Mugla Journal of Science and Technology

A PERSONALIZED ONCOLOGY MOBILE APPLICATION INTEGRATING CLINICAL AND GENOMIC FEATURES TO PREDICT THE RISK STRATIFICATION OF LUNG CANCER PATIENTS VIA MACHINE LEARNING

Mehmet Cihan SAKMAN,Department of Bioinformatics, Muğla Sıtkı Koçman University, Turkey, sakmancihan@gmail.com
(https://orcid.org/0000-0001-9541-123X)
Talip ZENGİN, Department of Bioinformatics, Muğla Sıtkı Koçman University,Turkey, talipzengin@posta.mu.edu.tr
(https://orcid.org/0000-0003-4764-4615)
Deniz KURŞUN, Department of Bioinformatics, Muğla Sıtkı Koçman University,Turkey, denizkursun@posta.mu.edu.tr
(https://orcid.org/0000-0002-1253-1242)
*Tuğba ÖNAL-SÜZEK, Department of Bioinformatics, Muğla Sıtkı Koçman University,Turkey, tugbasuzek@mu.edu.tr
(https://orcid.org/0000-0002-3243-1759)

Received: 23.08.2022, Accepted: 28.12.2022 Research Article
*Corresponding author DOI: 10.22531/muglajsci.1165634

Abstract

Predicting lung adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) risk status is a crucial step in precision oncology. In current clinical practice, clinicians, and patients are informed about the patient's risk group only with cancer staging. Although several machine learning approaches for stratifying LUAD and LUSC patients were described, the integrated machine learning model of clinical and genetic data of the two lung cancer types is not studied. In our work, we used a prognostic prediction model based on clinical and somatically altered gene features from 1026 patients to assess the relevance of features on risk classification. By integrating these two aspects, we achieved the highest accuracy; 93% for LUAD and 89% for LUSC, respectively. Our second finding is KEAP1 and CSMD3 are prognostic genes for LUAD and LUSC respectively and the site of resection is significantly associated with the risk stratification. We validated the risk stratification impact of CSMD3 on an independent RNAseq dataset from NCBI GEO (GSE81089) and finally integrated our model into a user-friendly mobile application. Using this machine learning model and mobile application, clinicians and patients can assess the survival risk of their patients using each patient's own clinical and molecular feature set.

Keywords: Machine learning, lung adenocarcinoma, lung squamous cell carcinoma, prognosis prediction model, the cancer genome atlas, multi-omics, data integration, electronic health records

MAKİNE ÖĞRENİMİ YOLUYLA AKCİĞER KANSERİ HASTALARININ RİSK SINIFLANDIRMASINI ÖNGÖRMEK İÇİN KLİNİK VE GENOMİK ÖZELLİKLERİ BÜTÜNLEŞTİREN KİŞİSELLEŞTİRİLMİŞ ONKOLOJİ MOBİL UYGULAMASI

Özet

Akciğer adenokarsinomu (LUAD) ve Akciğer Skuamöz Hücreli Karsinom (LUSC) hastalarının sağkalım riskini tahmin etmek, hassas onkolojide çok önemli bir adımdır. Mevcut klinik uygulamada klinisyenler ve hastalar hastanın risk grubu hakkında sadece kanser evrelemesi ile bilgilendirilmektedir. LUAD ve LUSC hastalarını sınıflandırmak için çeşitli makine öğrenimi yaklaşımları tanımlanmış olmasına rağmen, iki akciğer kanseri türünün klinik ve genetik verilerinin entegre eden bir makine öğrenimi modeli çalışılmamıştır. Çalışmamızda, özniteliklerin risk sınıflandırmasıyla ilişkisini değerlendirmek için 1026 hastadan alınan klinik ve somatik olarak değiştirilmiş gen özniteliklerine dayanan bir prognostik tahmin modeli kullandık. Bu iki yönü entegre ederek en yüksek doğruluğu elde ettik; LUAD için sırasıyla %93 ve LUSC için %89. İkinci bulgumuz KEAP1 ve CSMD3'ün sırasıyla LUAD ve LUSC için prognostik genler olması ve rezeksiyon bölgesinin sağkalım risk sınıflandırması ile önemli ölçüde ilişkili olmasıdır. Bulunan prognostik genlerden CSMD3'ün NCBI GEO'dan (GSE81089) bağımsız bir RNAseq veri kümesi üzerindeki risk sınıflandırma etkisini doğruladık ve son olarak modelimizi kullanıcı dostu bir mobil uygulamaya entegre ettik. Bu makine öğrenimi modelini ve mobil uygulamayı kullanarak klinisyenler ve hastalar, her hastanın kendi klinik ve moleküler özellik setini kullanarak hastalarının hayatta kalma riskini değerlendirebilir.

Anahtar Kelimeler: Makine öğrenimi, akciğer adenokarsinomu, akciğer skuamöz hücreli karsinomu, prognoz tahmin modeli, kanser genom atlası, çoklu omikler, veri entegrasyonu, elektronik sağlık kayıtları

Cite

Sakman, M. C., Zengin T., Kurşun D., Önal-Süzek T. (2022) "A Personalized Oncology Mobile Application Integrating Clinical and Genomic Features to Predict the Risk Stratification of Lung Cancer Patients via Machine Learning", 8(2), 90-99.

1. Introduction

Lung cancer is the most common type of cancer and the leading cause of death worldwide. The World Health Organization (WHO) reported that lung cancer is the second most frequently diagnosed cancer type, constituting 11.4% of all cancers, and the leading cause of cancer-related deaths (18%) in 2020 [1]. In the United States, the 5-year survival rate of patients diagnosed with lung cancer is only 14%. Although this survival rate increased over the last two decades significant improvements appear unlikely in the near future [2].

Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main types of lung cancer. Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) are two of the three subtypes of NSCLC that affect 85% of patients. Patient risk group-based treatment can be administered based on the course of the cancer if the genetic and clinic profile of the patients can be obtained at an earlier stage.

The emergence and effectiveness of machine learning (ML) techniques to process massive volumes of data are revolutionizing bioinformatics and conventional approaches to genetic diagnostics. There has been a significant rise in the amount of cancer studies utilizing ML and molecular data in recent years producing outcomes that are comparable to those of traditional techniques for searching of potential genetic biomarkers. Due to its success in classification and prediction in supporting clinicians, ML offers efficient solutions to decision-making processes and the rising cost of health care while also improving patient-clinician communication [3,4].

