Am J Cancer Res 2020;10(12):4624-4639
www.ajcr.us /ISSN:2156-6976/ajcr0126537

Original Article

Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models

Fei Deng[1], Lanlan Shen[2], He Wang[3], Lanjing Zhang[4,5,6,7]

[1]School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China; [2]Department of Pediatrics, Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, TX, USA; [3]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA; [4]Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA; [5]Department of Biological Sciences, Rutgers University, Newark, NJ; [6]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA; [7]Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA

Received November 18, 2020; Accepted November 25, 2020; Epub December 1, 2020; Published December 15, 2020

Abstract: Classification of multicategory survival-outcome is important for precision oncology. Machine learning (ML) algorithms have been used to accurately classify multi-category survival-outcome of some cancer-types, but not yet that of lung adenocarcinoma. Therefore, we compared the performances of 3 ML models (random forests, support vector machine [SVM], multilayer perceptron) and multinomial logistic regression (Mlogit) models for classifying 4-category survival-outcome of lung adenocarcinoma using the TCGA. Mlogit model overall performed similar to SVM and multilayer perceptron models (micro-average area under curve=0.82), while random forests model was inferior. Surprisingly, transcriptomic data alone and clinico-transcriptomic data appeared sufficient to accurately classify the 4-category survival-outcome in these patients, but no models using clinical data alone performed well. Notably, *NDUFS5, P2RY2, PRPF18, CCL24, ZNF813, MYL6, FLJ41941, POU5F1B,* and *SUV420H1* were the top-ranked genes that were associated with alive without disease and inversely linked to other outcomes. Similarly, *BDKRB2, TERC, DNAJA3, MRPL15, SLC16A13, CRHBP* and *ACSBG2* were associated with alive with progression and *GAL3ST3, AD2, RAB41, HDC*, and *PLEKHG1* associated with dead with disease, respectively, while also inversely linked other outcomes. These cross-linked genes may be used for risk-stratification and future treatment development.

Keywords: Lung adenocarcinoma, cause-specific mortality, survival, machine learning, multilabel classification, transcriptomic

Introduction

Lung cancer is the most common cause of cancer deaths among men or women in the U.S.A., accounting for 135,720 deaths in 2020 [1], although its trend in age-standardized, sex- and race-adjusted morality was downward in the past 5 years [2]. Thanks to better treatments, an increasing number of lung cancer patients died of non-cancer causes in the past decade [3], but the overall survival of lung cancer remains dismal. Besides development of additional targeted therapies, a better risk stratification and subsequent treatment decision-making are urgently needed to reduce the deaths of lung cancer. It is probably equally important to deescalate the treatment intensity to reduce non-cancer deaths in other lung cancer patients.

Many clinical and genomic features have been identified for the prognostication of lung adenocarcinoma [4-8]. The advances in machine learning (ML) algorithm also helped develop several gene-signatures for survival prediction in lung adenocarcinoma patients [9-15]. However, most of these works have been focused on binary survival outcomes, either disease-free survival or overall survival. To fully realize the potential power of ML, we may use

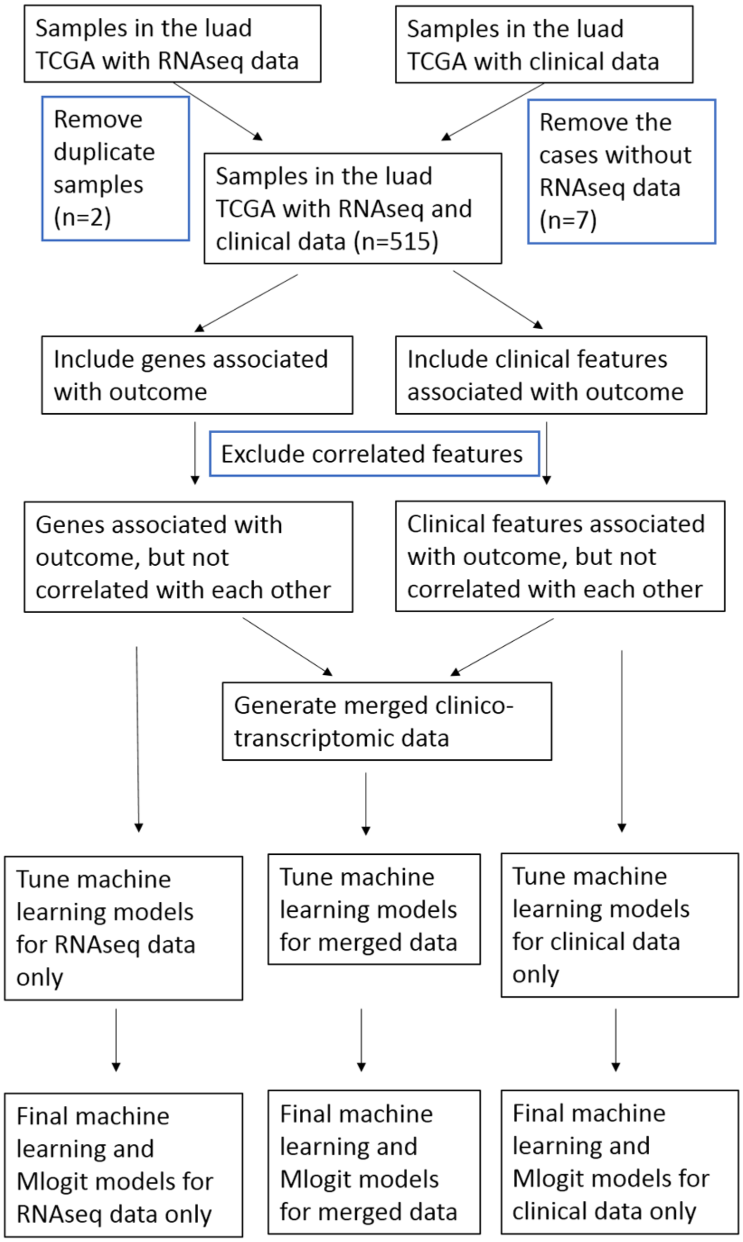

Samples in the luad TCGA with RNAseq data
Samples in the luad TCGA with clinical data
Remove duplicate samples (n=2)
Samples in the luad TCGA with RNAseq and clinical data (n=515)
Remove the cases without RNAseq data (n=7)
Include genes associated with outcome
Include clinical features associated with outcome
Exclude correlated features
Genes associated with outcome, but not correlated with each other
Clinical features associated with outcome, but not correlated with each other
Generate merged clinico-transcriptomic data
Tune machine learning models for RNAseq data only
Tune machine learning models for merged data
Tune machine learning models for clinical data only
Final machine learning and Mlogit models for RNAseq data only
Final machine learning and Mlogit models for merged data
Final machine learning and Mlogit models for clinical data only

Figure 1. Study flow. We extracted the lung adenocarcinoma cases in the cancer genome atlas (TCGA), and classified the patient survival into 4 categories, including alive with no progression, alive with disease, dead with no known disease and dead with disease. We used random forests, support vector machine, and multilayer perceptron to classify the 4-category outcomes. The 5-fold cross-validation approach was used during the tuning and modelling of the machine learning algorithms.

ML to stratify the risks of cancer death, non-cancer death and being alive among lung cancer patients. This is statistically a multilabel classification problem. One of the common approaches to classify the lung cancer patients is using clinical or transcriptomic data [5, 10, 11, 16-21]. Indeed, we showed that random forests (RF) model outperformed the conventional multinomial logistic regression (Mlogit) model in classifying 5-category outcomes of lung cancer patients, using clinical data of a large population-based dataset [22]. However, it is still unclear whether other ML models and the transcriptomic data alone are sufficient for classifying multicategory outcomes of lung cancer. Therefore, we compared the performances of 3 ML and Mlogit models in classifying 4-category outcomes of lung adenocarcinoma, using transcriptomic data lone, clinical data alone and combined clinic and transcriptomic data.

Material and methods

We extracted the lung adenocarcinoma cases from the cancer genome atlas (TCGA, legacy Version) from the cBioPortal website (**Figure 1**) [23]. No exclusion-criteria were used. The outcome was the 4-category survival-outcome based on the vital status and disease-free status, including alive with no progression, alive with disease, dead with no known disease, and dead with disease. We first dichotomized the RNAseq data (here referred as transcriptomic data) using the normalized Z scores based on their expression in all patients. We then identified and included only the genes that were statistically associated with the 4-category outcome. After identification and removal of the correlated genes and clinical features, we merged the two datasets into one single dataset (referred as clinico-transcriptomic data). The clinical, transcriptomic, and clinico-transcriptomic data were then subject to the tuning of ML models and conventional Mlogit models, respectively.

The informed consent could not be and was not obtained for the TCGA patients due to de-identified nature of the dataset. Because we used de-identified, publicly available cases, this study was deemed exempt from review by an institutional review board. Moreover, a set of policies were developed by National Cancer Institute and National Human Genome Research Institute to protect the privacy of participants donating specimens to TCGA, including the TCGA's informed consent policy, data access policy and information about Health Insurance Portability and Accountability Act Privacy Rule compliance (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies/tcga-human-subjects-data-policies.pdf). Thus, the TCGA data collection was supervised by respective funding agencies and their ethical review committees.

We used the RF, support vector machine (SVM) with linear or radial basis function (RBF) kernel, and multilayer perceptron (MLP) models of the sklearn library [24]. The linear support vector classifier (SVC) was used as a reference, and differed from linear SVM in their kernels (liblinear vs libsvm in SVM) [24]. Specifically, we tuned the number of estimators, and the number of splits for RF, the C and gamma values for SVM, and the C and number of hidden layers for MLP models. The performance metrics for all models were precision, accuracy, recall and F1. The cross_validation library was used to conduct 5-fold cross validation in the ML model-tuning process. Otherwise, the sample split library was used to spilt samples in a proportion of 4:1 (i.e., 80% of the samples for training, and 20% of the samples for testing) as described before [24-26]. The parameters that produced the best accuracy were chosen as the final ML model, while the default settings were also used to avoid over-tuning or missing the proper parameter range. The true positive rate and the true negative rate/sensitivity were calculated using the OneVsRestClassifier library (sklearn). The receiver operator characteristics curve and area under the curve were calculated using the ROC library. The micro-average and macro-average metrics were computed as defined before [24].

The ML and Mlogit processes were conducted using python version 3.6.9. The cut off of *P* value of 0.05 was used to select the clinical or transcriptomic features that were correlated with the 4-category outcome. The Rho of 0.9 was used to select the transcriptomic or clinical features that were correlated with each other. The first identified feature of the two correlated features would be removed.

We used the Erichr app to conduct gene set enrichment analyses (GSEA) to identify the pathways and related diseases [27], that were associated with the ranked up-regulated or down-regulated transcriptomic features. We interrogated the ranked gene list for their enrichment in 6 domains, including BioPlanet (2019), Kyoto Encyclopedia of Genes and Genomes (KEGG) Human (2019), UK Biobank GWAS v1, gene ontologies (GO) Molecular Function, GO Cellular Component, and GO Biological Process. A *P* value less than 0.05 was considered statistically significant.

