negative predictive value	83.3
F1-score	88.9

because of the model. The HIST1H1E gene had the highest predictor importance of 100%, followed by C12orf56 at 99.64%, DSTNP2 at 42.24%, ADAMDEC1 at 36.10, and HMGB2 at 19.43%.

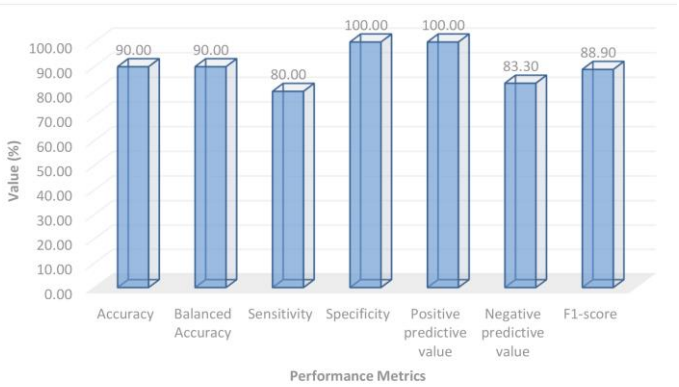

Figure 2. Graph of values for performance criteria obtained from XGBoost models

Table 4. Variable importance values related to the XGBoost model

Gene Name	Importance Values (%)
HIST1H1E	100
C12orf56	99.649
DSTNP2	42.247
ADAMDEC1	36.106
HMGB2	19.438

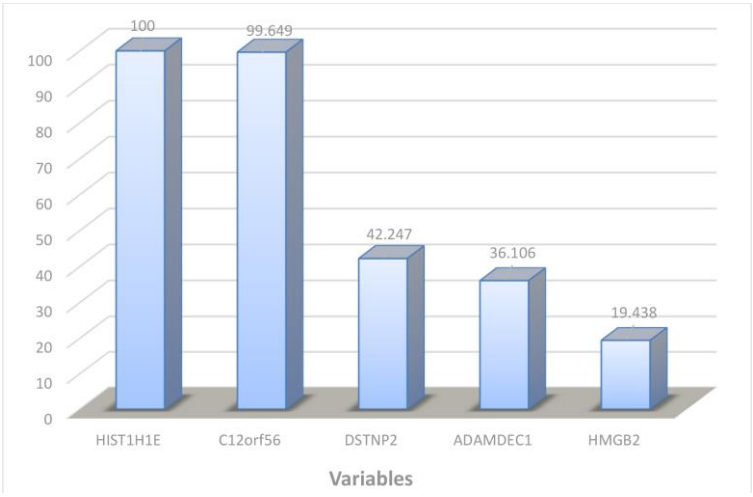

Figure 3. Graphic of gene importance values for predicting the output variable

5. Discussion

SCLC accounts for 15% of new lung cancer diagnoses. It is a particularly aggressive cancer with rapid development and early hematogenous dissemination. At the time of diagnosis, approximately one-third of patients had limited-stage disease, which can be treated with chemoradiation, whereas the remaining patients had an extensive-stage illness, which is normally treated with palliative chemotherapy. Although most patients show an initial response to chemotherapy and/or radiotherapy, virtually all patients relapse with resistant cancer, and the five-year overall survival rate is 5%-10% (24-27).

Comprehensive genome-wide profiling has significantly increased our understanding of the genomic landscapes of numerous cancer types over the last decade, leading to the identification of novel predictive/prognostic biomarkers and therapeutic targets (28, 29). However, in comparison to many other solid tumors, only a few studies have been conducted to investigate the genetic landscape of SCLC (30, 31). Genomic studies are inconclusive due to the lack of suitable tumor tissues, as most SCLC patients are not treated with surgical resection. Therefore, there is a need to determine the genomic structure of SCLS, reveal genomic profiles related to the disease, and explain the underlying genomic mechanisms (24). For this reason, differential