Among the recent studies in ML based lung cancer prediction, several of them used Computed Tomography (CT) based features [5] whereas other techniques are more specific, and used genomic or phenotypic feature sets for building classification models [6]. Jones et al. developed a prediction model that combines genetic and clinical factors to predict cancer recurrence to compare the likelihood of recurrence in patients with traditional TNM-based disease prediction models and those with completely resected stages I to III LUAD [7]. More recently, Yang et al. used genomic information, clinical status and demographics and how they influence the prediction of recurrence and survival for both early stage LUAD and LUSC by comparing the accuracy of three ML algorithms: decision tree methods, neural networks and support vector machines [8].

In this study, for the first time to our knowledge, we experiment incorporating both electronic health record and mutation data features of The Cancer Genome Atlas (TCGA) [9] lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) patients. For this purpose, we extracted the clinical and genes with somatically altered mutations as training features from 1026 patients to construct a prognostic prediction model for their impact on risk classification and validated our prediction in an independent dataset from NCBI GEO.

2. Materials and Methods

2.1. Data Collection

The genetic data and the corresponding electronic health record (EHR) information of LUAD(522 patients) and LUSC (504 patients) [Table 4] are downloaded via TCGABiolinks R package. High or low survival risk of each patient and the top 10 somatically mutated genes and their mutation counts of each of the LUAD and LUSC patients are obtained from our previous publication [10]. To validate the importance of the features found by our ML model, we download the median fragments per kilobase of exon per million fragments mapped (FPKM) values and clinical features of an independent Gene Expression Omnibus (GEO) dataset https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81089 with 52 LUSC samples by Djureinovic et al [11].

2.2. Preprocessing

The imbalanced and complicated nature of clinical records due to missing values, among other data quality limitations, is a common problem. Most biomedical data is not smooth, requiring a number of preprocessing strategies to remove noise and recover valuable biomedical data. For this purpose, clinical features with less than 80% clinical data content are eliminated from the training and test datasets. Uninformative columns such as submitter id, diagnosis id, exposure id, demographic id, treatment id, and BCR patient barcode are excluded. Redundant columns such as year of birth, disease state, updated datetime, and tissue or organ of origin columns are also eliminated. Data is partitioned into training (80%) and testing (20%) datasets using the sci-kit-learn's model_selection package and all subsequent exploratory data analysis and model training is performed only on the training portion of the dataset.

2.3. Missing Value Imputation

The most common two strategies to cope with the missing value problem are dropping the null values and filling out the missing values with the mean of the feature. Although the effectiveness of these approaches is questionable, we experimented with filling out the missing values in train and test data with the mean of the training data.

a) Days to death can be longer than days since the last follow-up, and unfortunately days to death for patients who are still alive are not available. As a result, we try three distinct approaches to dealing with the problem: deleting the days_to_last_follow_up, assigning 0 for all alive patients in days_to_death, and deleting the patients when a dead patient's days_to_death value is 0. As a result of this experiment, there is a 0.63 correlation between risk and the vital status with no link between them.

b) We experiment with the assumption that patients will live until 90 years old and fill the NaN values for alive patients. As a result of this experiment, there is a 0.83

correlation between vital state and risk, but no link between risk and vital status.

c) We transferred the alive patient information into the days_to_death column and dropped the days_to_last_follow_up column. As we can see, we could not find any correlation for days to death neither between vital status nor risk of the patient.

After computing the correlation coefficients of the features for these three strategies (Figure 1), alive patients' days_to_death values are filled with the days_to_last_follow_up feature.

2.4. Numeric Columns Imputation

The missing values in the columns (age_at_diagnosis, days_to_birth, years_smoked, cigarettes_per_day) of both LUAD and LUSC training and test datasets are filled with the mean values of the training data using the SimpleImputer library in scikit-learn. The ratio of missing values in the years_smoked feature is 62.83% in LUAD and 55.55% in LUSC therefore they are removed. For LUAD, cigarettes per day column contains data for 61% of the patients therefore we could only include this feature for LUSC (86%).

2.5. Categorical Columns Imputation

The fundamental issue with categorical features is that many ML algorithms cannot operate directly on label data. The majority of ML methods require all input and output variables to be numeric. Therefore, we used the scikit-learn library's LabelEncoder to convert each unique value into an integer value to encode our categorical data into numerical data. The integer representation is then encoded using OneHotEncoding. For each unique integer value, the integer encoded variable is removed and a new binary variable is added.

2.6. Balancing the Imbalanced Classes

For the LUAD and LUSC datasets, the number of patients surviving more than 5 years is approximately 9 times smaller than the opposite. Oversampling and undersampling lead to similar performances provided that the sampling is correctly implemented on the training and testing folds separately [6]. For some of the earlier TCGA studies, SMOTE was applied to the TCGA training set [6,12]. Therefore, we over-sampled both the training and testing patient set via Synthetic Minority Oversampling Technique (SMOTE). The class distributions before and after applying the SMOTE algorithm are presented in Figure 2.

2.7. Creation of the Machine Learning Models

After over-sampling, five different classification algorithms (Logistic Regression, Random Forest Classifier, Naive Bayes, SVC, and K-Neighbors Classifier) are implemented. The receiver operating characteristics (ROC) curves are plotted and Area Under The Curve (AUC) values are obtained to assess the performance of the learning algorithms (Figure 3, 4). In order to support the AUC [13,14] metrics, the F1-score, precision, and recall are calculated and shown in Table 1 and Table 2.