Results

Among the 522 patients in the TCGA lung adenocarcinoma dataset, 7 patients did not have any transcriptomic data. Among the 517 samples of RNA sequencing, 2 of them had duplicates and the duplicates were removed (**Figure 1**). Therefore, 515 cases/samples were included in the study (**Table 1**). Among the 17 clinical features, 16 were found uniquely correlated with the 4-category outcome as shown by the correlation study. In the 20,113 genes that were subject to the RNA sequencing, 2,887 genes were found significantly associated with the 4-category outcome and included in the analysis. The correlation study showed that 2,631 genes were uniquely associated with the outcome and used in the study after removing the first correlated genes.

We tuned the 3 ML models using the transcriptomic, clinical and clinico-transcriptomic data, respectively (**Figure 2**). During the tuning of RF and RBF-SVM models, the transcriptomic data alone and clinic-transcriptomic data produced similar accuracy heatmaps, while the MLP and linear SVM models had different accuracy metrics for the three different datasets. The best accuracy appeared to be present in the models using transcriptomic data, reaching to 0.528 in RF model, 0.608 in MPL model, 0.581 in RBF SVM model and 0.583 in linear SVM model.

We computed performance metrics of the three ML and Mlogic models based on the 5-fold cross validation, including accuracy, precision, recall/sensitivity, and F1 (**Table 2**). We also

Table 1. Baseline characteristics of the included patients with lung adenocarcinoma in the TCGA

	Alive no progression, %	Alive with disease, %	Dead with disease, %	Dead with no known disease, %	All, %
n	252	76	107	80	515
Sex					
Female	55.56	51.32	57.94	45.00	53.79
Male	44.44	48.68	42.06	55.00	46.21
Race					
Black	11.11	9.21	11.21	6.25	10.10
Other	13.49	19.74	8.41	21.25	14.56
White	75.40	71.05	80.37	72.50	75.34
Age (65+ yr)					
No	47.22	34.21	42.06	37.50	42.72
Yes	52.78	65.79	57.94	62.50	57.28
pT category					
T1	42.46	25.00	22.43	23.75	32.82
T2	48.81	61.84	61.68	55.00	54.37
T3	5.95	13.16	12.15	11.25	9.13
T4	2.78	0.00	3.74	10.00	3.69
pN category					
N0	73.41	75.00	48.60	46.25	64.27
N1	13.10	13.16	32.71	22.50	18.64
N2	10.32	9.21	18.69	26.25	14.37
N3	0.40	1.32	0.00	0.00	0.39
NX	2.78	1.32	0.00	5.00	2.33
pM category					
M0	65.48	63.16	73.83	67.50	67.18
M1	2.78	3.95	4.67	12.50	4.85
MX	31.75	32.89	21.50	20.00	27.96
Kras gene analysis					
No	50.79	31.58	55.14	47.50	48.35
Yes	11.90	13.16	13.08	10.00	12.04
Not Available	37.30	55.26	31.78	42.50	39.61
Kras mutation presence					
No	7.14	11.84	7.48	5.00	7.57
Yes	5.16	1.32	4.67	5.00	4.47
Not Available	87.70	86.84	87.85	90.00	87.96
Alk translocation presence					
No	42.86	22.37	50.47	40.00	40.97
Yes	5.95	6.58	9.35	5.00	6.60
Not Available	51.19	71.05	40.19	55.00	52.43
ECOG score					
0	19.05	18.42	23.36	7.50	18.06
1	16.67	21.05	19.63	26.25	19.42
2	3.97	5.26	2.80	5.00	4.08
3	0.40	0.00	0.93	1.25	0.58
Not Available	59.92	55.26	53.27	60.00	57.86
Radiation treatment, adjuvant					
No	31.35	26.32	21.50	25.00	27.57
Yes	1.19	1.32	6.54	2.50	2.52

Not Available	67.46	72.37	71.96	72.50	69.90
Targeted molecular therapy					
None given	26.59	17.11	21.50	17.50	22.72
Given	5.95	10.53	5.61	10.00	7.18
Not Available	67.46	72.37	72.90	72.50	70.10
Surgery					
No	30.16	25.00	17.76	18.75	25.05
Yes	69.84	75.00	82.24	81.25	74.95
History of other cancer					
No	82.94	80.26	83.18	81.25	82.33
Yes	17.06	19.74	16.82	18.75	17.67
Smoking history					
No	31.35	28.95	33.64	35.00	32.04
Yes	68.65	71.05	66.36	65.00	67.96

TCGA, the cancer genome atlas; ECOG, Eastern Cooperative Oncology Group.

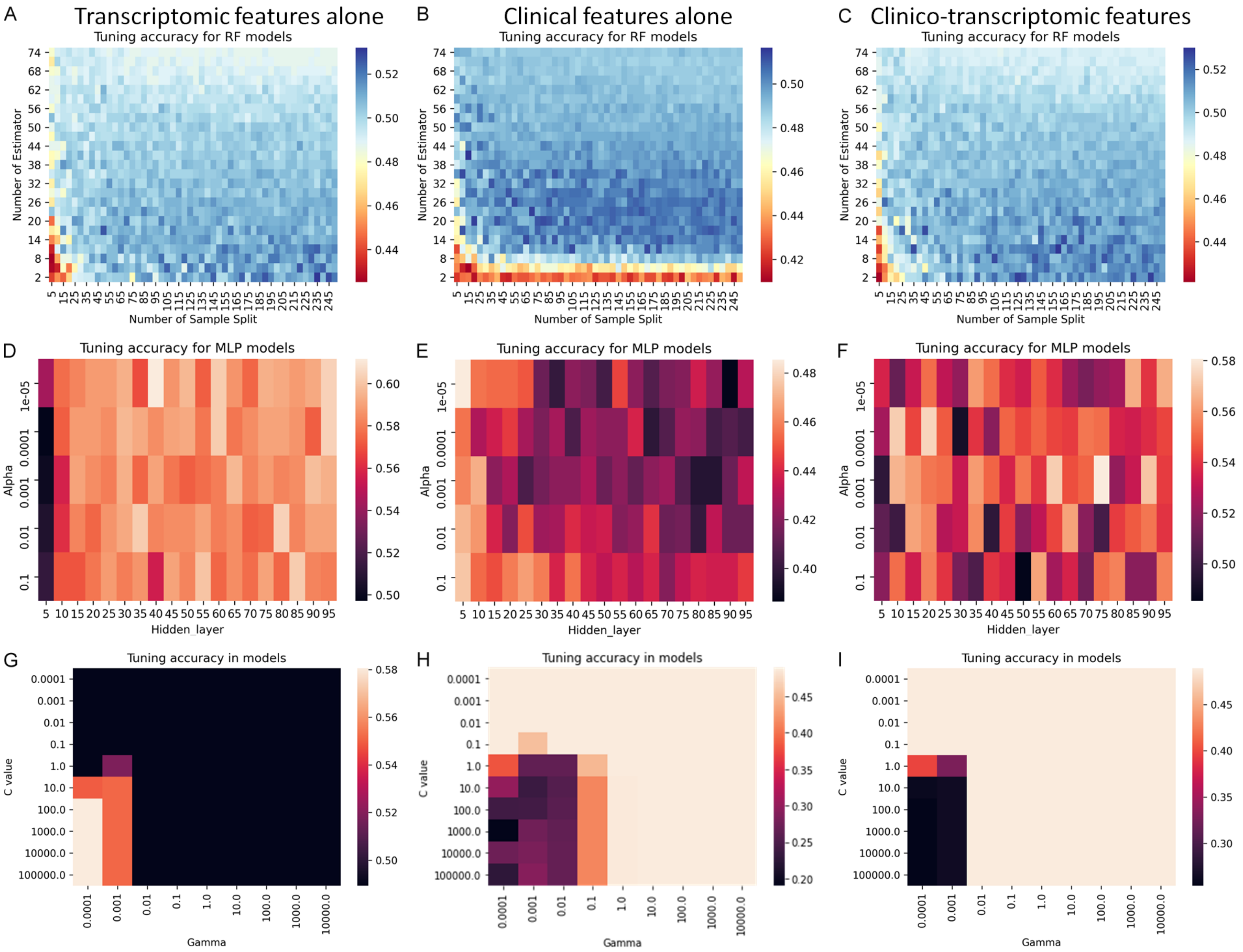
included the linear SVC model, whose kernel was different from the linear SVM model. For transcriptomic data, Mlogit had the best accuracy and linear SVC had the best recall. For clinical data, linear SVM and RBF SVM both had the best accuracy, while MLP had the best recall. For clinico-transcriptomic data, linear SVC and linear SVM had the best accuracy, and linear SVC had the best recall.

To more reliably assess the performance of these models, we used one over the rest classifier to generate receiver operator characteristics curve, and computed the AUC (**Figure 3**, next page). For transcriptomic and clinico-transcriptomic data, the Mlogit, linear SVM and MLP models all achieved a micro-average AUC of 0.82, while the RF model only reached the AUC of 0.75. For clinical data set, the RF model reached a micro-average AUC of 0.69, which was slightly higher than those of Mlogit and MLP models, but still lower than most of the AUC produced using transcriptomic or clinic-transcriptomic data.

There were 1340 genes (positively) linked to alive without progression, 1304 genes (positively) linked to alive with disease, 135 genes (positively) linked to dead with no known disease, and 1298 genes (positively) linked to dead with disease according to the Mlogit model (**Table 3**). To identify the genes that were important for classifying the 4-category outcome, we compared the top-ranked genes by their positive and inverse associations with 4-category outcome. Nine of the top 25 ranked genes in the alive without disease group were also in the list of the bottom-25 ranked genes in other outcome-groups, including *NDUFS5, P2RY2, PRPF18, CCL24, ZNF813, MYL6, FLJ41941, POU5F1B,* and *SUV420H1* in ranking order. Similarly, *BDKRB2, LOC100133738 (TERC), DNAJA3, MRPL15, SLC16A13, CRHBP* and *ACSBG2* in the alive with progression group were bottom-25 ranked genes in the other groups, so were *GAL3ST3, AD2, RAB41, HDC*, and *PLEKHG1* in the dead with disease group. Interestingly, only *FBXO15, IPMK,* and *PCDHB8* of the top-25 ranked genes in the death with no known disease were cross-linked to the bottom-25 ranked genes in other groups. The GSEA revealed the pathways, disease/clinical presentations and GO that were enriched in the 4 outcome-groups (**Figure 4**), as well as the related genes (Supplementary Figure 1A and 1B).

Discussion

We here compared the performances of the 3 ML and Mlogit models for classifying the 4-category survival outcomes of lung adenocarcinoma patients using the TCGA data. We found that Mlogit model overall performed similar to the 3 ML models (micro-average AUC=0.82). Surprisingly, transcriptomic data alone appeared to be sufficient to successfully classify the 4-category outcome in lung adenocarcinoma patients, and more useful than clinical data alone for the classification. We also identified a set of genes that were associated with one of the 4-category outcomes, and inversely linked to other outcomes.


