Research Article

ISSN 2771-9057

Recent Advances in Clinical Trials

Using Bioinformatics and Machine Learning Techniques to Identify Potential Genes that may be Associated with Lung Cancer and Facilitate the Screening of Such Genes

Zeypep Kucukakcali* and Ipek Balikci Cicek

Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, 44280, Malatya, Turkey.

***Correspondence:**

Zeynep Kucukakcali, Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, 44280, Malatya, Turkey.

Received: 20 Mar 2024; **Accepted:** 26 Apr 2024; **Published:** 03 May 2024

Citation: Kucukakcali Z, Balikci Cicek I. Using Bioinformatics and Machine Learning Techniques to Identify Potential Genes that may be Associated with Lung Cancer and Facilitate the Screening of Such Genes. Recent Adv Clin Trials. 2024; 4(2); 1-8.

ABSTRACT

Aim: *Lung cancer, the most frequently diagnosed cancer globally, is the leading cause of cancer-related deaths. Due to its increasing prevalence and low survival rates, new biomarkers are needed to diagnose the disease. Therefore, this study aims to identify potential genes that may be associated with lung cancer by bioinformatics methods using gene expression data of lung cancer and non-tumour tissues, and to classify the data with stochasting gradient boosting (SGB), one of the machine learning models, and to determine the genes that may be most associated with the disease with variable significance values obtained at the end of the model.*

Methods: *The data underwent bioinformatics analyses utilizing the limma package within the R programming language. During the modeling phase, the SGB model was utilized for classification purposes. The evaluation of classification performance was conducted by various measures, including accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score. Following the process of modeling, the variable importance values were utilized to ascertain the influential genes in relation to the target variable.*

Results: *Based on the outcomes of bioinformatic analysis, a total of 7098 expressions exhibited statistically significant variations in gene expression levels between the two groups. The performance metrics derived from the SGB model were accuracy (93.5%), balanced accuracy (94.1%), sensitivity (88.2%), specificity (100%), positive predictive value (100%), negative predictive value (87.5%), and F1-score (93.8%). Based on the findings pertaining to variable importance, it was determined that the AGTR1, TNXB///TNXA, and SPP1 genes exhibited significant efficacy in the process of tumorigenesis.*

Conclusion: *Lung cancer-associated genes have been identified through the utilization of bioinformatics and machine learning models. Through conducting thorough research on the discovered genes, it is possible to confirm the correctness of their association with the disease and subsequently design target-based treatment options for these genes.*

Keywords

Lung cancer, Tumor, Non-tumor, Machine learning, Biomarkers.

Introduction

Lung cancer, namely bronchogenic malignant tumors originating from airway epithelioma, is the most often diagnosed cancer globally and the leading cause of cancer-related mortality.

Annually, a global estimation of 1.8 million newly diagnosed cases of lung cancer is reported. In the year 2012, an estimated 1.6 million individuals succumbed to lung cancer, with projections indicating a potential rise to 3 million lung cancer-related fatalities by the year 2035 [1,2]. The prognosis for lung cancer is generally dismal; depending on the disease's stage at diagnosis, the 5-year survival rate might range from 4% to 17% [3]. Although the likelihood of finding lung cancer has increased due to advancements in non-invasive testing, only 10-15% of new cases are identified at an early stage of the disease [4]. When lung cancer is discovered at an advanced stage, there are few therapeutic choices available to 75% of patients. However, the 5-year survival rate for patients with clinical stage IA cancer according to the TNM (tumor-lymph nodes-metastasis) classification is only about 60%, suggesting that a significant proportion of patients have undetectable metastases at this stage of the disease [5-8]. Chest radiography and sputum cytology, the two diagnostic modalities currently in use, are not sensitive enough to diagnose non-small cell lung carcinoma (NSCLC), and tumor markers like carcinoembryonic antigen (CEA), CYFRA 21-1, neuron-specific enolase (NSE), or squamous cell carcinoma antigen (SCCA) prevent the diagnosis of lung cancer at an early stage [4]. These findings highlight the need for additional targeted, less intrusive biomarkers that might be added to or utilized in place of radiographic techniques to enhance the identification and staging of lung cancer [9].

Lung cancer arises from the last stage of multistage carcinogenesis, which has progressively more pronounced genetic and epigenetic alterations, rather than from the abrupt transformation of bronchia epithelioma [6,10]. Today, mutations and genetic mechanisms associated with the disease are being investigated through genetic-based studies that are specific to lung cancer and can enable the disease to be diagnosed at an early stage. Since the disease is diagnosed at an advanced stage and the survival rate is low, knowledge of the underlying mechanisms of the disease will be of great benefit to clinicians. The recent advancements and extensive adoption of next-generation sequencing (NGS) technologies have facilitated genomic analyses aimed at elucidating the etiology of cancer. These investigations have unveiled associations between diverse malignant tumors and genomic data, thereby enabling the discovery of novel molecular markers and intracellular pathways implicated in disease progression. Given these advancements, these methods have been widely employed to elucidate the complete genetic structure of lung cancer [7].