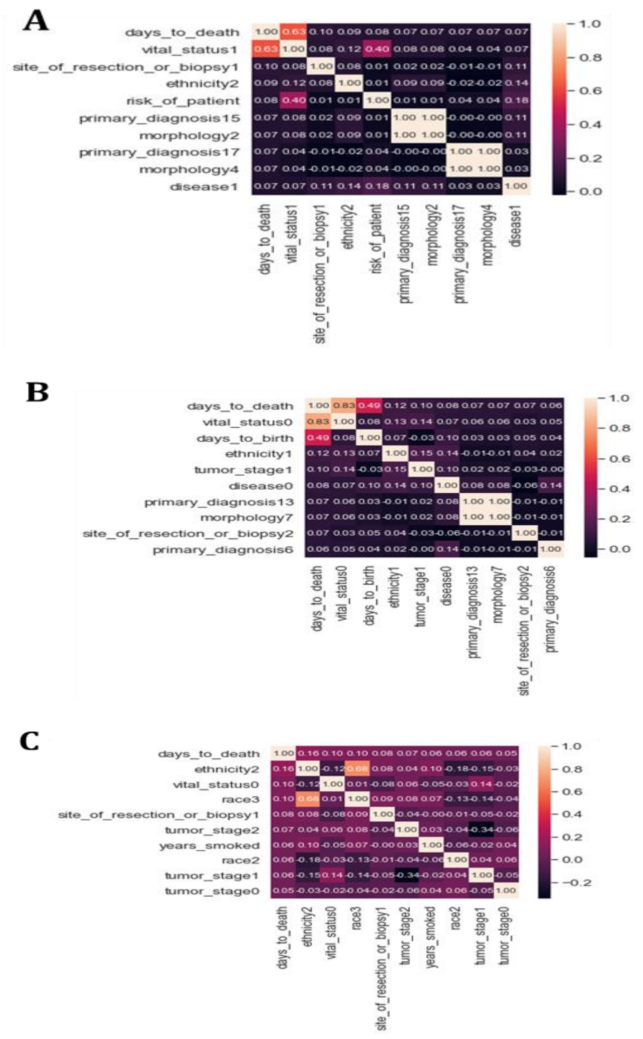


Figure 1. Correlation coefficients of three strategies for assigning missing values of days_to_death column (A) filling with zero (B) filling with days_until_90s (C)filling with days_to_last_follow_up.

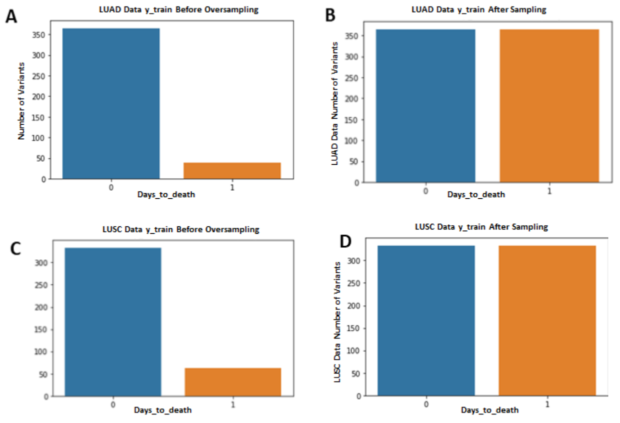


Figure 2. The Synthetic Minority Oversampling Technique (SMOTE) was used to address the class imbalance issue in the training and testing data sets separately. (A,B) LUAD data before and after oversampling, (C,D) LUSC data before and after oversampling.

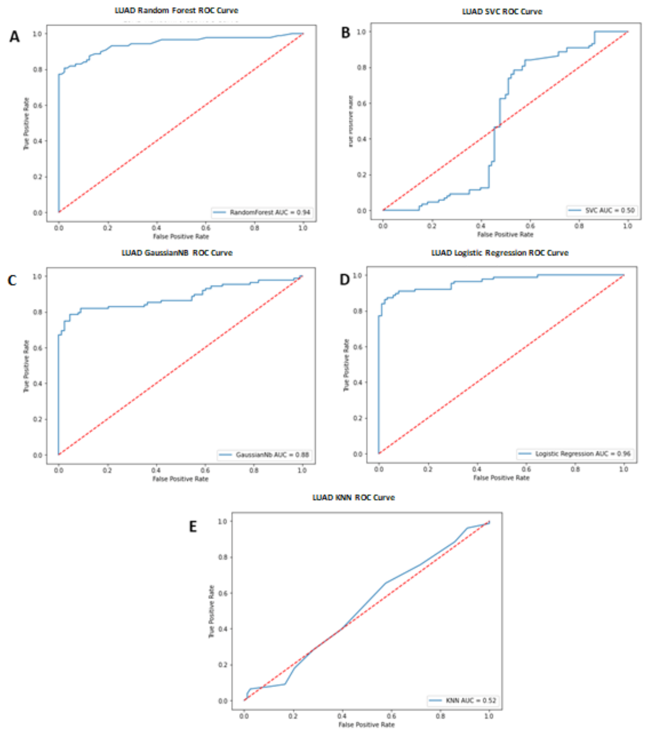


Figure 3. Receiver operating characteristics (ROC) curves are plotted and Area Under The Curve (AUC) values are obtained to evaluate the performance of five different algorithms for LUAD data set (A) Random Forest, (B) Support Vector Machine (SVC), (C) Gaussian Naive Bayes (GaussianNB), (D) Logistic Regression, (E) K-Neighbors Classifier (KNN).

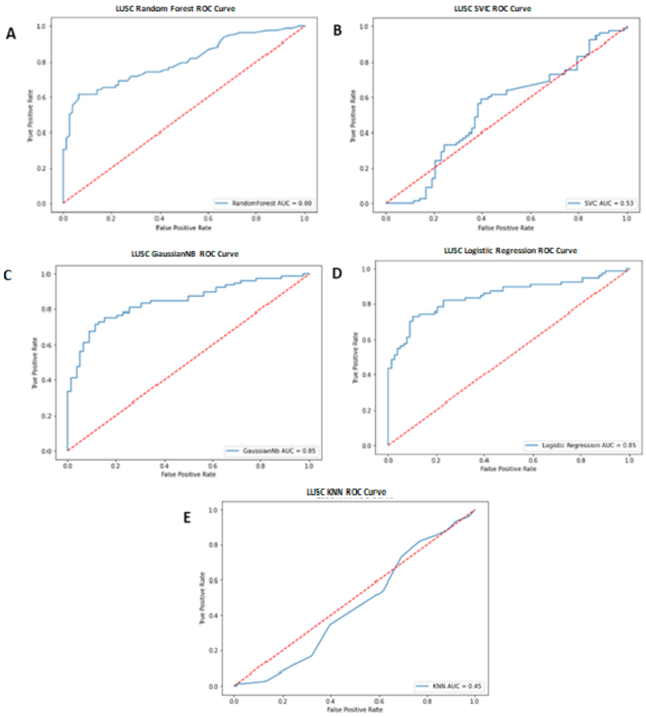


Figure 4. Receiver operating characteristics (ROC) curves are plotted and Area Under The Curve (AUC) values are obtained to evaluate the performance of five different algorithms for LUSC data set (A) Random Forest, (B) Support Vector Machine (SVC), (C) Gaussian Naive Bayes (GaussianNB), (D) Logistic Regression, (E) K-Neighbors Classifier (KNN).