A
Transcriptomic features alone
Tuning accuracy for RF models
Number of Estimator
74 68 62 56 50 44 38 32 26 20 14 8 2
Number of Sample Split
5 15 25 35 45 55 65 75 85 95 105 115 125 135 145 155 165 175 185 195 205 215 225 235 245
0.52 0.50 0.48 0.46 0.44
B
Clinical features alone
Tuning accuracy for RF models
Number of Estimator
Number of Sample Split
0.50 0.48 0.46 0.44 0.42
C
Clinico-transcriptomic features
Tuning accuracy for RF models
Number of Estimator
Number of Sample Split
0.52 0.50 0.48 0.46 0.44
D
Tuning accuracy for MLP models
Alpha
1e-05 0.0001 0.001 0.01 0.1
Hidden_layer
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
0.60 0.58 0.56 0.54 0.52 0.50
E
Tuning accuracy for MLP models
Alpha
Hidden_layer
0.48 0.46 0.44 0.42 0.40
F
Tuning accuracy for MLP models
Alpha
Hidden_layer
0.58 0.56 0.54 0.52 0.50
G
Tuning accuracy in models
C value
0.0001 0.001 0.01 0.1 1.0 10.0 100.0 1000.0 10000.0 100000.0
Gamma
0.0001 0.001 0.01 0.1 1.0 10.0 100.0 1000.0 10000.0
0.58 0.56 0.54 0.52 0.50
H
Tuning accuracy in models
C value
Gamma
0.45 0.40 0.35 0.30 0.25 0.20
I
Tuning accuracy in models
C value
Gamma
0.45 0.40 0.35 0.30

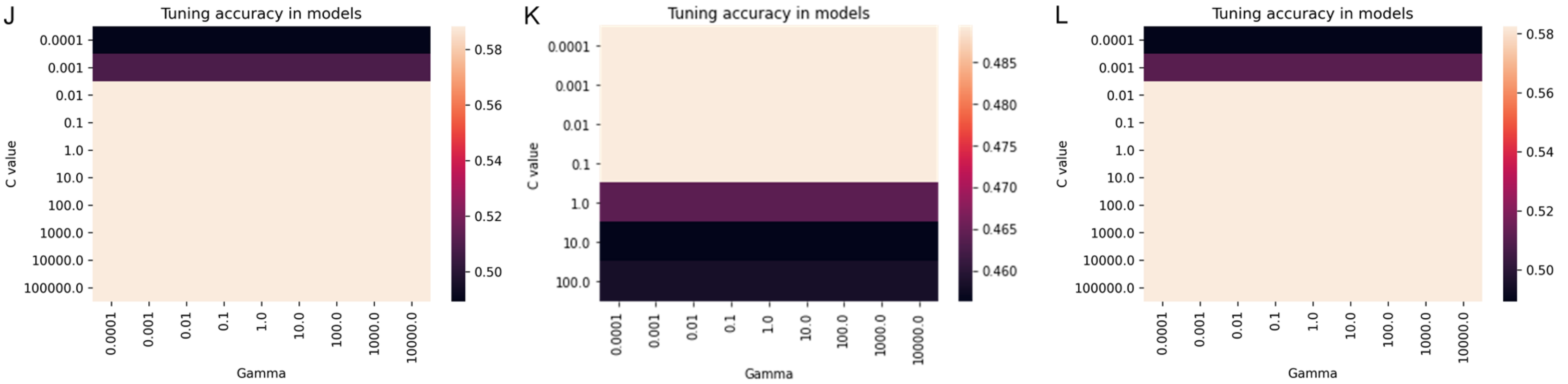


Figure 2. Tuning of the machine learning algorithms to classify the 4-category survival outcome. We tuned the random forests (RF), support vector machine (SVM) with linear or radial basis function kernel, and multilayer perceptron (MLP) to classify the 4-category outcomes using 5-fold cross-validation (Heatmap graphs: A-C, RF models; D-F, MLP models; G-I, radial basis function SVM; and J-L, linear SVM models). The left column was the data produced using transcriptomic features alone, middle column using clinical features alone, and the right column using clinico-transcriptomic features (merged dichotomized clinical and transcriptomic features).

Table 2. Performance of machine learning and multinomial logistic regression models in classifying 4-category survival of TCGA lung adenocarcinoma patients using transcriptomic, clinical or clinico-transcriptomic data

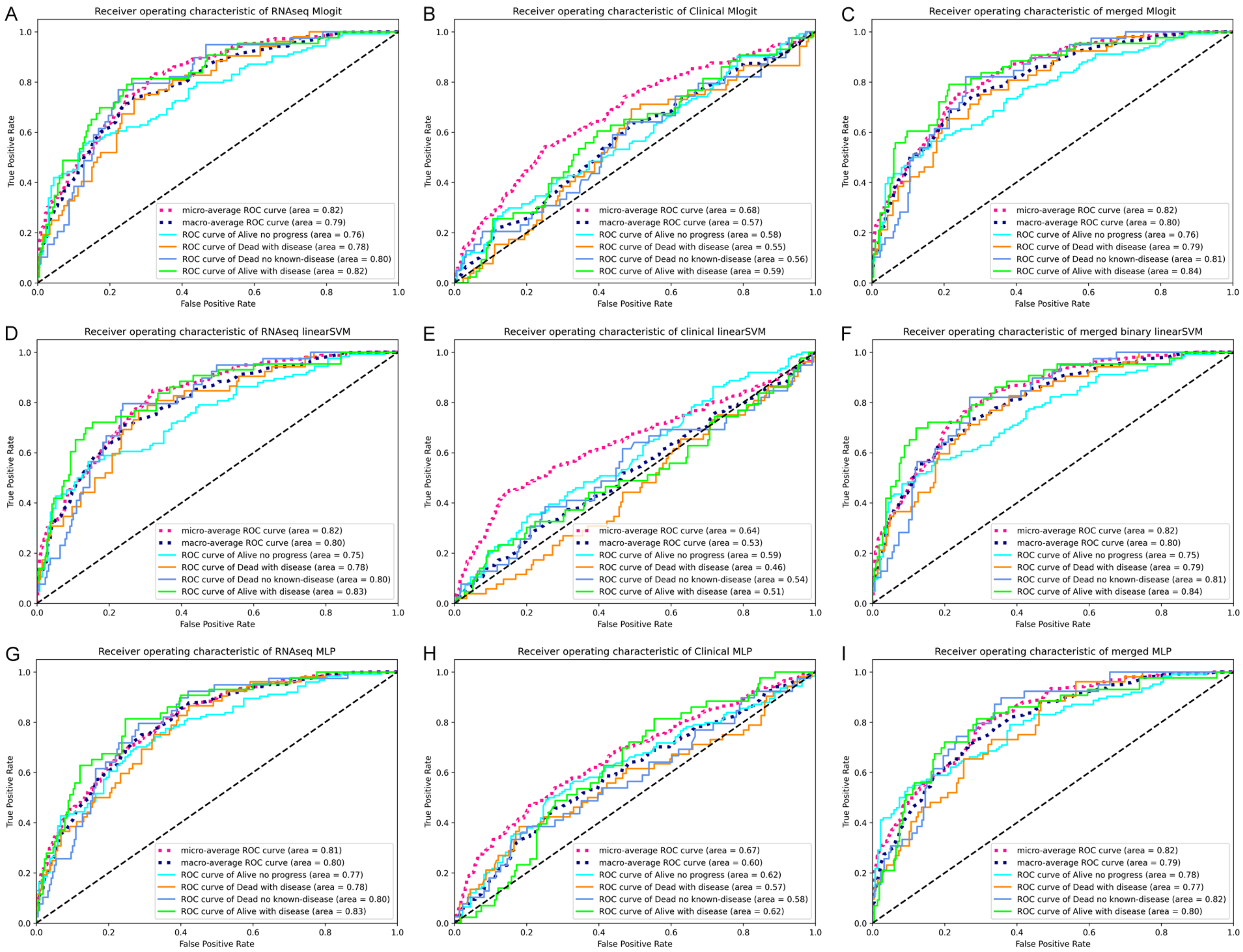
Model	Accuracy	Precision	Recall/sensitivity	F1
Transcriptomic data				
Mlogit	0.592±0.041	0.586±0.067	0.487±0.037	0.508±0.043
Random Forest	0.489±0.005	0.122±0.001	0.250±0.000	0.164±0.001
MLP	0.573±0.055	0.578±0.075	0.454±0.045	0.473±0.049
Linear SVC	0.577±0.043	0.544±0.056	0.490±0.052	0.502±0.055
Linear SVM	0.588±0.045	0.604±0.088	0.465±0.046	0.487±0.055
RBF SVM	0.581±0.041	0.614±0.096	0.448±0.046	0.470±0.059
Clinical data				
Mlogit	0.433±0.051	0.210±0.065	0.250±0.042	0.212±0.046
Random Forest	0.497±0.007	0.273±0.123	0.259±0.009	0.183±0.017
MLP	0.476±0.011	0.215±0.037	0.263±0.017	0.206±0.023
Linear SVC	0.441±0.033	0.174±0.030	0.241±0.023	0.191±0.022
Linear SVM	0.489±0.005	0.122±0.001	0.250±0.000	0.164±0.001
RBF SVM	0.489±0.005	0.122±0.001	0.250±0.000	0.164±0.001
Clinico-transcriptomic data				
Mlogit	0.573±0.049	0.582±0.084	0.476±0.046	0.496±0.046
Random Forest	0.489±0.005	0.122±0.001	0.250±0.000	0.164±0.001
MLP	0.546±0.071	0.599±0.104	0.443±0.098	0.445±0.096
Linear SVC	0.583±0.040	0.580±0.070	0.490±0.047	0.508±0.045
Linear SVM	0.583±0.040	0.608±0.094	0.460±0.046	0.475±0.053
RBF SVM	0.489±0.005	0.122±0.001	0.250±0.000	0.164±0.001

Note: Data presented as mean ± standard deviation from 5 cross-validation study. TCGA, the cancer genome atlas; Mlogit, multinomial logistic regression; MLP, multiple layer perceptron; SVC, support vector classifier; SVM, support vector machine; RBF, Radial basis function.

The major strength of this study is the first-time classification of the multicategory outcome in lung adenocarcinoma using ML models. The past studies have used ML and other models to classify the causes of death in lung adenocarcinoma patients, but only for the binary survival-outcomes [5, 10, 11, 16-21]. The in-depth outcome-classification will help select the patients who might need only lower doses of chemotherapy or radiotherapy, and increase treatment dosages or use other treatment modalities in the patients who died of cancer. Indeed, three genes were linked to death with disease and inversely linked to alive without disease, while 5 genes were linked to the alive without disease, and inversely linked to death with disease. These genes in our view could help classify cancer-recurrence risks for more effective prevention of cancer deaths and de-escalation of treatment intensity.