Machine learning (ML) is a specialized domain within the science of artificial intelligence that seeks to provide predictions regarding novel observations through the process of learning from pre-existing data, in contrast to conventional statistical methodologies. ML plays a crucial role in various health-related domains, serving as the fundamental framework for detecting genetic illnesses, facilitating early diagnosis of cancer, and identifying patterns in medical imaging. Over the past decade, the field of machine learning has witnessed significant advancements in performance, owing to the increased accessibility of extensive datasets and enhanced computational capabilities. This has resulted in notable achievements across several domains and scenarios [11,12]. In this study, the stochasting gradient boosting (SGB) method will be applied to transcript data obtained from tissues from patients with lung cancer tumour cells and non-tumour tissues in order to take advantage of the superior performance achievements of ML techniques.

The aim of this study is to identify potential genes that may be associated with lung cancer by bioinformatics methods using open-access gene expression data obtained from human tumor tissues and non-tumor tissues. Our second aim is to classify the disease with SGB, which is one of the ML models, using the data set to be obtained by determining the genes that show different regulation in diseased tissues compared to the non tumor group. Finally, we will determine the most important genes that may be associated with the target variable lung cancer through the SGB model.

Material and Methods

Dataset

The dataset used in the study is an open-access dataset created by taking tumor tissues and non-tumor tissues of patients with lung cancer. The dataset included in the research was acquired from the National Center for Biotechnology Information (NCBI). The data utilized in this study were acquired from the Gene Expression Omnibus (GEO) database, specifically identified by the accession code "GSE10072".

Bioinformatics and Gene Expression Analysis

Bioinformatics refers to the comprehensive process of gathering, retaining, arranging, preserving, scrutinizing, and presenting findings derived from the application of theoretical and practical principles within fields such as biology, medicine, behavioral sciences, and health sciences. Moreover, the primary focus of this endeavor lies in the investigation and advancement of computational tools and methodologies, with the aim of expanding the utilization and manipulation of data acquired through research endeavors or the implementation of established protocols. Acquired via the process of scholarly investigation or the use of established procedures. Bioinformatic analyses are conducted by choosing a suitable database and employing a tool that facilitates the execution of bioinformatic analysis, aligning with the specific biological query, molecule, or structure under investigation. The collected data and generated findings from the analyses are consolidated, and the subsequent assessments are critically examined in the context of the existing literatüre [13].

Alterations in the physiological state of an organism or cell are invariably followed by corresponding modifications in the gene expression profile. Consequently, the measurement of gene expression assumes significant importance across all domains of biological research. The DNA microarray technique, which is currently under development, is employed for the investigation of gene expression. This is achieved via the process of hybridization, where mRNA molecules are bound to a densely populated array of immobilized target sequences. Each of these target sequences corresponds to a distinct gene. The impact of chemical substances on the regulation of gene expression can provide insights into both functional and toxicological attributes. Investigations conducted on clinical samples, encompassing both those in good health and those afflicted with illness, have the potential to unveil previously undiscovered biomarkers [14].

Bioinformatics Analysis Phase

In this study, gene expression analyses were performed on trancriptomic data obtained from tumor samples and non-tumor samples. The inquiry utilized the limma package, a software tool available in the R programming language that enables expression analysis [15]. Limma, also known as Linear Models for Microarray Analysis, is a software package designed to assess gene expression microarray data. Its primary objective is to employ linear models to analyze specific experiments and identify differential expression. The functions of the packet may be applied to several gene expression methods, including microarrays, RNA-seq, and quantitative PCR. The Limma software offers the capability to yield consistent outcomes, even in scenarios with a limited number of sequences, owing to the utilization of Empirical Bayes methodologies. The bioinformatic study yielded Log2FC, a metric that quantifies the fold change in gene expression differences. This metric ranks the genes in descending order of significance. Genes that are up-regulated are identified by using a threshold of log2 fold change (log2FC) greater than 1, whereas genes that are down-regulated are identified by applying a threshold of log2FC less than -1.

The distribution of the data utilized in the study was shown through the utilization of box plot graphs and expression density graphs. The graphs depict samples possessing similar qualities, which are shown by the utilization of consistent colors. The researchers choose to utilize the Uniform Manifold Approximation and Projection (UMAP) graph as a means to visually represent the interrelationships among the samples under investigation. The utilization of the volcano plot was deemed preferable for the purpose of illustrating genes that are differently expressed, both in terms of upregulation and downregulation. The volcano plot illustrates the relationship between significance and fold-change, shown on the y- and x-axes respectively, in a logarithmic base of 2. This graphical representation facilitates the rapid identification of genes that exhibit differential expression. The graph displays gene expression levels, with the color red representing up-regulated genes, blue representing down-regulated genes, and black representing genes that exhibit no significant difference in expression.