Table 1. Comparison of the Precision, Recall, F1-score of 5 algorithms for LUAD.

Model Name	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.90	0.89	0.89	0.89
Random Forest	0.82	0.81	0.81	0.81
Naive Bayes	0.72	0.65	0.62	0.71
SVC	0.53	0.52	0.45	0.53
KNN	0.62	0.62	0.62	0.62

Table 2. Comparison of the Precision, Recall, F1-score of 5 algorithms for LUSC.

Model Name	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.75	0.74	0.73	0.74
Random Forest	0.71	0.69	0.68	0.68
Naive Bayes	0.70	0.62	0.58	0.71
SVC	0.56	0.53	0.48	0.56
KNN	0.53	0.53	0.52	0.53

2.8. Hyperparameter Tuning

Among the five different classification algorithms, the top two best scoring algorithms are found as Random Forest and Logistic Regression, thus hyperparameter tuning is applied to Random Forest and Logistic Regression based models. For hyperparameter tuning of these two algorithms, we implemented a 5-fold cross-validation where we first split the training set into 5 folds and then applied random oversampling on 4 folds which are used for training the classification model and then documented the model performance metrics on the remaining 1-fold using the GridSearchCV in scikit-learn.

3. Results and Discussion

All tumor stages are cconsolidated into four main tumor stages. During the course of the TCGA project, 188 patients died in LUAD and 220 patients died in LUSC. Table 4 summarizes the overall statistics of the LUAD and LUSC clinical features.

To identify the optimum parameters for the models, we employed GridSearchCV(scoring='f1', cv=5) in scikit-learn to Random Forest and Logistic Regression as it is the preferred method for adjusting hyperparameters [15] (Table 3).

Table 3. Comparison of performance metrics for Random Forest and Logistic Regression with and without hyperparameter tuning.

Model Name	Data	GridSearchCV	Precision	Recall	F1-Score	Accuracy
Logistic Regression	LUAD	No/Yes	0.92/0.94	0.92/0.93	0.92/0.93	%91.7/%93.1
Random Forest	LUAD	No/Yes	0.88/0.90	0.88/0.90	0.87/0.90	%87.9/%89.7
Logistic Regression	LUSC	No/Yes	0.77/0.82	0.74/0.78	0.74/0.77	%74.3/%77.5
Random Forest	LUSC	No/Yes	0.72/0.77	0.68/0.71	0.66/0.69	%66.6/%70.5

Table 4. Clinical properties of the LUAD and LUSC patients.

LUAD Patient Property Name	Number	LUSC Patient Property Name	Number
Age at diagnosis (median; range)	67 (33-89)	Age at diagnosis (median; range)	68 (39-90)
Gender		**Gender**	
Female	280	Female	131
Male	242	Male	373
Smoking Associated Features		**Smoking Associated Features**	
Number of cigarettes per day (mean; range)	2 (0-9)	Number of cigarettes per day (mean; range)	3 (0-13)
Number of years smoked (mean; range)	32 (2-64)	Number of years smoked (mean; range)	40 (8-63)
Tumor Stage		**Tumor Stage**	
I	279	I	245
II	124	II	163
III	85	III	85
IV	26	IV	7
NA	8	NA	4
Vital Status		**Vital Status**	
Alive	334	Alive	284
Dead	188	Dead	220

Table 5. Accuracies of each fold for LUSC for 5-fold cross-validation.

Model Name and Parameters	Metrics	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
LogisticRegression(C= 109.85411419875572, penalty='l1', solver='liblinear')	F1-Score	0.90	0.85	0.80	0.83	0.81	0.84	0.040
	Precision	0.91	0.86	0.84	0.86	0.81	0.86	0.038
	Recall	0.90	0.85	0.81	0.83	0.81	0.84	0.039
	Accuracy	0.90	0.85	0.81	0.83	0.81	0.84	0.039
	AUC	0.93	0.89	0.86	0.88	0.87	0.89	0.025
RandomForestClassifier(max_depth= 8, max_features='auto', min_samples_leaf=3, n_estimators=200)	F1-Score	0.78	0.83	0.82	0.69	0.62	0.75	0.090
	Precision	0.79	0.84	0.83	0.71	0.67	0.77	0.074
	Recall	0.78	0.83	0.82	0.70	0.64	0.75	0.083
	Accuracy	0.78	0.83	0.82	0.70	0.64	0.75	0.083
	AUC	0.79	0.92	0.92	0.76	0.72	0.82	0.092

3.1. K-Fold Cross-Validation

We next apply 5-fold cross-validation with Logistic Regression and Random Forest models with hyperparameters to avoid bias [16,17]. Random Forest and Logistic Regression models' results for each five-folds are documented in Table 5 and Table 6.

3.2. Random Forest Feature Importance

The feature selection techniques select a subset of the most relevant features according to the target feature. The main goal of choosing the most relevant features is running the algorithms more efficiently to overcome space and time complexity problems. Irrelevant input features can mislead the ML algorithms resulting in worse performance. In this work, we compared the feature importances on two alternative sets of features; 1) Full set of clinical features and 2) Clinical features with top 10 somatically mutated driver genes. Importance ranking of the features are provided by the fitted attribute feature_importances_of the scikit-learn Python ML library. The feature importances are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree (Figure 5A, 5B). We also plotted the top 9 most correlated features to days_to_death (Figure 6A, 6B).

3.3. Somatically Mutated Genes

In this experiment, we investigated which somatically mutated genes effect the survival by feature engineering. We added the mutation counts of the top ten most commonly mutated somatic driver genes along with clinical features to create the training model. For this purpose, we selected the top 10 most highly mutated somatic driver genes for LUAD, and LUSC identified in our previous publication [10] using SominaClust [18]. The top ten most somatically mutated driver genes for LUAD patients are CDH10, COL11A1, CSMD3, HMCN1, KEAP1, KRAS, LRP1B, SPTA1, TP53, and USH2A. Although KEAP1, TP53, USH2A, and CSMD3 popped up in the most important top 10 feature list, adding the genes had no discernible effect on the LUAD Logistic Regression and Random Forest models' accuracy (Figure 7A, Figure 8A). The performance of each fold of 5-fold cross-validation is presented in Table 7.