Moreover, in contrary to the common understanding that ML helps multilabel prediction, we for the first time showed that ML models were not all superior to Mlogit model in this sample set of 515 lung cancer cases and 2631 differential expressed features. This finding is consistent with some of the previous ML studies [28, 29], but in contrast with others on cancer-outcomes [22, 30, 31]. We believe that the differences might be attributable to the sample size of ~500, although we thought that the large number of features may benefit from ML algorithms, but it did not. Therefore, we recommend to compare the performances of ML models with Mlogit or other conventional statistical models. However, the potential overfitting of Mlogit regression model must also be noted and carefully evaluated in an external validation set. On the other hand, linear SVM and MLP models seemed very useful in classifying multilabel data in this study, but the RF model did not perform well likely due to the small sample size and feature composition.

Furthermore, transcriptomic data alone appear sufficient for classifying 4-category outcome in


A
Receiver operating characteristic of RNAseq Mlogit
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.79)
ROC curve of Alive no progress (area = 0.76)
ROC curve of Dead with disease (area = 0.78)
ROC curve of Dead no known-disease (area = 0.80)
ROC curve of Alive with disease (area = 0.82)
B
Receiver operating characteristic of Clinical Mlogit
micro-average ROC curve (area = 0.68)
macro-average ROC curve (area = 0.57)
ROC curve of Alive no progress (area = 0.58)
ROC curve of Dead with disease (area = 0.55)
ROC curve of Dead no known-disease (area = 0.56)
ROC curve of Alive with disease (area = 0.59)
C
Receiver operating characteristic of merged Mlogit
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.80)
ROC curve of Alive no progress (area = 0.76)
ROC curve of Dead with disease (area = 0.79)
ROC curve of Dead no known-disease (area = 0.81)
ROC curve of Alive with disease (area = 0.84)
D
Receiver operating characteristic of RNAseq linearSVM
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.80)
ROC curve of Alive no progress (area = 0.75)
ROC curve of Dead with disease (area = 0.78)
ROC curve of Dead no known-disease (area = 0.80)
ROC curve of Alive with disease (area = 0.83)
E
Receiver operating characteristic of clinical linearSVM
micro-average ROC curve (area = 0.64)
macro-average ROC curve (area = 0.53)
ROC curve of Alive no progress (area = 0.59)
ROC curve of Dead with disease (area = 0.46)
ROC curve of Dead no known-disease (area = 0.54)
ROC curve of Alive with disease (area = 0.51)
F
Receiver operating characteristic of merged binary linearSVM
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.80)
ROC curve of Alive no progress (area = 0.75)
ROC curve of Dead with disease (area = 0.79)
ROC curve of Dead no known-disease (area = 0.81)
ROC curve of Alive with disease (area = 0.84)
G
Receiver operating characteristic of RNAseq MLP
micro-average ROC curve (area = 0.81)
macro-average ROC curve (area = 0.80)
ROC curve of Alive no progress (area = 0.77)
ROC curve of Dead with disease (area = 0.78)
ROC curve of Dead no known-disease (area = 0.80)
ROC curve of Alive with disease (area = 0.83)
H
Receiver operating characteristic of Clinical MLP
micro-average ROC curve (area = 0.67)
macro-average ROC curve (area = 0.60)
ROC curve of Alive no progress (area = 0.62)
ROC curve of Dead with disease (area = 0.57)
ROC curve of Dead no known-disease (area = 0.58)
ROC curve of Alive with disease (area = 0.62)
I
Receiver operating characteristic of merged MLP
micro-average ROC curve (area = 0.82)
macro-average ROC curve (area = 0.79)
ROC curve of Alive no progress (area = 0.78)
ROC curve of Dead with disease (area = 0.77)
ROC curve of Dead no known-disease (area = 0.82)
ROC curve of Alive with disease (area = 0.80)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

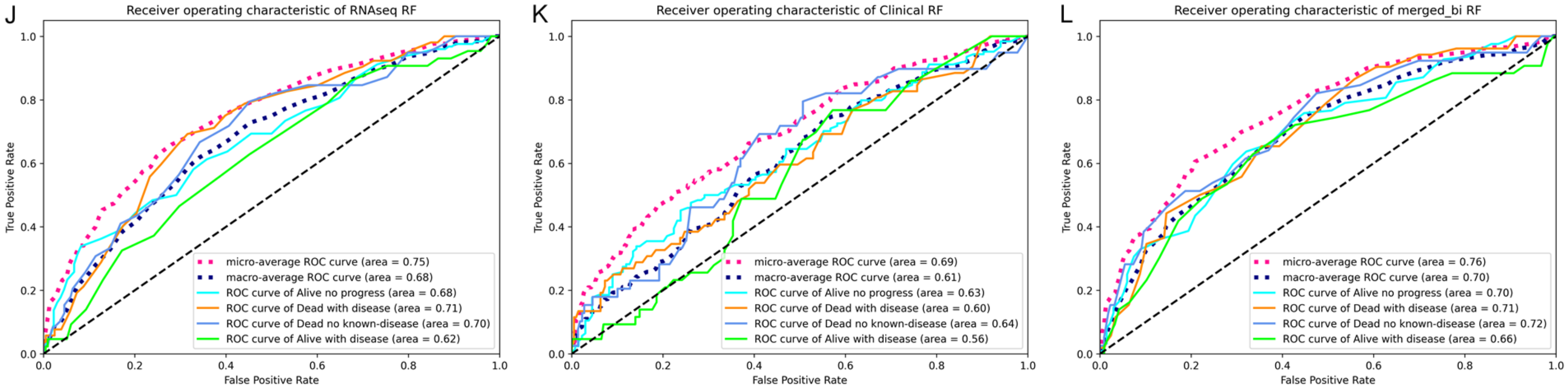


Figure 3. Receiver operator characteristics curves and the areas under the curve of the final models. Largely based on the tuning data, we chose the final models of multinomial logistic regression (Mlogit), linear support vector machine (SVM), multilayer perceptron (MLP) and random forests (RF) models to classify the 4-category outcomes. The operator characteristics curves and the areas under the curve were produced using 5-fold cross-validation and the OneVsRestClassifier function. (A-C, Mlogit model; D-F, linear SVM models; G-I, MLP models; and J-L, RF models). The left column was the data produced using transcriptomic features alone, middle column using clinical features alone, and the right column using clinico-transcriptomic features (merged dichotomized clinical and transcriptomic features). The micro-average was the calculated metrics globally by counting the total true positives, false negatives and false positives. The macro-average was the calculated metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

Table 3. The genes of top- and bottom-ranked coefficients for their associations with the 4-category outcomes in the multinomial logistic regression model

Class 1 (1340 positive)		Class 2 (1304 positive)		Class 3 (135 positive)		Class 4 (1298 positive)	
ID	coefficient	ID	coefficient	ID	coefficient	ID	coefficient
Top 25 genes							
NDUFS5	0.461	*LOC389613*	0.356	*FBXO15*	0.398	*TMEM8C*	0.390
P2RY2	0.428	*BDKRB2*	0.291	*THAP10*	0.394	*GAL3ST3*	0.377
PRPF18	0.418	*NFATC4*	0.290	*REP15*	0.351	*HSPA2*	0.374
C19ORF57	0.411	*LOC100133738*#	0.289	*SHISA7*	0.340	*AD2*	0.363
OR11H6	0.406	*DBNDD2*	0.277	*OSIL*	0.338	*D12S53E*	0.356
CCL24	0.406	*GCKR*	0.256	*SQLE*	0.337	*OTOA*	0.346
ZNF813	0.395	*LOC654185*	0.254	*IPMK*	0.336	*CCDC155*	0.344
COX14	0.390	*LIPT2*	0.253	*PCDHGA11*	0.311	*RAB41*	0.334
MYL6	0.389	*KBF2*	0.249	*LRRC66*	0.308	*PYY2*	0.326
FLJ41941	0.367	*FOPNL*	0.247	*TRMT10B*	0.306	*BMF*	0.318
LRRC24	0.364	*ASAH3*	0.245	*LOC92033*	0.304	*HIST1H3I*	0.303
DLG6	0.361	*HAR1A*	0.241	*TGIF1*	0.298	*LOC728613*	0.301
LOC125688	0.360	*DNAJA3*	0.238	*PCDHB8*	0.295	*MBL1P*	0.299
GPR143	0.346	*BDKRB1*	0.234	*FUT4*	0.287	*HDC*	0.293
POU5F1B	0.338	*SCML1*	0.233	*MOV10*	0.285	*TMED4*	0.285
GSTA5	0.334	*LINC00662*	0.233	*EVL*	0.277	*ACTRT3*	0.282
MIR4697HG	0.328	*MRPL15*	0.232	*TGM4*	0.270	*FLJ45829*	0.276
CD86	0.324	*LIMCH1*	0.232	*HCN2*	0.269	*HNRNPA0*	0.275
ADCK5	0.324	*SLC16A13*	0.231	*TCN1*	0.263	*KLRAP1*	0.273
IBM3	0.322	*PPP1R7*	0.230	*NOX4*	0.261	*IL13*	0.265
LOC286359	0.319	*LIMS3-LOC440895*	0.230	*EPGN*	0.260	*SPATA33*	0.262
FGF20	0.318	*CRHBP*	0.228	*MIC2Y*	0.258	*CLDN20*	0.261
SUV420H1	0.317	*RRP7A*	0.227	*FGFBP1*	0.258	*LMNB2*	0.258
FNBP1	0.316	*ARSG*	0.226	*C15orf6*	0.249	*PLEKHG1*	0.256
SNF8	0.315	*ACSBG2*	0.225	*ELP5*	0.243	*LOC100131869*	0.254
Bottom 25 genes							
PPP1R2P3	-0.402	*SLC1A7*	-0.342	*AMTN*	-0.356	*SUV420H1*	-0.438
RNPEPL1	-0.393	*ID4*	-0.337	*CCBL2*	-0.304	*CCL24*	-0.425
OR52N2	-0.387	*PCSK1N*	-0.309	*MYL6*	-0.289	*SPINK1*	-0.379
SLC16A13	-0.380	*CRH*	-0.307	*NDUFS5*	-0.281	*AKR1B11*	-0.362
LINC00518	-0.361	*MAPK1IP1L*	-0.289	*MTMR8*	-0.274	*FAM222A-AS1*	-0.341
IPMK	-0.358	*LOC284940*	-0.286	*ACSBG2*	-0.272	*P2RY2*	-0.322
PCK1	-0.338	*LOC155254*	-0.281	*SHISA4*	-0.269	*HCG18*	-0.320
SH2D3A	-0.336	*ABCC6P1*	-0.276	*TEX14*	-0.267	*KAPPA-200*	-0.314
GAL3ST3	-0.335	*MIA*	-0.274	*LOC102724786*	-0.266	*ZNF79*	-0.306
ATP4B	-0.332	*HBA2*	-0.269	*CCDC22*	-0.266	*MIR4697HG*	-0.301
MECT1	-0.331	*SLC14A2*	-0.268	*LOC100133738*	-0.263	*MRE11A*	-0.300
AD2	-0.327	*C17ORF51*	-0.250	*MRPL15*	-0.261	*DNALI1*	-0.300
C1ORF105	-0.326	*PCDHB8*	-0.249	*HTSS*	-0.257	*PRPF18*	-0.293
TMEM165	-0.324	*HOXB9*	-0.248	*ARRDC3-AS1*	-0.250	*ARIH2*	-0.288
INCA1	-0.323	*LRRC24*	-0.246	*LOC115830*	-0.249	*LOC100421021*	-0.285
RAB41	-0.320	*POU5F1B*	-0.243	*CRTAC1*	-0.247	*HSD17B3*	-0.285
REM2	-0.319	*RFX1*	-0.241	*COASY*	-0.246	*S100A1*	-0.284
FBXO15	-0.318	*CLIC5*	-0.240	*BDKRB2*	-0.245	*MLLT10*	-0.282
SEMA3A	-0.317	*HNRNPA3P1*	-0.234	*PLEKHG1*	-0.240	*CENPV*	-0.280
PRO1873	-0.316	*LMAN1*	-0.233	*HIST1H4J*	-0.240	*KCTD3*	-0.277
CACNG3	-0.316	*ANKRD55*	-0.233	*LOC732239*	-0.239	*SLC25A14*	-0.275
DNAJA3	-0.316	*KCNF1*	-0.231	*OS9*	-0.238	*TMEM126A*	-0.275