Feature Selection Phase

The process of variable selection has significant importance in predictive modeling procedures, and a crucial aspect of constructing a statistical model involves making informed decisions on the inclusion of data throughout the modeling phase. Prior to engaging with large datasets and computationally expensive models, it is imperative to undertake a process of identifying the most significant elements within the dataset. This endeavor is crucial in order to optimize the efficiency and effectiveness of the study's outcomes. Feature selection is a process that aims to identify the most influential characteristics that have an impact on the dependent variable of a given data collection. The presence of an excessive number of explanatory factors might result in prolonged calculation durations and the potential for overfitting the data, so yielding biased outcomes. Furthermore, the interpretation of models constructed with a multitude of variables is a significant challenge. Prior to engaging in statistical modeling, it is advisable to carefully choose the pertinent factors that exert an influence on the dependent variable [16]. Many machine learning and data mining techniques may yield suboptimal outcomes when applied to large datasets. Hence, these methodologies yield more efficient outcomes as the dimensionality is decreased [17].

Gene expression datasets are of considerable size. The computational efficiency of modeling analyses is sometimes hindered by the high size of gene expression datasets, hence resulting in prolonged study durations. The model's performance may be negatively impacted due to the issue of excessive dimensionality. If gene expression datasets contain an excessive number of genes, a classification method has the potential to overfit the training examples and fail to generalize well to new samples. In this work, the feature selection approach known as Lasso was employed to address the aforementioned issues. The LASSO method requires that the sum of the model parameters' absolute values be less than a fixed value (upper limit). This approach is accomplished by imposing penalties on the regression coefficients of the variables, resulting in the elimination of some coefficients, reducing them to zero. The presence of several variables and few observations in a dataset renders this particular circumstance particularly advantageous. Moreover, by the elimination of extraneous variables that bear no relevance to the response variable, the LASSO technique enhances the interpretability of the model and effectively mitigates the issue of overfitting [18].

Stochastic Gradient Boosting

Fridman developed SGB by including randomization into the gradient boosting strategy. Using the permutation sampling methodology, a sub-sample is randomly picked in each iteration of this process. Instead of all students, this sub-sample is utilized to compute the current state of the model, minimizing the correlation between the constructed trees [19]. Unlike previous ensemble learning approaches, this method summarizes each tree (about 100 to 200 trees) formed as the process runs, rather of constructing massive massive trees, and each observation is classified based on the most prevalent categorization across trees . The SGB model is distinguished from other augmentation strategies by this type of separation. Furthermore, this discriminating approach is less sensitive to outliers and imbalanced datasets. This approach is 5 times quicker than other existing algorithms and has a far higher predictive power. Another key element of the model is the inclusion of a set of regularization algorithms that can increase overall performance and decrease over-fitting and over-learning [19,20].

Modeling Phase

Before modeling, variable selection was made using the Lasso variable selection method. To model with SGB, the data set is divided into 70% training data and 30% test data.

In this study, the n-fold cross-validation technique, which is a type of resampling approach, was employed to guarantee the validity of the model. Specifically, the n-fold cross-validation approach involves initially dividing the dataset into n subsets, followed by the application of the model to each of these subsets. In the subsequent stage, a single component out of the total n components is allocated for testing purposes, while the remaining n-1 components are utilized for training. In the last stage, the cross-validation technique is assessed by computing the average of the values obtained from the models.

The performance of the modeling was evaluated using several metrics, including accuracy (ACC), balanced accuracy (b-ACC), sensitivity (SE), specificity (SP), positive predictive value (ppv), negative predictive value (npv), and F1-score. Finally, as a result of the modeling test, variable importance values were calculated to determine the genes that were most effective on the target variable.

Results

According to the results of biostatistical analysis, the mean age was found to be 66.39±7.91 in the tumor group and 65.59±7.67 in the non-tumor group. There were 23 women and 35 men in the tumor group and 15 women and 34 men in the non-tumor group. In addition, there were 24 current smokers, 18 former smokers, and 16 never smokers in the tumor group, while these numbers were 16, 18, and 15 for current smokers, former smokers, and never smokers, respectively, in the non-tumor. Additionally, no significant relationship was found between the categories of the target variable and the categories of the smoking variable (p=0.644).

Distribution plots of 58 tumor tissues and 49 non-tumor tissues used in the study are given in Figure 1 and Figure 2.