Moreover, KEAP1 mutation has higher feature importance followed by TP53, USH2A, CSMD3, LRP1B, SPTA1, CDH10, HMCN1, KRAS and COL11A1. Mutated genes have higher importance than many clinical features. To examine how much a patient's clinical variants and somatic mutation profile affect a patient's survival risk, we set up separate machine learning models that incorporate these two types of features separately and together. Evaluating the accuracy of these machine learning models using combinations of clinical and mutation features, we showed that clinical variables are more effective in predicting patient survival than mutation data. Site of resection, morphology, primary diagnosis, smoking amount have importance along with the gene mutations.

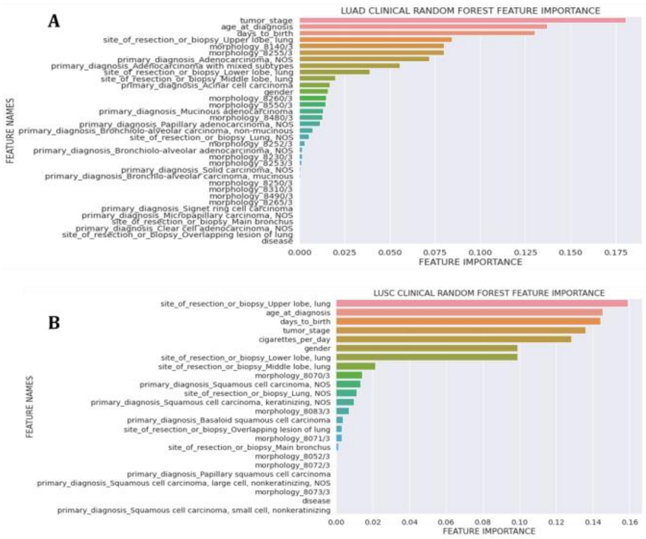


Figure 5. Importance ranks of LUAD (A) and LUSC (B) clinical features only

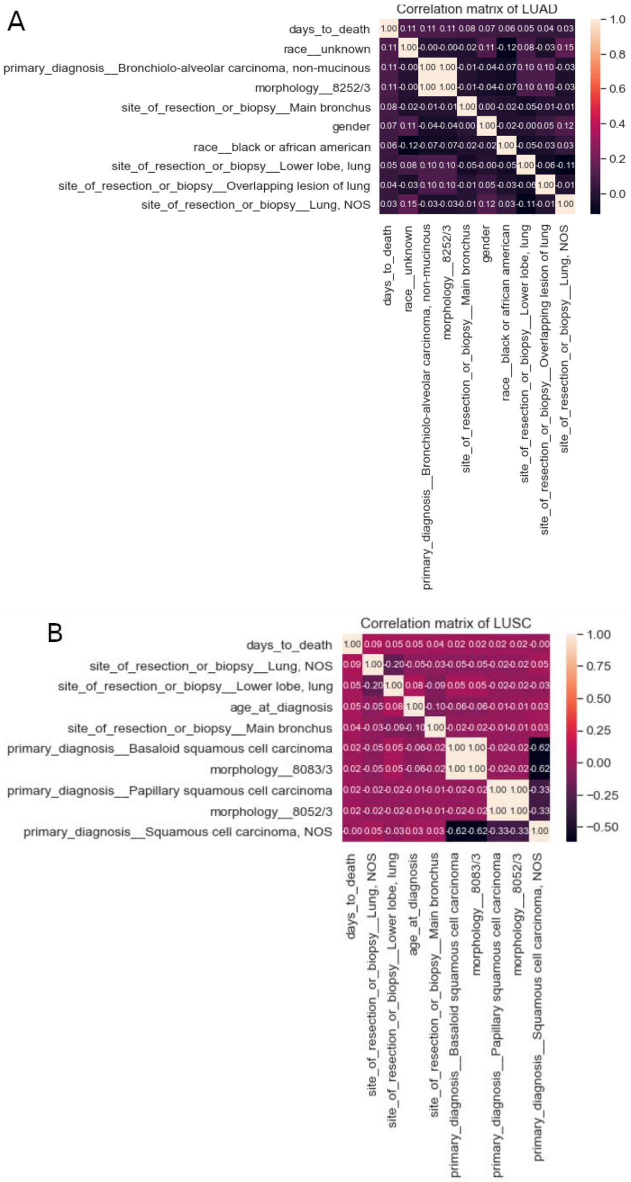


Figure 6. Top 9 most correlated LUAD (A) and LUSC (B) clinical features to days_to_death

Incorporating gene mutation features did not change the performance for LUAD model yet we included it in the final LUAD model integrated into the mobile application. For example, loss of function mutations in KEAP1 gene, promote KRAS-driven lung tumorigenesis [19] that may be reason of correlation of KEAP1 with risk of patients, therefore KEAP1 and KRAS can be used along with clinical variables.

Unlike LUAD, addition of the top 10 most highly mutated somatic driver genes to the feature set vastly improved the classification model of LUSC patients (Table 8). Top 10 somatically mutated genes of LUSC patients reported in our previous publication are CDKN2A, CSMD3, FAT1, KEAP1, KMT2C, KMT2D, NF1 NFE2L2, PIK3CA, TP53 [10].

Integrating the given ten genes improved the accuracy of the learning model most for the Logistic Regression and Random Forest in LUSC (Figure 7B, Figure 8B). Among these 10 genes, CSMD3 is known as one of the most frequently mutated genes in lung cancer and a potential tumor-suppressor [20]. Following CSMD3, smoking related features ranked the highest importance. This finding is also in agreement with LUSC pathogenesis. Unlike LUAD, LUSC pathogenesis is strongly associated with airway lesions that arise with smoking and is mostly located in the central parts of the lung [21]. Following smoking, age, tumor stage and TP53, KEAP1, NFE2L2, KMT2D, KMT2C, FAT1, CDKN2A, NF1 and PIK3CA gene mutations appeared in top feature list of LUSC.