CRIP1	-0.316	*TENM1*	-0.231	*CHRNA1*	-0.237	*AK9*	-0.274
CRHBP	-0.314	*C9ORF40*	-0.227	*LOC641367*	-0.236	*KIF12*	-0.272
NUP62CL	-0.312	*FAH*	-0.224	*ARSF*	-0.233	*KLRG1*	-0.267

Note: Duplicated genes are highlighted by the classes which they were (positively) associated with. Class 1, Alive no progression; class 2, Alive with disease; class 3, Dead with no known disease; class 4, Dead with disease. A total of 2631 genes were subject to the analyses; #, the new gene ID is *TERC*.

this study, and more useful than the clinical data alone. To our surprise, we were not able to identify synergistic effects of combining the clinical and transcriptomic data, and the models using only clinical data performed poorly. It is noteworthy that clinical data could be used to reach a reasonably good prediction accuracy using large population-based datasets [25, 30] and others [32-34]. The difference may be attributable to the availability of more targeted therapies and better understanding of molecular aspects of lung adenocarcinoma than prostate cancer [35]. Indeed, the guidelines for non-small cell carcinoma of the lung, including lung adenocarcinoma, recently expanded the already-long list of targeted- and immunotherapies [35].

In addition, few of the previous ML studies on transcriptomic data of lung adenocarcinoma used cross-validation approach [26], while some used small-size external validation cohorts [9-11, 19]. We here used k-fold (k=5 in our case) cross-validation to increase rigor of our study. Briefly, in the k-fold cross-validation, one randomly and evenly splits the samples into k-portions, and after shuffling conducts k rounds of modelling using the k-1 portions as the training set and the last one portion as the test set. It has been used to effectively measure the performance and validate findings of large datasets [22, 25, 36, 37]. The use of k-fold cross-validation and subsequent evaluation of performance metrics were methodologically rigorous for the randomized repeats. Thus, this approach as used in this study is particularly useful when external validation set is unavailable or not feasible.

Finally, we identified several genes and biological processes that were associated with or useful for classifying the 4-category outcomes of lung adenocarcinoma. For example, *P2RY2* is a gene important for lung fibrosis [38], and regulating the proliferation of lung carcinoma cells [39], but its roles in lung cancer development or progression are otherwise unknown. Additional studies on these cross-linked genes are needed, and might reveal novel targets of lung adenocarcinoma therapy. Another example is the genes that are linked to some clinical, socioeconomic and behavioral characteristics in the UK Biobank GWAS study such as smoking, and self-reported eye/eyelid problem, which have been shown important for all-cause mortality or other deaths [40, 41].

This study has several limitations. First, we could not validate our findings in another large-size transcriptomic data set with the 4-category outcome, which to our knowledge is yet available. Our internal 5-fold cross-validation approach may in part address this limitation. The cross-link to the large sample size study such as UK Biobank also confirmed some of our findings. Nonetheless, additional validation studies are needed. Second, the sample size of 500 is large for transcriptomic studies, but might be too small for optimal performance of some ML models. We would like to confirm our findings using another large transcriptomic dataset with the 4-category outcome. However, large size clinico-transcriptomic studies of multicategory outcomes are expensive to carry out and largely unavailable. Thus, additional transcriptomic studies are recommended to have a large sample-size and detailed clinical-outcomes such as specific causes of death. Finally, treatment data were included in the TCGA but might be limited or misinterpreted by the data collectors. Despite our slight concern in this regard, the TCGA data have been widely used for external validation, primary study or in silico (secondary) analysis [40, 42-45]. Nonetheless, caution should be used when applying our findings to patient care.

In summary, we here show that transcriptomic features alone could be used to accurately classify 4-category survival-outcome in the patients with lung adenocarcinoma using Mlogit regression or ML models such as linear SVM and MLP. These findings may help better classify these patients for choosing the right treatment options. Our findings also reveal several genes and pathways that are important for