Table 1: Transcripts found to be up-regulated in tumourous tissue samples compared to non-tumourous tissues.

ID	Adj.P Val	P Value	t	B	Log2FC	Gene name
209875_s_at	7,76E-37	5,22E-40	20,99937	80,59894	4,364415	SPP1
37892_at	1,88E-21	3,04E-23	12,72311	42,27889	3,061522	COL11A1
204475_at	5,80E-15	2,47E-16	9,674348	26,46668	2,862036	MMP1
206239_s_at	1,35E-11	9,83E-13	8,079641	18,25491	2,7584	SPINK1
218469_at	6,28E-23	8,05E-25	13,43449	45,88895	2,548363	GREM1
217428_s_at	7,58E-20	1,52E-21	11,96602	38,39089	2,463475	COL10A1
201292_at	2,43E-24	2,41E-26	14,13133	49,37731	2,455963	TOP2A
218468_s_at	2,15E-21	3,51E-23	12,69522	42,13645	2,441458	GREM1
214774_x_at	3,82E-17	1,15E-18	10,69731	31,80171	2,43796	TOX3
201291_s_at	9,97E-22	1,57E-23	12,85174	42,93501	2,422077	TOP2A
204580_at	6,93E-19	1,58E-20	11,51596	36,0616	2,377325	MMP12
212353_at	1,51E-24	1,45E-26	14,23397	49,88675	2,36008	SULF1
202310_s_at	1,18E-17	3,24E-19	10,93855	33,05919	2,328322	COL1A1
216623_x_at	7,82E-18	2,10E-19	11,02106	33,48896	2,311316	TOX3
201884_at	2,31E-11	1,78E-12	7,963586	17,66856	2,285407	CEACAM5

Table 2: Transcripts found to be down-regulated in tumourous tissue samples compared to non-tumourous tissues

ID	Adj.P Val	P Value	t	B	Log2FC	Gene name
210081_at	2,23E-37	7,00E-41	-21,4864	82,58584	-4,41747	AGER
214387_x_at	8,67E-24	9,58E-26	-13,8562	48,00591	-4,04718	SFTPC
203980_at	1,25E-35	1,43E-38	-20,2111	77,32083	-3,83854	FABP4
210096_at	7,41E-24	7,86E-26	-13,8955	48,20255	-3,7098	CYP4B1
204712_at	9,54E-24	1,07E-25	-13,8347	47,89866	-3,68671	WIF1
209613_s_at	1,50E-26	1,09E-28	-15,2246	54,74244	-3,67847	ADH1B
219230_at	1,79E-27	1,15E-29	-15,689	56,97886	-3,57031	TMEM100
205866_at	3,27E-30	1,36E-32	-17,1137	63,67134	-3,5031	FCN3
214135_at	4,15E-30	1,77E-32	-17,0576	63,41307	-3,44641	CLDN18
205200_at	2,59E-37	1,28E-40	-21,34	81,99142	-3,41185	EXOSC7///CLEC3B
209074_s_at	7,20E-41	3,23E-45	-24,014	92,44007	-3,3718	FAM107A
209612_s_at	5,62E-28	3,31E-30	-15,9489	58,2187	-3,3469	ADH1B
204273_at	1,12E-36	8,06E-40	-20,8949	80,16895	-3,16126	EDNRB
213317_at	5,61E-29	2,87E-31	-16,4631	60,64699	-3,08355	CLIC5
217046_s_at	6,22E-37	3,91E-40	-21,0692	80,8856	-3,03769	AGER