expression analysis was performed in the current study to identify genes that can be potential biomarkers for SCLS by comparing SCLS tissue to normal tissue. Afterward, XGBoost, one of the ML methods, was used to determine the most important genes associated with SCLS. In the dataset examined in the current study, genomic data from samples acquired from the lungs of 18 patients with SCLC tissue and normal lung tissue (adjacent tissue) were used for related analyses. The samples were obtained by surgical resection, and the microarray method made the gene expression profile. According to the Log2FC values used to define the expression fold changes between the two groups from the findings of the bioinformatics analysis (detailed in Table 2), the UBE2T gene has 12.90-fold higher gene expression in SCLC tissue than in normal lung tissue. Similarly, NUF2 gene has 18.63-fold, EXO1 gene has 8.87-fold, HEPACAM2 gene has 63.55-fold, ZWINT gene has 9.25-fold, ORC6 gene has 9.00-fold, GINS1 gene has 11.00-fold, TPX2 gene has 14.92-fold, TOP2A gene has 19.69-fold, and TTK gene has 16.33-fold higher gene expression. Because of their sheer quantity, gene expression data provide unique challenges for modeling. Therefore, the most crucial genes linked with the output variable were chosen using the Lasso variable selection approach before modeling using the current dataset. To construct XGBoost, 18 genes were chosen using the Elastic Net technique. The AC, BAC, Sens, Spec, PPV, NPV, and F1 scores from the XGBoost model were 90%, 90%, 80%, 100%, 100%, 83.3%, and 88.9%, respectively. The performance metrics indicated that the proposed XGBoost could correctly classify the two groups of tissue. According to the variable importance obtained from the XGBoost method, HIST1H1E, C12orf56, DSTNP2, ADAMDEC1, and HMGB2 genes can be used as potential predictive biomarkers for SCLC. The statistical analysis revealed that 18 genes acquired through variable selection exhibited statistically significant variations between the two groups. The proposed bioinformatic model for the detection of expression profiles in the current small sample of lung cancer data revealed that the upregulation of HIST1H1E mRNA is the highest predictor of cancer in humans. HIST1H1E is a gene that encodes histone H1.4, one of 11 H1 linker histones. HIST1H1E protein has a role in the formation of higher chromatin structures and the accessibility of proteins, which are related to chromatin remodeling or histone modifications (32). HIST1H1E mutations are mainly related to Rahman Syndrome, but the processes it involves have been implicated in cancer pathogenesis. Lee et al. reported that HIST1H1E expression levels were low in endometrial cancer cell lines compared to immortalized endometrium epithelial cells and were upregulated in response to calcitriol treatment in cancerous cells, suggesting possible antitumor activity (33). On the other hand, the bioinformatic data suggested conflicting reports depending on the type of cancer. For instance, Kumar et al. (also reviewed in Chang et al.) (34) suggested that HIST1H1E acted as an oncogene in diffuse large-B cell lymphoma, while it was proposed as a tumor suppressor gene in liver hepatocellular carcinoma (35). In another report based on the TCGA data, HIST1H1E was overexpressed in esophageal cancer and associated with a poor prognosis (36). Here we analyzed differentially expressed genes in an SCLC cohort compared to adjacent normal tissues and established HIST1H1E expression level as a reliable biomarker using our ML method, XGBoost.

A previous study showed that the C12orf56 gene was differentially expressed in lung squamous cell carcinoma, which is one of the lung cancer types (37). It was determined that the C12orf56 gene was associated with cancer and showed differential expression in cancer by the upregulation of this gene in the case of ovarian cancer (38). It is also known that the ADAMDEC1 gene plays an important role in the pathogenesis of many known diseases, including cancer. In one study, in bioinformatic analyses performed to determine the functions of the ADAMDEC1 gene in NSCLC, it was found that this gene was up-regulated in NSCL and associated with poor prognosis in the disease through the PI3K/AKT pathway (39). Another study found that HMGB2 may be a biomarker that reflects the disease characteristics and prognosis of NSCLC and is useful for improving clinical efficacy in the case of NSCLC (40).

The present study has several limitations. The first is the inability to obtain the necessary demographic and clinical data of the patients. This is a limitation for studies using such datasets. In addition, since the patients did not have clinical information, the integration of the obtained genes with other information was not possible.

6. Conclusion

Using gene expression data from SCLC tissue and normal lung tissue (adjacent tissue), this study discovered possible genetic biomarkers for SCLC. Future, more in-depth investigations will evaluate the accuracy of these genes, permitting the development of targeted therapies and elucidating their clinical relevance. The discovery and application of these gene-based therapeutic approaches for this disease, which is determined according to the RCRA and is in the category of intractable cancer, will be very valuable and will shed light on reducing mortality and revealing the preclinic stages of the disease. Therefore, target gene-directed therapies can be set out for SCLC, which does not have an approved treatment due to the lack of clear genomic targets, and can be

obtained with a good response from the treatment, as in NSCLS.

Acknowledgments

We would like to thank Prof. Dr. Cemil Çolak for the support.