3.4. Kaplan Meier Analysis

As we want to validate our finding that CSMD3 can be a prognostic biomarker gene for LUSC, we perform a Kaplan-Meier survival analysis on an independent dataset of GSE81089. FPKM values and the electronic health records of 52 LUSC patients by Djureinovic et al. [11] are downloaded from GEO web site. After sorting the patients by their FPKM values, we compared the survival of the patients with FPKM values above third quartile (high) against the patients with FPKM values in the first quartile (low). These 26 LUSC samples are labeled as low and high expression groups. The prognostic value of genes for OS are performed with the hazard ratios (HR) and Log-Rank p-values. Our analysis results indicated that the patients with low expression of CSMD3 had significantly better prognosis (p=0.037) (Figure 9).

Lastly, best performing ML model has been integrated to a user-friendly mobile interface (Figure 10) enabling both clinicians and lung cancer patients to assess the patients' risk stratification.

4. Conclusion

Main goal of our study was to investigate the effect of clinical features and the biomarker genes most helpful the in prediction of survival stratification of LUAD and LUSC patients. For this purpose, we implement several ML studies investigating the vast clinical feature set and top 10 most somatically mutated gene set of TCGA lung adenocarcinoma and lung squamous carcinoma patients and rank the features contributing to the risk stratification. An important finding of our study is that, in general, clinical features have more survival predictive power than somatic mutations, therefore, we recommend gene mutations and clinical features to be evaluated together by the clinicians for predicting risk of survival.

New genes such as KEAP1 for LUAD, CSMD3 for LUSC and new clinical features such as site of resection are discovered as potential features to be added to clinical decision process. We predict that LUSC patients with low CSMD3 mutation rates have a favorable prognosis in the TCGA dataset using our proposed ML method integrating clinical and mutation factors. Analyzing an independent dataset from NCBI GEO, we confirmed that low expression level of CSMD3 for LUSC patients is an indication of a significantly favorable survival.

Future work of research involves integrating more cancers into our mobile interface and deploying our mobile application to public dissemination for both IOS and Android platforms.

Table 6. Accuracies of each fold for LUAD for 5-fold cross- validation.

Model Name and Parameters	Metrics	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
LogisticRegression(C=16.768329368110066, penalty='l2', solver='newton-cg')	F1-Score	0.91	0.93	0.89	0.91	0.88	0.904	0.017
	Precision	0.91	0.94	0.90	0.91	0.90	0.912	0.014
	Recall	0.91	0.94	0.89	0.91	0.89	0.908	0.018
	Accuracy	0.91	0.93	0.89	0.91	0.89	0.906	0.015
	AUC	0.95	0.95	0.94	0.95	0.89	0.936	0.023
RandomForestClassifier(max_depth=8, max_features='log2', min_samples_leaf=3, n_estimators=50)	F1-Score	0.86	0.86	0.87	0.87	0.73	0.838	0.054
	Precision	0.86	0.88	0.88	0.88	0.79	0.858	0.035
	Recall	0.86	0.86	0.87	0.87	0.74	0.84	0.050
	Accuracy	0.86	0.86	0.87	0.87	0.74	0.84	0.050
	AUC	0.94	0.91	0.95	0.91	0.82	0.906	0.046

Table 7. Accuracies of each fold for LUAD model with clinical features and top 10 somatically mutated genes for 5-fold cross-validation

Model Name	**Metrics**	**Feature Type**	**Fold-1**	**Fold-2**	**Fold-3**	**Fold-4**	**Fold-5**	**mean**	**std**
Logistic Regression	**F1-Score**	Clinical/ Mutation	0.91/0.89	0.93/0.90	0.89/0.90	0.91/0.88	0.88/0.89	0.90/0.89	0.017/0.009
	Precision	Clinical/ Mutation	0.91/0.90	0.94/0.91	0.90/0.90	0.91/0.88	0.90/0.89	0.91/0.90	0.014/0.011
	Recall	Clinical/ Mutation	0.91/0.89	0.94/0.90	0.89/0.90	0.91/0.88	0.89/0.89	0.91/0.89	0.018/0.009
	Accuracy	Clinical/ Mutation	0.91/0.89	0.93/0.90	0.89/0.90	0.91/0.88	0.89/0.89	0.90/0.89	0.015/0.009
	AUC	Clinical/ Mutation	0.95/0.91	0.95/0.98	0.94/0.93	0.95/0.92	0.89/0.95	0.93/0.94	0.023/0.028
Random Forest	**F1-Score**	Clinical/ Mutation	0.86/0.81	0.86/0.90	0.87/0.77	0.87/0.80	0.73/0.81	0.84/0.82	0.054/0.052
	Precision	Clinical/ Mutation	0.86/0.85	0.88/0.91	0.88/0.85	0.88/0.82	0.79/0.87	0.86/0.86	0.035/0.031
	Recall	Clinical/ Mutation	0.86/0.82	0.86/0.90	0.87/0.78	0.87/0.79	0.74/0.81	0.84/0.82	0.050/0.049
	Accuracy	Clinical/ Mutation	0.86/0.81	0.86/0.91	0.87/0.78	0.87/0.79	0.74/0.81	0.84/0.82	0.050/0.049
	AUC	Clinical/ Mutation	0.94/0.93	0.91/0.96	0.95/0.88	0.91/0.89	0.82/0.92	0.90/0.92	0.046/0.031

Table 8. Accuracies of each fold for LUSC model with clinical features and top 10 somatically mutated genes for 5-fold cross-validation