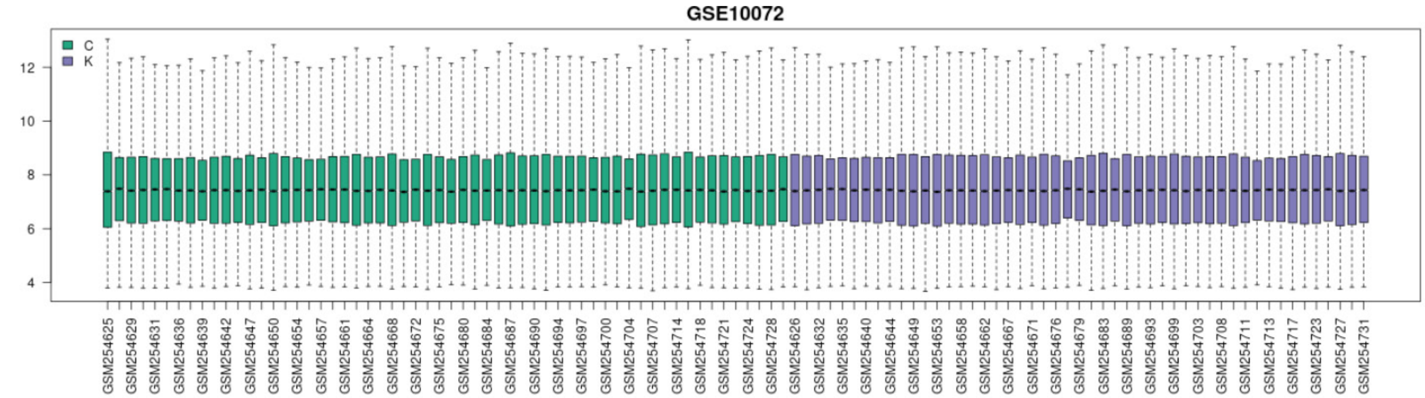

Figure 1: Distribution plot of the samples.

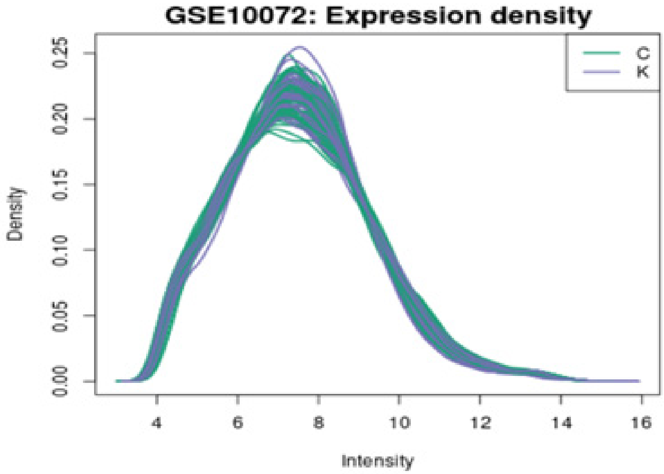

Figure 2: The expression density graph of the samples.

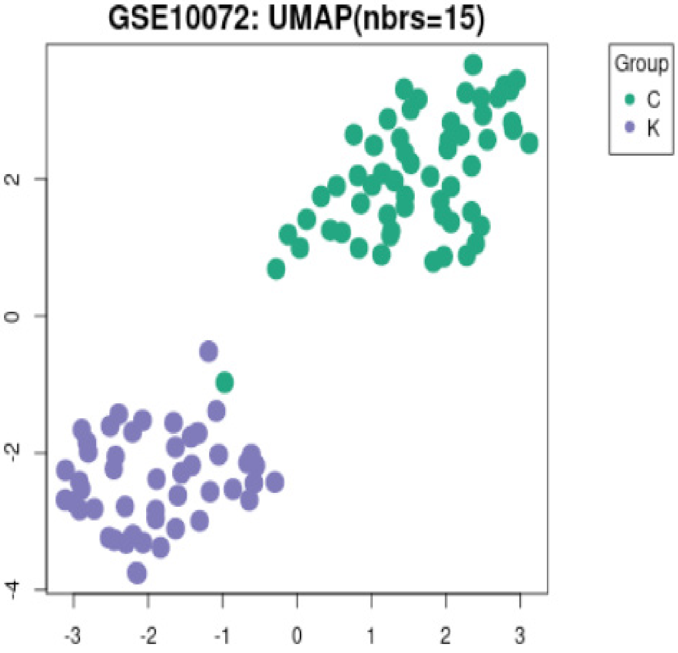

Figure 3: UMAP plot of the samples (Green Dots: Tumor tissues, Purple dots: non-tumor tissues).

Figure 3 presents the UMAP graph, which visually represents the interrelationships among the samples. The graph illustrates that samples with similar features are observed to be grouped together. The graph displays tumor tissues samples represented by green dots, and non-tumor tissues represented by purple dots.

Based on the examination of gene expression, a total of 7098 gene expressions were identified to exhibit statistically significant changes in their expression levels between the two groups (|log2FC| > 1.0, p 0.05). Tables 1 and 2 provide details on the up-regulation and down-regulation of gene expression in the top 15 genes, respectively, observed between the two groups.

Figure 4 illustrates the volcano plot, which serves as a visual representation of the genes that exhibit differential expression between the several groups.