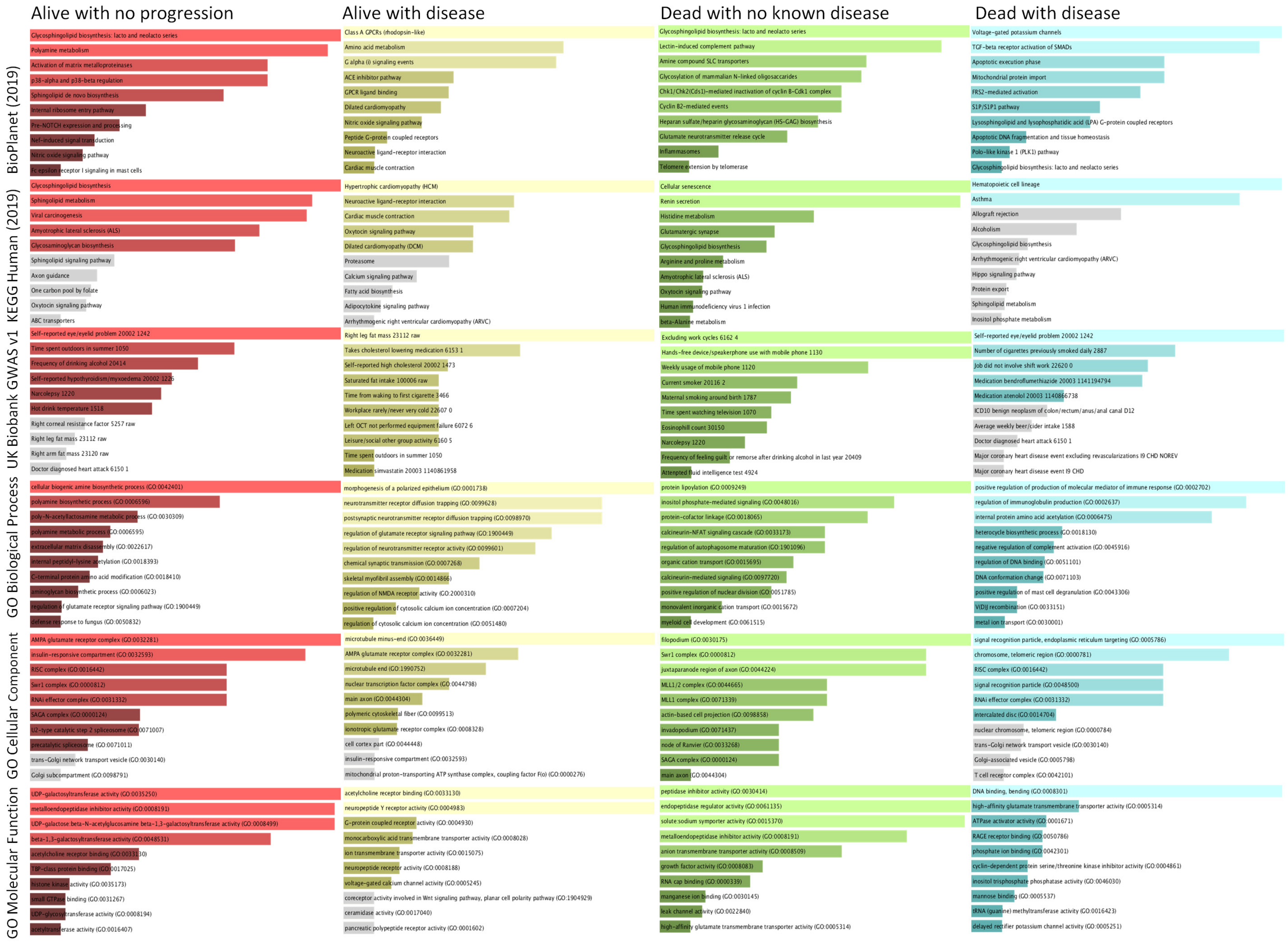

Alive with no progression
Alive with disease
Dead with no known disease
Dead with disease
BioPlanet (2019)
KEGG Human (2019)
UK Biobank GWAS v1
GO Biological Process
GO Cellular Component
GO Molecular Function
Glycosphingolipid biosynthesis: lacto and neolacto series
Polyamine metabolism
Activation of matrix metalloproteinases
p38-alpha and p38-beta regulation
Sphingolipid de novo biosynthesis
Internal ribosome entry pathway
Pre-NOTCH expression and processing
Nef-induced signal transduction
Nitric oxide signaling pathway
Fc epsilon receptor I signaling in mast cells
Glycosphingolipid biosynthesis
Sphingolipid metabolism
Viral carcinogenesis
Amyotrophic lateral sclerosis (ALS)
Glycosaminoglycan biosynthesis
Sphingolipid signaling pathway
Axon guidance
One carbon pool by folate
Oxytocin signaling pathway
ABC transporters
Self-reported eye/eyelid problem 20002 1242
Time spent outdoors in summer 1050
Frequency of drinking alcohol 20414
Self-reported hypothyroidism/myxoedema 20002 1226
Narcolepsy 1220
Hot drink temperature 1518
Right corneal resistance factor 5257 raw
Right leg fat mass 23112 raw
Right arm fat mass 23120 raw
Doctor diagnosed heart attack 6150 1
cellular biogenic amine biosynthetic process (GO:0042401)
polyamine biosynthetic process (GO:0006596)
poly-N-acetyllactosamine metabolic process (GO:0030309)
polyamine metabolic process (GO:0006595)
extracellular matrix disassembly (GO:0022617)
internal peptidyl-lysine acetylation (GO:0018393)
C-terminal protein amino acid modification (GO:0018410)
aminoglycan biosynthetic process (GO:0006023)
regulation of glutamate receptor signaling pathway (GO:1900449)
defense response to fungus (GO:0050832)
AMPA glutamate receptor complex (GO:0032281)
insulin-responsive compartment (GO:0032593)
RISC complex (GO:0016442)
Swr1 complex (GO:0000812)
RNAi effector complex (GO:0031332)
SAGA complex (GO:0000124)
U2-type catalytic step 2 spliceosome (GO:0071007)
precatalytic spliceosome (GO:0071011)
trans-Golgi network transport vesicle (GO:0030140)
Golgi subcompartment (GO:0098791)
UDP-galactosyltransferase activity (GO:0035250)
metalloendopeptidase inhibitor activity (GO:0008191)
UDP-galactose:beta-N-acetylglucosamine beta-1,3-galactosyltransferase activity (GO:0008499)
beta-1,3-galactosyltransferase activity (GO:0048531)
acetylcholine receptor binding (GO:0033130)
TBP-class protein binding (GO:0017025)
histone kinase activity (GO:0035173)
small GTPase binding (GO:0031267)
UDP-glycosyltransferase activity (GO:0008194)
acetyltransferase activity (GO:0016407)
Class A GPCRs (rhodopsin-like)
Amino acid metabolism
G alpha (i) signaling events
ACE inhibitor pathway
GPCR ligand binding
Dilated cardiomyopathy
Nitric oxide signaling pathway
Peptide G-protein coupled receptors
Neuroactive ligand-receptor interaction
Cardiac muscle contraction
Hypertrophic cardiomyopathy (HCM)
Neuroactive ligand-receptor interaction
Cardiac muscle contraction
Oxytocin signaling pathway
Dilated cardiomyopathy (DCM)
Proteasome
Calcium signaling pathway
Fatty acid biosynthesis
Adipocytokine signaling pathway
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Right leg fat mass 23112 raw
Takes cholesterol lowering medication 6153 1
Self-reported high cholesterol 20002 1473
Saturated fat intake 100006 raw
Time from waking to first cigarette 3466
Workplace rarely/never very cold 22607 0
Left OCT not performed equipment failure 6072 6
Leisure/social other group activity 6160 5
Time spent outdoors in summer 1050
Medication simvastatin 20003 1140861958
morphogenesis of a polarized epithelium (GO:0001738)
neurotransmitter receptor diffusion trapping (GO:0099628)
postsynaptic neurotransmitter receptor diffusion trapping (GO:0098970)
regulation of glutamate receptor signaling pathway (GO:1900449)
regulation of neurotransmitter receptor activity (GO:0099601)
chemical synaptic transmission (GO:0007268)
skeletal myofibril assembly (GO:0014866)
regulation of NMDA receptor activity (GO:2000310)
positive regulation of cytosolic calcium ion concentration (GO:0007204)
regulation of cytosolic calcium ion concentration (GO:0051480)
microtubule minus-end (GO:0036449)
AMPA glutamate receptor complex (GO:0032281)
microtubule end (GO:1990752)
nuclear transcription factor complex (GO:0044798)
main axon (GO:0044304)
polymeric cytoskeletal fiber (GO:0099513)
ionotropic glutamate receptor complex (GO:0008328)
cell cortex part (GO:0044448)
insulin-responsive compartment (GO:0032593)
mitochondrial proton-transporting ATP synthase complex, coupling factor F(o) (GO:0000276)
acetylcholine receptor binding (GO:0033130)
neuropeptide Y receptor activity (GO:0004983)
G-protein coupled receptor activity (GO:0004930)
monocarboxylic acid transmembrane transporter activity (GO:0008028)
ion transmembrane transporter activity (GO:0015075)
neuropeptide receptor activity (GO:0008188)
voltage-gated calcium channel activity (GO:0005245)
coreceptor activity involved in Wnt signaling pathway, planar cell polarity pathway (GO:1904929)
ceramidase activity (GO:0017040)
pancreatic polypeptide receptor activity (GO:0001602)
Glycosphingolipid biosynthesis: lacto and neolacto series
Lectin-induced complement pathway
Amine compound SLC transporters
Glycosylation of mammalian N-linked oligosaccarides
Chk1/Chk2(Cds1)-mediated inactivation of cyclin B-Cdk1 complex
Cyclin B2-mediated events
Heparan sulfate/heparin glycosaminoglycan (HS-GAG) biosynthesis
Glutamate neurotransmitter release cycle
Inflammasomes
Telomere extension by telomerase
Cellular senescence
Renin secretion
Histidine metabolism
Glutamatergic synapse
Glycosphingolipid biosynthesis
Arginine and proline metabolism
Amyotrophic lateral sclerosis (ALS)
Oxytocin signaling pathway
Human immunodeficiency virus 1 infection
beta-Alanine metabolism
Excluding work cycles 6162 4
Hands-free device/speakerphone use with mobile phone 1130
Weekly usage of mobile phone 1120
Current smoker 20116 2
Maternal smoking around birth 1787
Time spent watching television 1070
Eosinophill count 30150
Narcolepsy 1220
Frequency of feeling guilt or remorse after drinking alcohol in last year 20409
Attenpted fluid intelligence test 4924
protein lipoylation (GO:0009249)
inositol phosphate-mediated signaling (GO:0048016)
protein-cofactor linkage (GO:0018065)
calcineurin-NFAT signaling cascade (GO:0033173)
regulation of autophagosome maturation (GO:1901096)
organic cation transport (GO:0015695)
calcineurin-mediated signaling (GO:0097720)
positive regulation of nuclear division (GO:0051785)
monovalent inorganic cation transport (GO:0015672)
myeloid cell development (GO:0061515)
filopodium (GO:0030175)
Swr1 complex (GO:0000812)
juxtaparanode region of axon (GO:0044224)
MLL1/2 complex (GO:0044665)
MLL1 complex (GO:0071339)
actin-based cell projection (GO:0098858)
invadopodium (GO:0071437)
node of Ranvier (GO:0033268)
SAGA complex (GO:0000124)
main axon (GO:0044304)
peptidase inhibitor activity (GO:0030414)
endopeptidase regulator activity (GO:0061135)
solute:sodium symporter activity (GO:0015370)
metalloendopeptidase inhibitor activity (GO:0008191)
anion transmembrane transporter activity (GO:0008509)
growth factor activity (GO:0008083)
RNA cap binding (GO:0000339)
manganese ion binding (GO:0030145)
leak channel activity (GO:0022840)
high-affinity glutamate transmembrane transporter activity (GO:0005314)
Voltage-gated potassium channels
TGF-beta receptor activation of SMADs
Apoptotic execution phase
Mitochondrial protein import
FRS2-mediated activation
S1P/S1P1 pathway
Lysosphingolipid and lysophosphatidic acid (LPA) G-protein coupled receptors
Apoptotic DNA fragmentation and tissue homeostasis
Polo-like kinase 1 (PLK1) pathway
Glycosphingolipid biosynthesis: lacto and neolacto series
Hematopoietic cell lineage
Asthma
Allograft rejection
Alcoholism
Glycosphingolipid biosynthesis
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Hippo signaling pathway
Protein export
Sphingolipid metabolism
Inositol phosphate metabolism
Self-reported eye/eyelid problem 20002 1242
Number of cigarettes previously smoked daily 2887
Job did not involve shift work 22620 0
Medication bendroflumethiazide 20003 1141194794
Medication atenolol 20003 1140866738
ICD10 benign neoplasm of colon/rectum/anus/anal canal D12
Average weekly beer/cider intake 1588
Doctor diagnosed heart attack 6150 1
Major coronary heart disease event excluding revascularizations I9 CHD NOREV
Major coronary heart disease event I9 CHD
positive regulation of production of molecular mediator of immune response (GO:0002702)
regulation of immunoglobulin production (GO:0002637)
internal protein amino acid acetylation (GO:0006475)
heterocycle biosynthetic process (GO:0018130)
negative regulation of complement activation (GO:0045916)
regulation of DNA binding (GO:0051101)
DNA conformation change (GO:0071103)
positive regulation of mast cell degranulation (GO:0043306)
V(D)J recombination (GO:0033151)
metal ion transport (GO:0030001)
signal recognition particle, endoplasmic reticulum targeting (GO:0005786)
chromosome, telomeric region (GO:0000781)
RISC complex (GO:0016442)
signal recognition particle (GO:0048500)
RNAi effector complex (GO:0031332)
intercalated disc (GO:0014704)
nuclear chromosome, telomeric region (GO:0000784)
trans-Golgi network transport vesicle (GO:0030140)
Golgi-associated vesicle (GO:0005798)
T cell receptor complex (GO:0042101)
DNA binding, bending (GO:0008301)
high-affinity glutamate transmembrane transporter activity (GO:0005314)
ATPase activator activity (GO:0001671)
RAGE receptor binding (GO:0050786)
phosphate ion binding (GO:0042301)
cyclin-dependent protein serine/threonine kinase inhibitor activity (GO:0004861)
inositol trisphosphate phosphatase activity (GO:0046030)
mannose binding (GO:0005537)
tRNA (guanine) methyltransferase activity (GO:0016423)
delayed rectifier potassium channel activity (GO:0005251)

Figure 4. Gene set enrichment analyses on the top ranked genes based on the coefficient of the multinomial logistic regression model. We conducted the gene-set enrichment analyses for the top-ranked genes based on their associations with respective outcome/groups (coefficient as the metric), using the web-based Enrichr algorithm and *p*-value based ranking. The length of the bar indicates the degree of the gene-enrichment. The top two rows show the results of pathway analyses using Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioPlanet (2019) algorithms, the third row shows the result of disease-related analysis using algorithm of the UK Biobank genome-wide association study (GWAS) v1, and the bottom 3 rows show the result of analyses using the gene ontologies (GO) algorithms. The far-left column was the alive with no progression group, middle-left column the alive with disease, the middle-right column the dead with no known disease, and the far-right column the dead with disease group, respectively. The lighter shade indicates P<0.01 for gene-enrichment, darker shade indicates P<0.05, and gray shade indicates P≥0.05.

the different, specific survival-outcome in these patients, and their potential biological significance. Additional studies are warranted to confirm and understand our findings.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Lanjing Zhang, Department of Pathology, Princeton Medical Center, 1 Plainsboro Rd, Plainsboro, NJ 08536, USA. Tel: 609-853-6833; Fax: 609-853-6841; E-mail: lanjing.zhang@rutgers.edu; ljzhang@hotmail.com

References

[1] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020; 70: 7-30.

[2] Hu X, Lin Y, Qin G and Zhang L. Underlying causes of death with changing mortality among adults in the United States, 2013-2017.

