Contents lists available at ScienceDirect

Translational Oncology

journal homepage: www.elsevier.com/locate/tranon

Early lung cancer diagnostic biomarker discovery by machine learning methods

Ying Xie[a,1], Wei-Yu Meng[a,1], Run-Ze Li[a,1], Yu-Wei Wang[a], Xin Qian[b], Chang Chan[b], Zhi-Fang Yu[b], Xing-Xing Fan[a], Hu-Dan Pan[a], Chun Xie[a], Qi-Biao Wu[a], Pei-Yu Yan[a], Liang Liu[a], Yi-Jun Tang[b], Xiao-Jun Yao[a,*], Mei-Fang Wang[b,*], Elaine Lai-Han Leung[a,b,**]

[a] State Key Laboratory of Quality Research in Chinese Medicine/Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau (SAR), China

[b] Respiratory Medicine department of Taihe Hospital, Hubei University of Medicine, Hubei, China

ARTICLE INFO

Keywords:
Lung cancer
Metabolites
Biomarker
Early diagnosis
Machine learning

ABSTRACT

Early diagnosis has been proved to improve survival rate of lung cancer patients. The availability of blood-based screening could increase early lung cancer patient uptake. Our present study attempted to discover Chinese patients' plasma metabolites as diagnostic biomarkers for lung cancer. In this work, we use a pioneering interdisciplinary mechanism, which is firstly applied to lung cancer, to detect early lung cancer diagnostic biomarkers by combining metabolomics and machine learning methods. We collected total 110 lung cancer patients and 43 healthy individuals in our study. Levels of 61 plasma metabolites were from targeted metabolomic study using LC-MS/MS. A specific combination of six metabolic biomarkers note-worthily enabling the discrimination between stage I lung cancer patients and healthy individuals (AUC = 0.989, Sensitivity = 98.1%, Specificity = 100.0%). And the top 5 relative importance metabolic biomarkers developed by FCBF algorithm also could be potential screening biomarkers for early detection of lung cancer. Naïve Bayes is recommended as an exploitable tool for early lung tumor prediction. This research will provide strong support for the feasibility of blood-based screening, and bring a more accurate, quick and integrated application tool for early lung cancer diagnostic. The proposed interdisciplinary method could be adapted to other cancer beyond lung cancer.

Introduction

In worldwide, lung carcinoma is the leading cause of cancer death in the past few decades. In January 2019, National Central Cancer Registry of China (NCCRC) released its latest nationwide tumor statistics of population-based tumor registry data gathered from 368 tumor registries in 2015. According to this report, lung cancer ranks top for its incidence among malignant tumors in China. What's more, the 5-year survival rate for patients with lung tumors was low which is at 18%. However, the survival rate can increase to approximately 55% if early diagnosis of lung cancer was achieved. Moreover, it has been reported the early stage patients who received proper treatment could have a 5 year survival rate around 40% [1]. Unfortunately, over 70% patients are diagnosed when their tumor are developed to the advanced stages, and most of them are not suitable for receiving operation [2]. This is partly related with the early stage diagnostic methods which are still not sensitive and specific enough. Therefore, it appears to be an important step to find out the cogent and powerful diagnostic biomarkers of lung cancer, particularly for the diagnosis of early lung tumor progression.

Metabolomics study have been used to recognize the metabolic pathways and metabolites that regulate tumor progression and physiological function [3-5]. Metabolomics could provide the information of cellular metabolic processes that drive tumorigenesis and tumor progression. These metabolites also could be helpful for distinguishing the tumor stage, histological types, and even the response to drug treatment [6]. These changes in metabolite pattern had be used to evaluate the clinical characteristics of colorectal tumor [7], ovarian tumor [8], renal tumor [9], oral tumor [10], and pancreatic tumor [11]. Even so, for lung cancer, more specific and sensitive biomarkers were needed to be revealed with metabolic analysis.

[*] Corresponding authors.

[**] Corresponding authors at: State Key Laboratory of Quality Research in Chinese Medicine/Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau (SAR), China.

E-mail addresses: xjyao@must.edu.mo (X.-J. Yao), wmfpps02@hotmail.com (M.-F. Wang), lhleung@must.edu.mo (E.L.-H. Leung).

[1] These authors contribute equally.

https://doi.org/10.1016/j.tranon.2020.100907
Received 4 June 2020; Received in revised form 21 August 2020; Accepted 25 September 2020

Artificial intelligence (AI) is the competence for machines to imitate human behavior, which is extremely adept at handling extensive amounts of data. Machine learning is the application of AI, which allows computer systems could be trained automatically from experience without explicitly programmed [12]. Fundamentally, machine learning means learning from the practice of using algorithms to parse data, and then making a prediction or decision about the future situation of any new data sets [13]. In cancer, machine learning has already been used to explore survival and prognostic prediction models in pancreatic cancer, bladder cancer, advanced nasopharyngeal carcinoma and breast cancer [14-17]. In some cases, their performance had attained comparable to that of human experts [18]. Machine learning models could be seemed as an approach of designing the model by learning from experience and improving its performance [19]. These models aim at finding out effective variables and the relationship between them. Over the past few years, the field of AI has moved from largely theoretical studies to real-world applications [20,21]. The application of AI in several domains is now associated with great expectations and at the same time exists a great vacancy in cancer research especially lung cancer.

In this study, the major aim of metabolomics research on lung cancer was to discover clinical metabolic biomarkers that had representative alterations between lung tumor patients and healthy individuals. Moreover, we also focused on biomarkers for distinguishing each histological subtypes and disease stages, especially for early stages. For the first time, based on the plasma metabolite features, we applied machine learning to develop the diagnostic model for early stages of lung cancer.