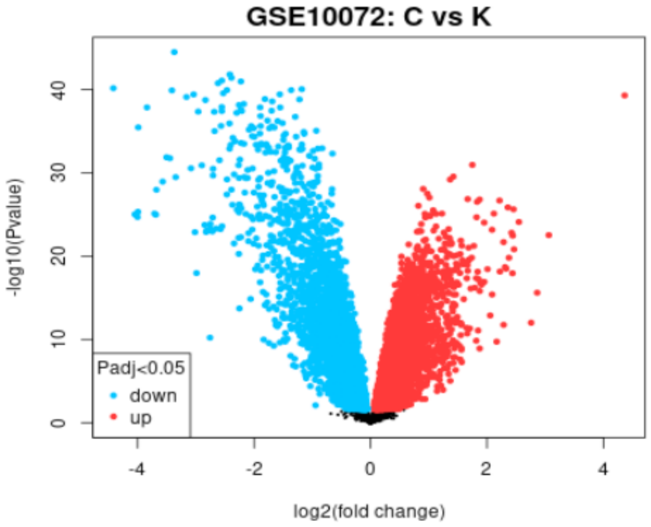

Figure 4: Volcano plot for transcripts in tumor and non-tumor tissues (Red dots indicate transcripts that increased, blue dots decreased, and black dots showed transcripts whose expression level did not change).

LASSO feature selection method was applied to 22283 transcripts together with the target variable (tumor and non-tumor tissues), and as a result, eighteen genes that were determined to explain the target variable the most were selected.

Modelling results using 18 genes selected by the SGB method are given in Table 3.

Table 3: Values of performance metrics of the SGB model.

Metric	Value (%)
Accuracy	93.5
Balanced Accuracy	94.1
Sensitivity	88.2
Specificity	100
Positive predictive value	100
Negative predictive value	87.5
F1 score	93.8

The following performance metrics were calculated for the SGB model: accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1 score. These metrics were 93.5%, 94.1%, 88.2%, 100%, 100%, 87.5% and 93.8% respectively.

The graph of the variable importance of the genes used in the modelling showing their success in explaining the target variable is given in Figure 5.

According to the ROC analysis performed with the 3 most effective genes on the target variable according to the variable significance values, the AUC value of AGTR1, which is the power to distinguish between tumour and non-tumour tissues, was 0.997, TNXB///TNXA was 0.994 and SPP1 was 0.981. ROC analysis results were shown in Figure 6.

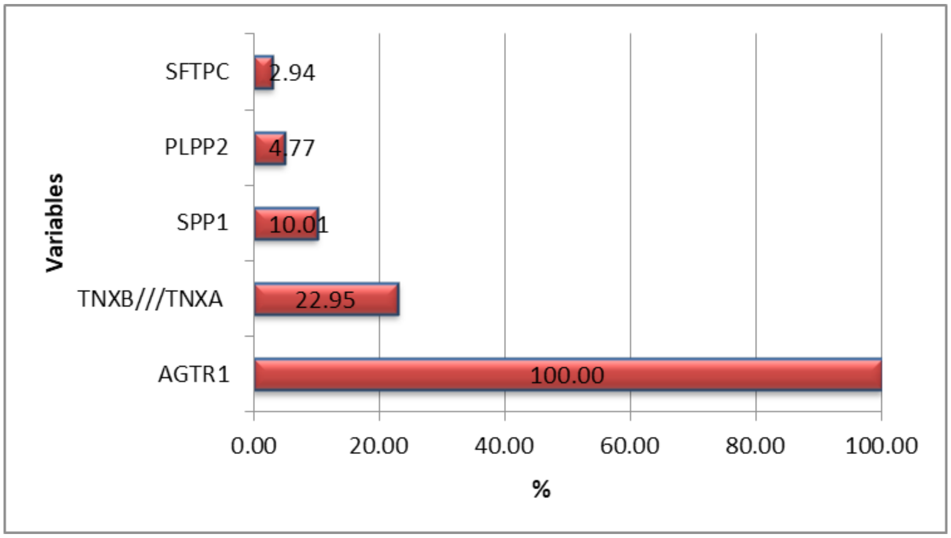

Figure 5: Variable importance graph for the most effective variables in explaining the target variable.