[3] Zaorsky NG, Churilla TM, Egleston BL, Fisher SG, Ridge JA, Horwitz EM and Meyer JE. Causes of death among cancer patients. Ann Oncol 2017; 28: 400-407.

[4] Shang G, Jin Y, Zheng Q, Shen X, Yang M, Li Y and Zhang L. Histology and oncogenic driver alterations of lung adenocarcinoma in Chinese. Am J Cancer Res 2019; 9: 1212-1223.

[5] Yu J, Hu Y, Xu Y, Wang J, Kuang J, Zhang W, Shao J, Guo D and Wang Y. LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features. BMC Cancer 2019; 19: 263.

[6] Gomez-Rueda H, Martinez-Ledesma E, Martinez-Torteya A, Palacios-Corona R and Trevino V. Integration and comparison of different genomic data for outcome prediction in cancer. BioData Min 2015; 8: 32.

[7] Kim D, Shin H, Song YS and Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. J Biomed Inform 2012; 45: 1191-1198.

[8] Fernandez FG, Force SD, Pickens A, Kilgo PD, Luu T and Miller DL. Impact of laterality on early and late survival after pneumonectomy. Ann Thorac Surg 2011; 92: 244-249.

[9] Xue L, Bi G, Zhan C, Zhang Y, Yuan Y and Fan H. Development and validation of a 12-gene immune relevant prognostic signature for lung adenocarcinoma through machine learning strategies. Front Oncol 2020; 10: 835.

[10] Ma B, Geng Y, Meng F, Yan G and Song F. Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. J Cancer 2020; 11: 1288-1298.

[11] Kim JS, Chun SH, Park S, Lee S, Kim SE, Hong JH, Kang K, Ko YH and Ahn YH. Identification of novel microRNA prognostic markers using cascaded Wx, a neural network-based framework, in lung adenocarcinoma patients. Cancers (Basel) 2020; 12: 1890.

[12] Bao X, Shi R, Zhao T and Wang Y. Immune landscape and a novel immunotherapy-related gene signature associated with clinical outcome in early-stage lung adenocarcinoma. J Mol Med (Berl) 2020; 98: 805-818.

[13] Zhang J, Shi K, Huang W, Weng W, Zhang Z, Guo Y, Deng T, Xiang Y, Ni X, Chen B and Zhou M. The DNA methylation profile of non-coding RNAs improves prognosis prediction for pancreatic adenocarcinoma. Cancer Cell Int 2019; 19: 107.

[14] Wang Y, Zhang Q, Gao Z, Xin S, Zhao Y, Zhang K, Shi R and Bao X. A novel 4-gene signature for overall survival prediction in lung adenocarcinoma patients with lymph node metastasis. Cancer Cell Int 2019; 19: 100.

[15] Cho HJ, Lee S, Ji YG and Lee DH. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PLoS One 2018; 13: e0207204.

[16] Xu D, Zhang J, Xu H, Zhang Y, Chen W, Gao R and Dehmer M. Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. BMC Genomics 2020; 21: 650.

[17] Rana HK, Akhtar MR, Islam MB, Ahmed MB, Lio P, Huq F, Quinn JMW and Moni MA. Machine learning and bioinformatics models to

identify pathways that mediate influences of welding fumes on cancer progression. Sci Rep 2020; 10: 2795.

[18] Kweon S, Lee JH, Lee Y and Park YR. Personal health information inference using machine learning on RNA expression data from patients with cancer: algorithm validation study. J Med Internet Res 2020; 22: e18387.

[19] Wang Y, Fu J, Wang Z, Lv Z, Fan Z and Lei T. Screening key lncRNAs for human lung adenocarcinoma based on machine learning and weighted gene co-expression network analysis. Cancer Biomark 2019; 25: 313-324.

[20] Wang Y, Deng H, Xin S, Zhang K, Shi R and Bao X. Prognostic and predictive value of three DNA methylation signatures in lung adenocarcinoma. Front Genet 2019; 10: 349.

[21] R K and R GR. Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed gaussian classifier system. J Med Syst 2019; 43: 201.

[22] Deng F, Zhou H, Lin Y, Heim J, Shen L, Li Y and Zhang L. Predict multicategory causes of death in lung cancer patients using clinicopathologic factors. medRxiv 2020; 2020.2009.2025.20201095.

[23] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C and Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013; 6: pl1.

[24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R and Dubourg V. Scikit-learn: machine learning in Python. J Machine Learning Res 2011; 12: 2825-2830.

[25] Deng F, Huang J, Yuan X, Cheng C and Zhang L. Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. bioRxiv 2020; 2020.2009.2013.295592.

[26] Herrmann M, Probst P, Hornung R, Jurinovic V and Boulesteix AL. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform 2020; [Epub ahead of print].

[27] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW and Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016; 44: W90-7.

[28] van der Ploeg T, Austin PC and Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014; 14: 137.

[29] Kim S, Park T and Kon M. Cancer survival classification using integrated data sets and intermediate information. Artif Intell Med 2014; 62: 23-31.

[30] Wang J, Deng F, Zeng F, Shanahan AJ, Li WV and Zhang L. Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. Am J Cancer Res 2020; 10: 1344-1355.

[31] Hanson HA, Martin C, O'Neil B, Leiser CL, Mayer EN, Smith KR and Lowrance WT. The relative importance of race compared to health care and social factors in predicting prostate cancer mortality: a random forest approach. J Urol 2019; 202: 1209-1216.

[32] Bartholomai JA and Frieboes HB. Lung cancer survival prediction via machine learning regression, classification, and statistical techniques. Proc IEEE Int Symp Signal Proc Inf Tech 2018; 2018: 632-637.

[33] Lynch CM, van Berkel VH and Frieboes HB. Application of unsupervised analysis techniques to lung cancer patient data. PLoS One 2017; 12: e0184370.

[34] Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, van Berkel VH and Frieboes HB. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform 2017; 108: 1-8.

[35] Ettinger DS, Wood DE, Aggarwal C, Aisner DL, Akerley W, Bauman JR, Bharat A, Bruno DS, Chang JY, Chirieac LR, D'Amico TA, Dilling TJ, Dobelbower M, Gettinger S, Govindan R, Gubens MA, Hennon M, Horn L, Lackner RP, Lanuti M, Leal TA, Lin J, Loo BW Jr, Martins RG, Otterson GA, Patel SP, Reckamp KL, Riely GJ, Schild SE, Shapiro TA, Stevenson J, Swanson SJ, Tauer KW, Yang SC and Gregory K; OCN, Hughes M. NCCN guidelines insights: non-small cell lung cancer, version 1.2020. J Natl Compr Canc Netw 2019; 17: 1464-1472.

[36] Hussain L, Ahmed A, Saeed S, Rathore S, Awan IA, Shah SA, Majid A, Idris A and Awan AA. Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. Cancer Biomark 2018; 21: 393-413.

[37] Agranoff D, Fernandez-Reyes D, Papadopoulos MC, Rojas SA, Herbster M, Loosemore A, Tarelli E, Sheldon J, Schwenk A, Pollok R, Rayner CF and Krishna S. Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Lancet 2006; 368: 1012-1021.

[38] Muller T, Fay S, Vieira RP, Karmouty-Quintana H, Cicko S, Ayata K, Zissel G, Goldmann T, Lungarella G, Ferrari D, Di Virgilio F, Robaye B, Boeynaems JM, Blackburn MR and Idzko M. The

purinergic receptor subtype P2Y2 mediates chemotaxis of neutrophils and fibroblasts in fibrotic lung disease. Oncotarget 2017; 8: 35962-35972.

[39] Schafer R, Sedehizade F, Welte T and Reiser G. ATP- and UTP-activated P2Y receptors differently regulate proliferation of human lung epithelial tumor cells. Am J Physiol Lung Cell Mol Physiol 2003; 285: L376-385.

[40] Li D, Yang W, Zhang Y, Yang JY, Guan R, Xu D and Yang MQ. Genomic analyses based on pulmonary adenocarcinoma in situ reveal early lung cancer signature. BMC Med Genomics 2018; 11: 106.

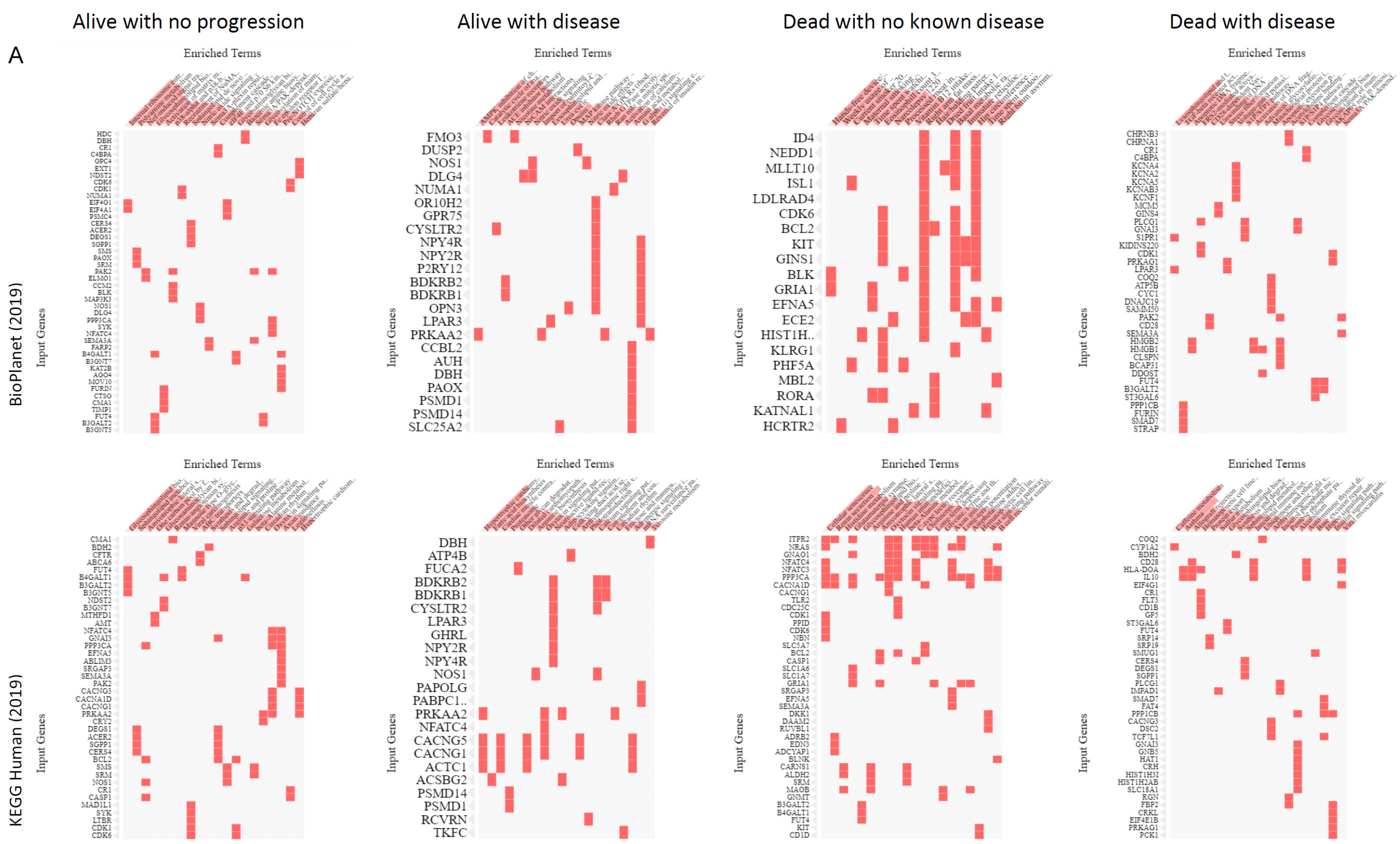
[41] Gomez SL, Chang ET, Shema SJ, Fish K, Sison JD, Reynolds P, Clément-Duchêne C, Wrensch MR, Wiencke JL and Wakelee HA. Survival following non-small cell lung cancer among Asian/Pacific Islander, Latina, and Non-Hispanic white women who have never smoked. Cancer Epidemiol Biomarkers Prev 2011; 20: 545-554.