Footnotes

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;**68**(6):394-424. doi: 10.3322/caac.21492. [PubMed: 30207593].
2. Schabath MB, Cote ML. Cancer progress and priorities: lung cancer. *Cancer Epidemiol Biomarkers Prev*. 2019;**28**(10):1563-79. doi: 10.1158/1055-9965.EPI-19-0221. [PubMed: 31575553].
3. Wahbah M, Boroumand N, Castro C, El-Zeky F, Eltorky M. Changing trends in the distribution of the histologic types of lung cancer: a review of 4,439 cases. *Ann Diagn Pathol*. 2007;**11**(2):89-96. doi: 10.1016/j.anndiagpath.2006.04.006. [PubMed: 17349566].
4. Rami-Porta R, Bolejack V, Giroux DJ, Chansky K, Crowley J, Asamura H, et al. The IASLC lung cancer staging project: the new database to inform the eighth edition of the TNM classification of lung cancer. *J Thorac Oncol*. 2014;**9**(11):1618-24. doi: 10.1097/JTO.0000000000000334. [PubMed: 25436796].
5. Tsoukalas N, Aravantinou-Fatorou E, Baxevanos P, Tolia M, Tsapakidis K, Galanopoulos M, et al. Advanced small cell lung cancer (SCLC): new challenges and new expectations. *Ann Transl Med*. 2018;**6**(8):145. doi: 10.21037/atm.2018.03.31. [PubMed: 29862234].
6. Rudin CM, Brambilla E, Faivre-Finn C, Sage J. Small-cell lung cancer. *Nat Rev Dis Primers*. 2021;**7**(1):3. doi: 10.1038/s41572-020-00235-0. [PubMed: 33446664].
7. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*. 2010;**463**(7278):184-90. doi: 10.1038/nature08629. [PubMed: 20016488].
8. Karachaliou N, Pilotto S, Lazzari C, Bria E, de Marinis F, Rosell R. Cellular and molecular biology of small cell lung cancer: an overview. *Transl Lung Cancer Res*. 2016;**5**(1):2-15. doi: 10.3978/j.issn.2218-6751.2016.01.02. [PubMed: 26958489].
9. Shtivelman E, Hensing T, Simon GR, Dennis PA, Otterson GA, Bueno R, et al. Molecular pathways and therapeutic targets in lung cancer. *Oncotarget*. 2014;**5**(6):1392-433. doi: 10.18632/oncotarget.1891. [PubMed: 24722523].
10. Byers LA, Rudin CM. Small cell lung cancer: where do we go from here? *Cancer*. 2015;**121**(5):664-72. doi: 10.1002/cncr.29098. [PubMed: 25336398].
11. Kalemkerian GP. Advances in pharmacotherapy of small cell lung cancer. *Expert Opin Pharmacother*. 2014;**15**(16):2385-96. doi: 10.1517/14656566.2014.957180. [PubMed: 25255939].
12. Drapkin BJ, Rudin CM. Advances in small-cell lung cancer (SCLC) translational research. *Cold Spring Harb Perspect Med*. 2021;**11**(4):a038240. doi: 10.1101/cshperspect.a038240. [PubMed: 32513672].
13. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
14. Akman M, Genç Y, Ankarali H. Random Forests Yöntemi ve Saglik Alaninda Bir Uygulama/Random forests methods and an application in health science. *Turk Klin Biyoistatistik*. 2011;**3**(1):36-48.
15. Cai L, Liu H, Huang F, Fujimoto J, Girard L, Chen J, et al. Cell-autonomous immune gene expression is repressed in pulmonary neuroendocrine cells and small cell lung cancer. *Commun Biol*. 2021;**4**(1):314. doi: 10.1038/s42003-021-01842-7. [PubMed: 33750914].
16. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;**23**(19):2507-17. doi: 10.1093/bioinformatics/btm344. [PubMed: 17720704].
17. Fodor IK. A survey of dimension reduction techniques. Lawrence Livermore National; 2002.
18. Fonti V. Research Paper in Business Analytics: Feature Selection with LASSO. Amsterdam: VU Amsterdam; 2017.
19. Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. *Applied Sci*. 2018;**8**(5):689. doi: 10.3390/app8050689.
20. Dikker J. Boosted tree learning for balanced item recommendation in online retail. Eindhoven University of Technology; 2017.
21. Salam Patrous Z. Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors. KTH Royal Institute of Technology; 2018.
22. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: Springer; 2005. p. 397-420.
23. Yan H, Zheng G, Qu J, Liu Y, Huang X, Zhang E, et al. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. *J Cell Physiol*. 2019;**234**(12):23785-97. doi: 10.1002/jcp.28947. [PubMed: 31215027].
24. Nong J, Gong Y, Guan Y, Yi X, Yi Y, Chang L, et al. Circulating tumor DNA analysis depicts subclonal architecture and genomic evolution of small cell lung cancer. *Nat Commun*. 2018;**9**(1):1-8. doi: 10.1038/s41467-018-05327-w
25. Murray N, Coy P, Pater JL, Hodson I, Arnold A, Zee B, et al. Importance of timing for thoracic irradiation in the combined modality treatment of limited-stage small-cell lung cancer. The national cancer institute of canada clinical trials group. *J Clin Oncol*. 1993;**11**(2):336-44. doi: 10.1200/JCO.1993.11.2.336. [PubMed: 8381164].
26. Johnson BE, Grayson J, Makuch RW, Linnoila RI, Anderson MJ, Cohen MH, et al. Ten-year survival of patients with small-cell lung cancer treated with combination chemotherapy with or without irradiation. *J Clin Oncol*. 1990;**8**(3):396-401. doi: 10.1200/JCO.1990.8.3.396. [PubMed: 2155310].
27. Lassen U, Osterlind K, Hansen M, Dombernowsky P, Bergman B, Hansen HH. Long-term survival in small-cell lung cancer: posttreatment characteristics in patients surviving 5 to 18+ years--an analysis of 1,714 consecutive patients. *J Clin Oncol*. 1995;**13**(5):1215-20. doi: 10.1200/JCO.1995.13.5.1215. [PubMed: 7738624].
28. CGAR. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;**511**(7511):543. doi: 10.1038/nature13385. [PubMed: 25079552].
29. CGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;**487**(7407):330-7. doi: 10.1038/nature11252. [PubMed: 22810696].
30. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. *Nature*. 2015;**524**(7563):47-53. doi: 10.1038/nature14664.
31. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet*. 2012;**44**(10):1111-6. doi: 10.1038/ng.2405. [PubMed: 22941189].
32. Tatton-Brown K, Loveday C, Yost S, Clarke M, Ramsay E, Zachariou A, et al. Mutations in epigenetic regulation genes are a major cause of overgrowth with intellectual disability. *Am J Hum Genet*. 2017;**100**(5):725-36. doi: 10.1016/j.ajhg.2017.03.010. [PubMed: 28475857].
33. Lee LR, Teng PN, Nguyen H, Hood BL, Kavandi L, Wang G, et al. Progesterone enhances calcitriol antitumor activity by upregulating vitamin D receptor expression and promoting apoptosis in endometrial cancer cells. *Cancer Prev Res (Phila)*.