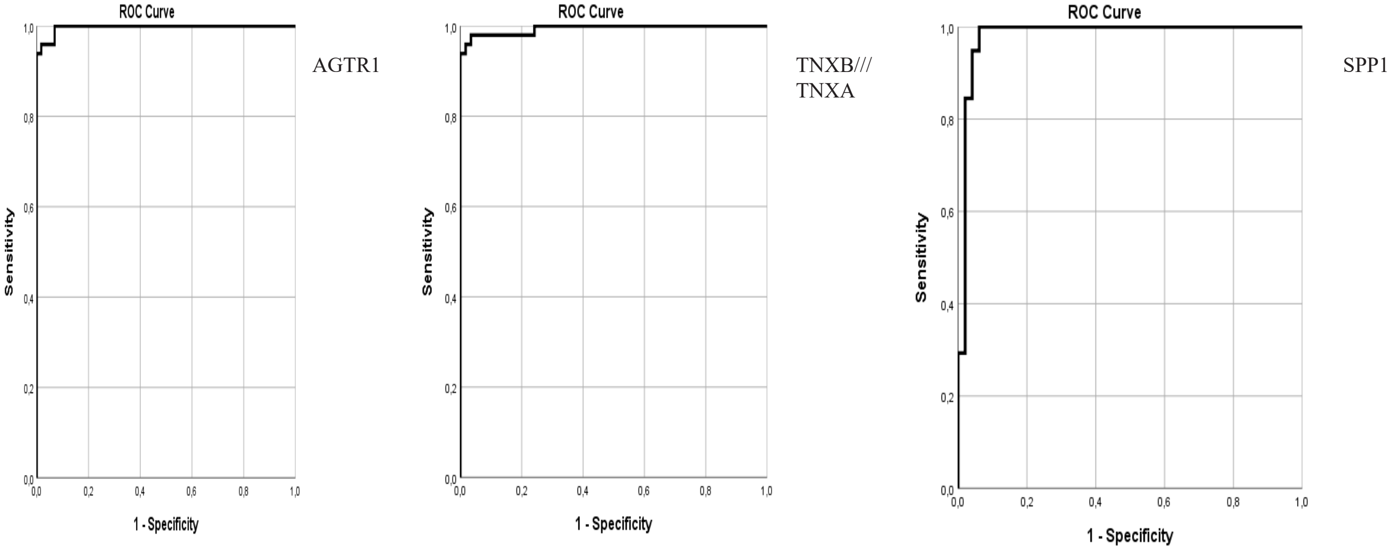

Figure 6: Graph of ROC analysis results.

The variable importance values and the results obtained from ROC analysis support each other and these three genes can be biomarkers.

Discussion

Lung cancer is the most often diagnosed form of cancer globally and represents the leading cause of cancer-related mortality. The prognosis for lung cancer is generally unfavorable, with a significant proportion of patients (about 75%) being diagnosed during the advanced stages of the disease. The diagnostic instruments now in use exhibit insufficient sensitivity and fail to facilitate early-stage disease diagnosis. Hence, the exploration of novel approaches for the timely and precise detection of lung cancer is imperative in order to optimize its therapeutic interventions. Lung cancer is a disease that arises from a complex process known as multistage carcinogenesis, characterized by a progressive accumulation of genetic and epigenetic alterations over time. The utilization of screening methods to detect certain genetic markers has the potential to facilitate the early-stage detection of lung cancer [7]. The progress in molecular techniques and analytical platforms has facilitated the examination of genomic alterations that contribute to the onset of cancer, namely the identification of potential biomarkers associated with lung cancer.

The aim of this study is to identify the genes that are important in the development of lung cancer by using bioinformatic analyses and ML methods, which can produce highly successful results in many fields. For this purpose, bioinformatics analyses were performed using an open access dataset and genes that are differentially expressed in cancerous tissue compared to non-tumour tissue were identified and modelled with SGB, one of the machine learning methods.

When the results of bioinformatic analyses were examined, it was determined that 7098 genes showed different regulation (up or down) in lung tumours compared to non-tumor tissues. SPP1 gene showed 20.53 fold up-regulation in lung tumor tissues compared to normal tissue samples. Likewise, COL11A1, MMP1, SPINK1, GREM1, COL10A1, TOP2A, GREM1, TOX3, TOP2A, MMP12, SULF1, COL1A1, TOX3, and CEACAM5 genes had up-regulated gene expression of 18.33, 7.26, 6.72, 5.81, 5.50, 5.46, 5.42, 5.38, 5.35, 5.16, 5.13, 4.99, 4.95, and 4.85 fold, respectively. AGER gene showed 21.25 fold down-regulation in gastric tumor samples compared to normal tissue samples. Likewise, SFTPC, FABP4, CYP4B1, WIF1, ADH1B, TMEM100, FCN3, CLDN18, EXOSC7///CLEC3B, FAM107A, ADH1B, EDNRB, CLIC5, and AGER genes had down-regulated gene expression of 16.44, 14.22, 12.99, 12.81, 12.72, 11.87, 11.31, 10.85, 10.62, 10.33, 10.12, 8.93, 8.45, and 8.16 fold, respectively.

Before modelling with SGB, 18 genes were selected by the variable selection method and these genes were used in the modelling. As a result of modelling with these genes The model gave various performance indicators such as accuracy (93.5%), balanced accuracy (94.1%), sensitivity (88.2%), specificity (100%), positive predictive value (100%), negative predictive value (87.5%) and F1-score (93.8%).

As a result of modelling, variable importance values were calculated to determine the genes that are effective on lung cancer. The results obtained according to the calculated variable importance values are as follows. AGTR1, TNXB///TNXA, SPP1, PLPP2, SFTPC genes were found to be the genes most commonly associated with lung cancer, respectively. According to the results of the ROC analysis performed with the 3 of these genes with the highest variable importance values, it can be said that these three genes distinguish cancerous and non-cancerous tissues quite well. When bioinformatics and modelling results and ROC analysis results are considered together, it can be said that AGTR1, TNXB///TNXA, SPP1 genes are potential genes that can be biomarkers.