Model Name	**Metrics**	**Feature Type**	**Fold-1**	**Fold-2**	**Fold-3**	**Fold-4**	**Fold-5**	**mean**	**std**
Logistic Regression	**F1-Score**	Clinical/ Mutation	0.90/0.89	0.85/0.80	0.80/0.87	0.83/0.82	0.81/0.85	0.84/0.85	0.040/0.036
	Precision	Clinical/ Mutation	0.91/0.90	0.86/0.82	0.84/0.90	0.86/0.83	0.81/0.86	0.86/0.87	0.038/0.037
	Recall	Clinical/ Mutation	0.90/0.89	0.85/0.80	0.81/0.87	0.83/0.82	0.81/0.85	0.84/0.84	0.039/0.036
	Accuracy	Clinical/ Mutation	0.90/0.89	0.85/0.80	0.81/0.87	0.83/0.83	0.81/0.85	0.84/0.85	0.039/0.034
	AUC	Clinical/ Mutation	0.93/0.93	0.89/0.88	0.86/0.93	0.88/0.90	0.87/0.91	0.89/0.91	0.025/0.021
Random Forest	**F1-Score**	Clinical/ Mutation	0.78/0.85	0.83/0.75	0.82/0.76	0.69/0.81	0.62/0.73	0.75/0.78	0.090/0.048
	Precision	Clinical/ Mutation	0.79/0.87	0.84/0.81	0.83/0.80	0.71/0.82	0.67/0.78	0.77/0.82	0.074/0.032
	Recall	Clinical/ Mutation	0.78/0.85	0.83/0.76	0.82/0.77	0.70/0.81	0.64/0.74	0.75/0.79	0.083/0.044
	Accuracy	Clinical/ Mutation	0.78/0.85	0.83/0.76	0.82/0.77	0.70/0.81	0.64/0.74	0.75/0.79	0.083/0.044
	AUC	Clinical/ Mutation	0.79/0.91	0.92/0.86	0.92/0.82	0.76/0.88	0.72/0.84	0.82/0.86	0.092/0.037

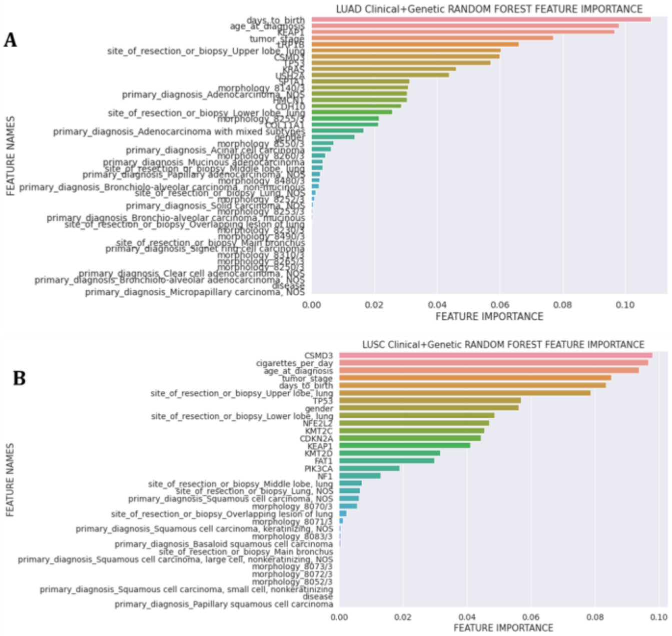


Figure 7. Importance ranks of LUAD (A) and LUSC (B) top 10 somatically mutated genes and clinical features.

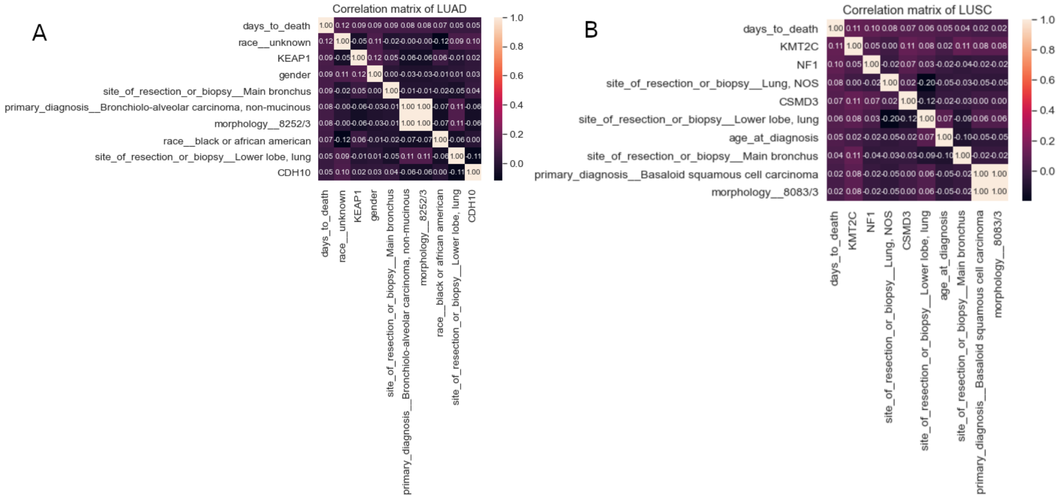
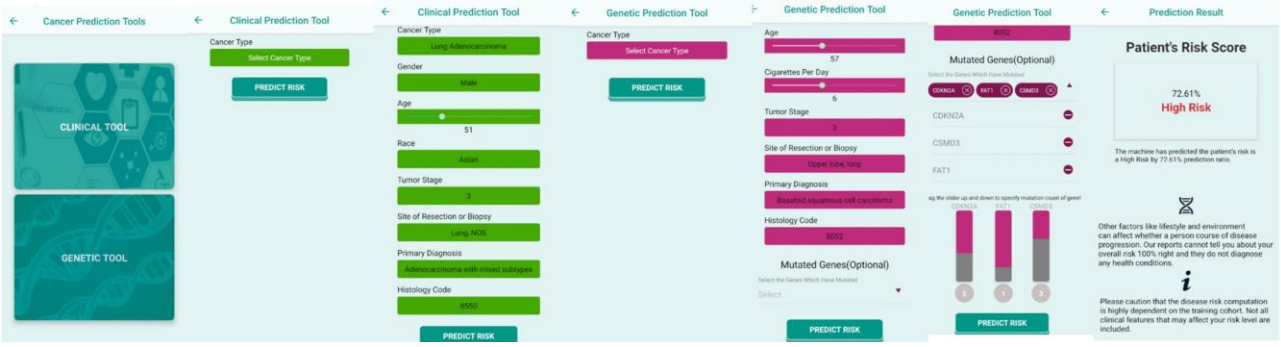


Figure 8. Top 9 most correlated LUAD (A) and LUSC (B) clinical features and top10.