2013;**6**(7):731-43. doi: 10.1158/1940-6207.CAPR-12-0493. [PubMed: 23682076].
34. Chang S, Yim S, Park H. The cancer driver genes IDH1/2, JARID1C/ KDM5C, and UTX/ KDM6A: crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism. *Exp Mol Med*. 2019;**51**(6):1-17. doi: 10.1038/s12276-019-0230-6. [PubMed: 31221981].
35. Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics*. 2015;**31**(22):3561-8. doi: 10.1093/bioinformatics/btv430. [PubMed: 26209800].
36. Dai J, Reyimu A, Sun A, Duoji Z, Zhou W, Liang S, et al. Establishment of prognostic risk model and drug sensitivity based on prognostic related genes of esophageal cancer. *Sci Rep*. 2022;**12**(1):8008. doi: 10.1038/s41598-022-11760-1. [PubMed: 35568702].
37. Zhang F, Chen X, Wei K, Liu D, Xu X, Zhang X, et al. Identification of key transcription factors associated with lung squamous cell carcinoma. *Med Sci Monit.* 2017;**23**:172-206. doi: 10.12659/msm.898297. [PubMed: 28081052].
38. Yang X, Zhu S, Li L, Zhang L, Xian S, Wang Y, et al. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. *Onco Targets Ther*. 2018;**11**:1457-74. doi: 10.2147/OTT.S152238. [PubMed: 29588600].
39. Zhu W, Shi L, Gong Y, Zhuo L, Wang S, Chen S, et al. Upregulation of ADAMDEC1 correlates with tumor progression and predicts poor prognosis in non-small cell lung cancer (NSCLC) via the PI3K/AKT pathway. *Thorac Cancer*. 2022;**13**(7):1027-39. doi: 10.1111/1759-7714.14354. [PubMed: 35178875].
40. Lou N, Zhu T, Qin D, Tian J, Liu J. High-mobility group box 2 reflects exacerbated disease characteristics and poor prognosis in non-small cell lung cancer patients. *Ir J Med Sci*. 2022;**191**(1):155-62. doi: 10.1007/s11845-021-02549-8. [PubMed: 33635447].