In a study, the relationship between AGRT1 and lung adenocarcinoma was investigated. The findings from enrichment analysis followed by in vitro validation suggest that AGTR1 may be involved in the pathogenesis of LUAD through the PI3K/AKT3 signalling pathway [21]. In a separate investigation, it was observed that the SPP1 gene had significantly elevated levels of expression in lung cancer tissues in comparison to normal tissues. The elevated expression of SPP1 was additionally correlated with tumor grade and worse clinical outcome [22]. A separate investigation has demonstrated a associated between SPP1 and unfavorable prognosis as well as chemoresistance in cases of lung cancer [23].

In conclusion, using gene expression data from tumor tissues and non-tumor tissues, this study identified potential genomic biomarkers for lung cancer. Given the forthcoming extensive study and investigations on these genetic factors, it is plausible that the development of targeted medicines and the incorporation of novel treatment approaches into the existing repertoire may be feasible.

References

1. Didkowska J, Wojciechowska U, Mańczuk M, et al. Lung cancer epidemiology contemporary and future challenges worldwide. Ann Transl Med. 2016; 4: 150.
2. Rahal Z, El Nemr S, Sinjab A, et al. Smoking and Lung Cancer A Geo-Regional Perspective. Front Oncol. 2017; 7: 194.
3. Hirsch FR, Scagliotti GV, Mulshine JL, et al. Lung cancer current therapies and new targeted treatments. The Lancet. 2017; 389: 299-311.
4. Xi KX, Zhang XW, Yu XY, et al. The role of plasma miRNAs in the diagnosis of pulmonary nodules. J Thorac Dis. 2018; 10: 4032-4041.
5. Lu S, Kong H, Hou Y, et al. Two plasma microRNA panels for diagnosis and subtype discrimination of lung cancer. Lung cancer. 2018; 123: 44-51.
6. Hirsch FR, Franklin WA, Gazdar AF, et al. Early detection of lung cancer: clinical perspectives of recent advances in biology and radiology. Clin Cancer Res. 2001; 7: 5-22.

7. Wadowska K, Bil-Lula I, Trembecki Ł, et al. Genetic markers in lung cancer diagnosis A review. Int J Mol Sci. 2020; 21: 4569.
8. Jakubek Y, Lang W, Vattathil S, et al. Genomic landscape established by allelic imbalance in the cancerization field of a normal appearing airway. Cancer res. 2016; 76: 3676-3683.
9. Santarpia M, Liguori A, D'Aveni A, et al. Liquid biopsy for lung cancer early detection. J Thorac Dis. 2018; 10: S882-S897.
10. Gazdar AF, Brambilla E. Preneoplasia of lung cancer. Cancer Biomark. 2011; 9: 385-396.
11. Polikar R. Ensemble learning. Ensemble machine learning. 2012; 1-34.
12. Akman M, Genç Y, Ankarali H. Random Forests Methods and an Application in Health Science. Turkiye Klinikleri J Biostat. 2011; 3: 36.
13. Akalın PK. Introduction to bioinformatics. Mol Nutr Food Res. 2006; 50: 610-619.
14. Van Hal NL, Vorst O, van Houwelingen AM, et al. The application of DNA microarrays in gene expression analysis. J Biotechnol. 2000; 78: 271-280.
15. Smyth GK. Limma linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: Springer. 2005; 397-420.
16. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23: 2507-2517.
17. Fodor IK. A survey of dimension reduction techniques. Lawrence Livermore National Lab. 2002.
18. Fonti V. Research Paper in Business Analytics Feature Selection with LASSO. Amsterdam: VU Amsterdam. 2017.
19. Friedman JH. Stochastic gradient boosting. Computational statistics and data analysis. 2002; 38: 367-378.
20. Lawrence R, Bunn A, Powell S, et al. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment. 2004; 90: 331-336.
21. Xiong L, Wei Y, Zhou X, et al. AGTR1 inhibits the progression of lung adenocarcinoma. Cancer Manag Res. 2021; 13: 8535-8550.
22. Tang H, Chen J, Han X, et al. Upregulation of SPP1 is a marker for poor lung cancer prognosis and contributes to cancer progression and cisplatin resistance. Front Cell Dev Biol. 2021; 9: 646390.
23. Matsubara E, Yano H, Pan C, et al. The Significance of SPP1 in Lung Cancers and Its Impact as a Marker for Protumor Tumor-Associated Macrophages. Cancers. 2023; 15: 2250.