Cancer Type: Select Cancer Type

Cancer Type: Lung Adenocarcinoma
Gender: Male
Age: 51
Race: Asian
Tumor Stage: 3
Site of Resection or Biopsy: Lung, NOS
Primary Diagnosis: Adenocarcinoma with mixed subtypes
Histology Code: 8550

Cancer Type: Select Cancer Type

Age: 57
Cigarettes Per Day: 6
Tumor Stage: 3
Site of Resection or Biopsy: Upper lobe, lung
Primary Diagnosis: Basaloid squamous cell carcinoma
Histology Code: 8052
Mutated Genes(Optional)
Select the Genes Which Have Mutated
Select

8052
Mutated Genes(Optional)
Select the Genes Which Have Mutated
CDKN2A FAT1 CSMD3
CDKN2A
CSMD3
FAT1
ag the slider up and down to specify mutation count of gene!
CDKN2A FAT1 CSMD3
2 1 3

Patient's Risk Score

72.61%
High Risk

The machine has predicted the patient's risk is a High Risk by 72.61% prediction ratio.

Other factors like lifestyle and environment can affect whether a person course of disease progression. Our reports cannot tell you about your overall risk 100% right and they do not diagnose any health conditions.

Please caution that the disease risk computation is highly dependent on the training cohort. Not all clinical features that may affect your risk level are included.

Figure 10. Mobile application user interfaces.

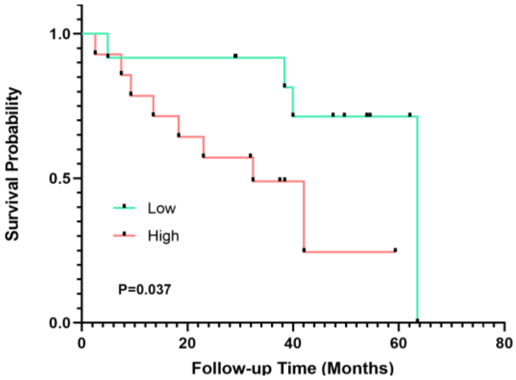


Figure 9. Survival plot of the low and high risk LUSC patients (P=0.037).

5. Data Availability Statement

The datasets supporting the conclusions of this article are publicly available from TCGA data portal (https://portal.gdc.cancer.gov). All code is available from https://github.com/cihansakman/MyCancerApp

6. Author Contributions

Methodology: MCS; formal analysis: MCS; mobile application development: MCS; writing—original draft preparation: MCS; visualization: MCS,DK; writing—review and editing: MCS,DK,TZ,TÖS; resources: MCS, TÖS; data curation: TZ,TÖS; project administration: TZ, TÖS. All authors have read and agreed to the published version of the manuscript.

7. Acknowledgment

DK was funded by YOK 100/2000. TZ, TÖS and DK are partially funded by TÜSEB 4583. MCS was funded by TÜBİTAK 2209A and Kosgeb Research & Innovation grant.

8. References

[1] IARC. “Globocan 2020 - Cancer Today.” *Int Agency Res Cancer* 2022;

[2] DeVita, V. T., Lawrence, T. S., & Rosenberg SA. DeVita, Hellman, and Rosenberg’s cancer: principles & practice of oncology. 10th ed. Lippincott Williams & Wilkins; 2015.

[3] Liñares-Blanco J, Pazos A, Fernandez-Lozano C. “Machine learning analysis of TCGA cancer data.” *PeerJ Comput Sci* 2021;7:e584.

[4] Bhargava N, Sharma S, Purohit R, et al. “Prediction of recurrence cancer using J48 algorithm.” *2017 2nd Int Conf Commun Electron Syst* 2017;386–390.

[5] Baskar S, Shakeel PM, Sridhar KP, et al. “Classification system for lung cancer nodule using machine learning technique and CT images.” *2019 Int Conf Commun Electron Syst* 2019;1957–1962.

[6] Sherafatian M, Arjmand F. “Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data.” *Oncol Lett* 2019;18:2125–2131.

[7] Jones GD, Brandt WS, Shen R, et al. “A Genomic-Pathologic Annotated Risk Model to Predict Recurrence in Early-Stage Lung Adenocarcinoma.” *JAMA Surg* 2021;156:e205601.

[8] Yang Y, Xu L, Sun L, et al. “Machine learning application in personalised lung cancer recurrence and survivability prediction.” *Comput Struct Biotechnol J* 2022;20:1811–1820.

[9] Liu J, Lichtenberg T, Hoadley KA, et al. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.” *Cell* 2018;173:400-416.e11.

[10] Zengin T, Önal-Süzek T. “Comprehensive profiling of genomic and transcriptomic differences between risk groups of lung adenocarcinoma and lung squamous cell carcinoma.” *J Pers Med* 2021;11:154.

[11] Djureinovic D, Hallström BM, Horie M, et al. “Profiling cancer testis antigens in non–small-cell lung cancer.” *JCI Insight* 2019;1:1–18.

[12] Yu L, Tao G, Zhu L, et al. “Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis.” *BMC Cancer* 2019;19:1–12.

[13] Provost F, Fawcett T. “Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions.” *KDD-97 Proc* 1997;43–48.

[14] Provost F, Fawcett T. “Robust classification for imprecise environments.” *Mach Learn* 2001;42:203–231.

[15] Zhao S, Mao X, Lin H, et al. “Machine Learning Prediction for 50 Anti-Cancer Food Molecules from 968 Anti-Cancer Drugs.” *Int J Intell Sci* 2020;10:1–8.

[16] Ramezan C, Warner T MA. “Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification.” *Remote Sens* 2019;11:185.

[17] Bengio Y, Grandvalet Y. “No unbiased estimator of the variance of k-fold cross-validation.” *Adv Neural Inf Process Syst* 2003;16:.

[18] Van den Eynden J, Fierro AC, Verbeke LPC, et al. “SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering.” *BMC Bioinformatics* 2015;16:1–12.

[19] Romero R, Sayin VI, Davidson SM, et al. “Keap1 loss promotes Kras-driven lung cancer and results in dependence on glutaminolysis.” *Nat Med* 2017;23:1362–1368.

[20] Liu P, Morrison C, Wang L, et al. “Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing.” *Carcinogenesis* 2012;33:1270–1276.

[21] Anusewicz D, Orzechowska M, Bednarek AK. “Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of Notch, Hedgehog, Wnt, and ErbB signalling.” *Sci Rep* 2020;10:21128.