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

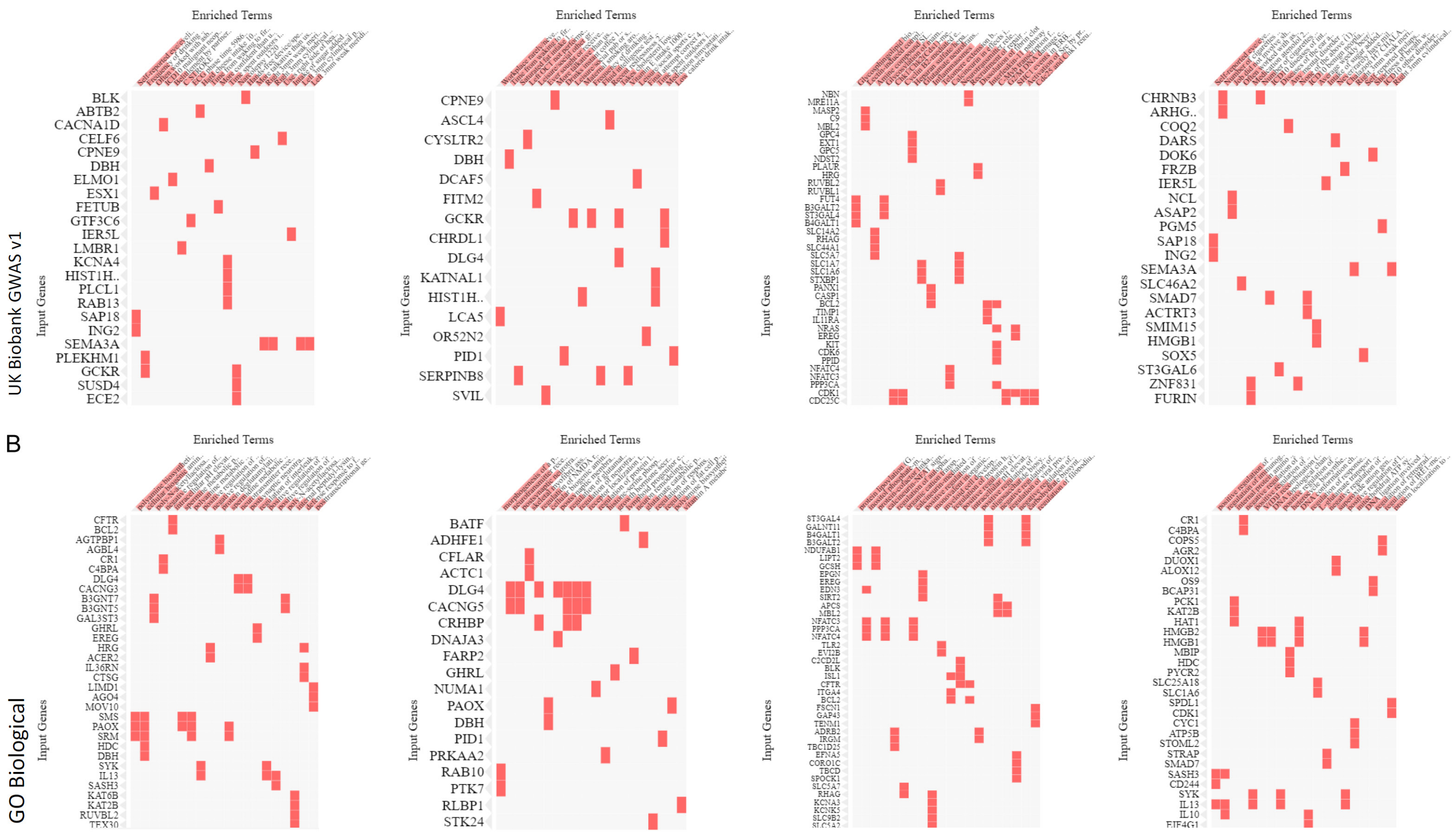
[43] Li Y, Ge D, Gu J, Xu F, Zhu Q and Lu C. A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies. BMC Cancer 2019; 19: 886.

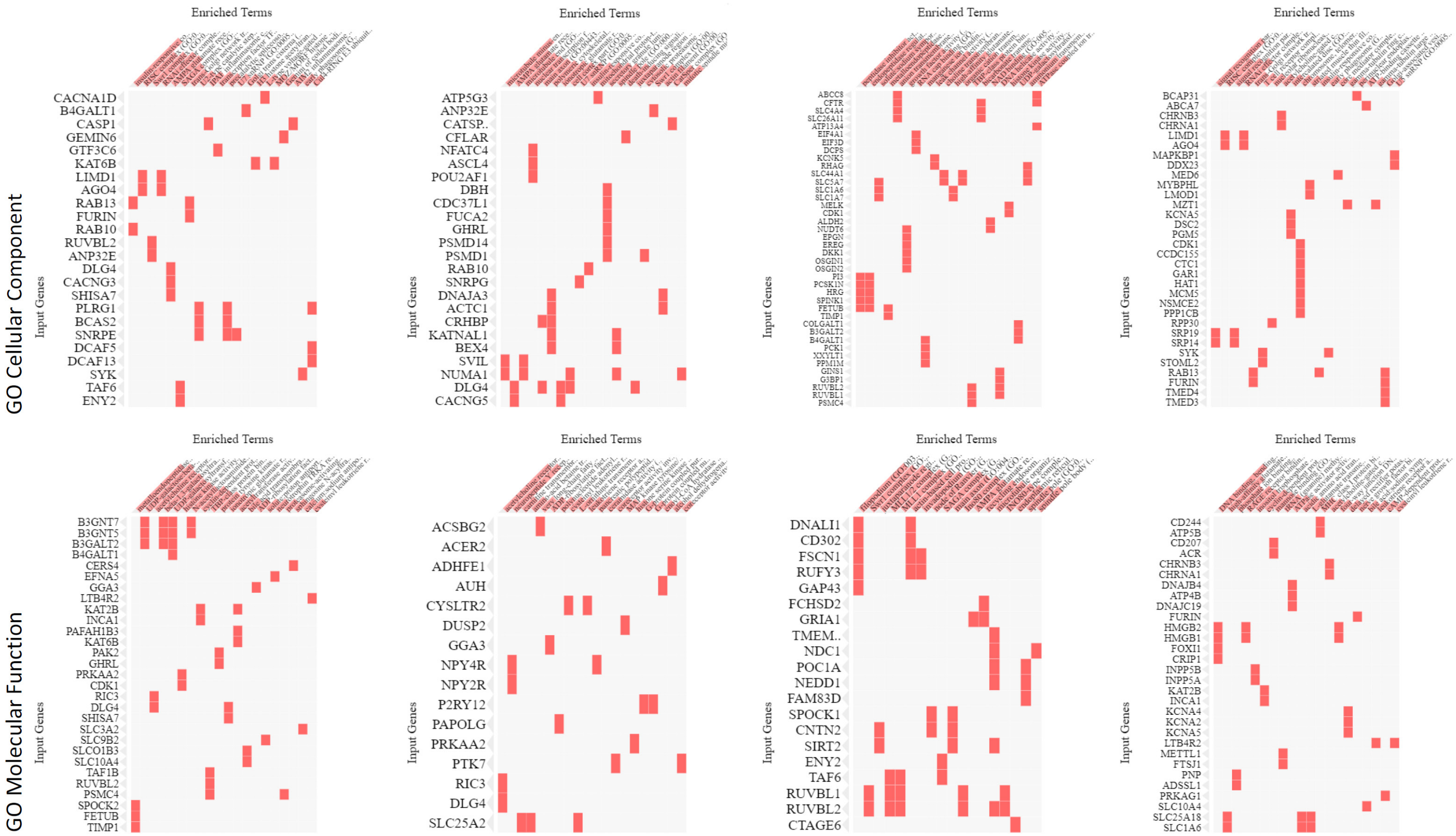
[44] Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K, Hou J, Zhang H, Yi S, Jafari M, Lin D, Chung C, Caughey BA, Li G, Dhar D, Shi W, Zheng L, Hou R, Zhu J, Zhao L, Fu X, Zhang E, Zhang C, Zhu JK, Karin M, Xu RH and Zhang K. DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci U S A 2017; 114: 7414-7419.

[45] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV and Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015; 13: 8-17.


A
Alive with no progression
Alive with disease
Dead with no known disease
Dead with disease
BioPlanet (2019)
KEGG Human (2019)
Enriched Terms
Input Genes
FMO3
DUSP2
NOS1
DLG4
NUMA1
OR10H2
GPR75
CYSLTR2
NPY4R
NPY2R
P2RY12
BDKRB2
BDKRB1
OPN3
LPAR3
PRKAA2
CCBL2
AUH
DBH
PAOX
PSMD1
PSMD14
SLC25A2
ID4
NEDD1
MLLT10
ISL1
LDLRAD4
CDK6
BCL2
KIT
GINS1
BLK
GRIA1
EFNA5
ECE2
HIST1H..
KLRG1
PHF5A
MBL2
RORA
KATNAL1
HCRTR2
DBH
ATP4B
FUCA2
BDKRB2
BDKRB1
CYSLTR2
LPAR3
GHRL
NPY2R
NPY4R
NOS1
PAPOLG
PABPC1..
PRKAA2
NFATC4
CACNG5
CACNG1
ACTC1
ACSBG2
PSMD14
PSMD1
RCVRN
TKFC

Machine learning for multicategory outcome of lung cancer





Supplementary Figure 1. The genes enriched in the specific terms of pathways and disease, and gene ontologies, respectively. The enriched terms are the columns, input genes are the rows, and cells in each matrix indicate whether the gene was associated with a term. The length of the shade at the top of each column indicates the degree of gene-enrichment in that term.