Materials and methods

Patients and groups

A total of 110 patients and 43 healthy individuals of the Hubei Taihe Hospital were included in this study. The Institutional Review Board of Taihe Hospital, Hubei University of Medicine approved the study involving patients and healthy individuals. All individuals have written informed consent prior to participation in the investigation, with permission of sample collection, usage and data analysis. Final diagnosis was ascertained by clinical symptoms and histopathological examination of operative specimens. According to the TNM staging system, patients were classified as stage I ($n = 54$) stage II ($n = 31$), stage III ($n = 25$). Based on the WHO classification of tumors [22], the tumors have been classified as adenocarcinomas ($n = 63$), squamous carcinomas ($n = 41$) and other histological types ($n = 6$).

Targeted metabolomic study using LC-MS/MS

Targeted metabolomic study was performed with previous reported LC-MS methods [23,24]. In brief, plasma samples (200 μL) were thawed on ice, mixed with N-ethylmaleimide PBS buffer (10 mM, 200 μL) and 1000 μl of methanol containing 10 ng/ml internal standards (IS) Phe-d5.The mixed solution then incubated 20 mins at −20 °C and centrifuged at 13,000 rpm for 10 min at 4 °C. The supernatants were dried under nitrogen flow at 4 °C and reconstituted with 30% methanol for LC-MS analysis.

The chromatographic separation was carried out with a Waters X Bridge^{TM} BEH C18 analytical column (2.5 μm, 3.0 × 100 mm; Waters, Torrance, CA) using a Waters ACQUITY UPLC coupled with a 4000 Q-TRAP mass spectrometer. The mobile phase was composed of 0.1% formic acid water (solvent A) and methanol (solvent B) which was running in a gradient program: 0–3.0 min (0%–1% B); 3.0–10.0 min (1–3% B); 10.0–14.0 min (3–50% B); 14.0–18.0 min (50–95% B); 18.0–22.0 min (95–0% B); followed by a 3-min re-equilibration step. The flow rate was 0.6 ml/min, and 10 μl were injected in LC-MS. The mass spectra were acquired in both the negative and positive ion voltage modes for electrospray (ESI) with the following parameters: gas temperature, 450 °C; the ion spray voltage, ±4500 V; ion source gas 1 (nebulizer gas) 40 psi (N2); ion source gas 2 (auxiliary gas), 40 psi (N2); curtain gas: 20 psi. Targeted MS/MS (MRM) mode were used with the collision energy ranging from 10 V to 40 V. All LC-MS data were obtained by AB Analyst Software (Version 1.6.2). The intensity of each ion was normalized to the peak area of IS prior to multivariate statistical analysis. For the metabolomic assay, principal component analysis (PCA) and orthogonal projection to latent structures discriminant analysis (OPLS-DA) analysis were analyzed with SIMCA-p14 software (Umetrics AB, UMEÅ, Sweden) for tested groups. Variables were screened by VIP values first, and values exceeding 1 were considered eligible for group discrimination. The selected metabolites were further confirmed by a between normal control and disease control with P-value less than 0.05. MetaboAnalyst 3.0 was used for integrated analysis, pathway impact and enrichment.

Statistical analysis

All statistical data were analyzed by using SPSS Statistics 22.0 (SPSS, Chicago, IL, United States), GraphPad Prism 8.0 (GraphPad Software, La Jolla, CA, United States), TBtools v0.6735 [25] and Orange software 3.23 [26]. Receiver operating characteristic (ROC) curve analysis was established to evaluate the diagnostic performance of metabolites. P values < 0.05 was considered statistically significant.

Machine learning methods

The six machine learning techniques of K-nearest neighbor (KNN), Naïve Bayes, AdaBoost, Support Vector Machine (SVM), Random Forest, and Neural Network with 10-cross fold technique were used for the early lung tumor prediction based on the metabolomic biomarkers features. SVM, which is the classification algorithm, intends to invent a decision boundary between two categories that enables the prediction of labels from feature vectors [27]. K-nearest-neighbor (KNN) is the preferred selection when there is tiny prior knowledge of data, which is elementary and plain nonparametric method for classification [28]. Random Forest is an ensemble tree method that trees are grown by binary recursive splitting of right-censored data [29]. Naïve Bayes is a statistical classifier, which was used to predict class membership probability [30]. It hypothesizes all variables participate in classification independently and provides the result for prediction [31]. Neural Network purposes to simulate the neuron and human brain. The artificial neuron of Neural Network uses particular input features to assign suitable mathematical weights that are eventually able to predict some output object [32].

80% samples including stage I lung tumor patients ($n = 43$) and healthy individuals ($n = 35$) were selected by stratified sampling from each group as the training set to uniformly train the models. The rest 20% samples including stage I lung cancer patients ($n = 11$) and healthy individuals ($n = 8$) making up the test set were used for evaluation. The training set was used to generate the prediction model that could predict the diagnosis of the test set. To test and compare the models, sensitivity, specificity, precision, classification accuracy, and AUC (area under the curve) value from each model were used to measure the performance. The sensitivity, specificity, and classification accuracy values were obtained from true negative (TN), false negative (FN), true positive (TP), and false positive (FP). The following terms are essential information of them:

TP = Lung tumor patients correctly diagnosed as patients.
FP = Healthy individuals incorrectly identified as patients.
TN = Healthy individuals correctly identified as healthy.
FN = Lung tumor patients incorrectly identified as healthy.

Results

Metabolic biomarkers for detection of early lung tumor

Our studies aim to identify metabolites that could act as promising biomarkers for distinguishing lung tumor patients with healthy individ-

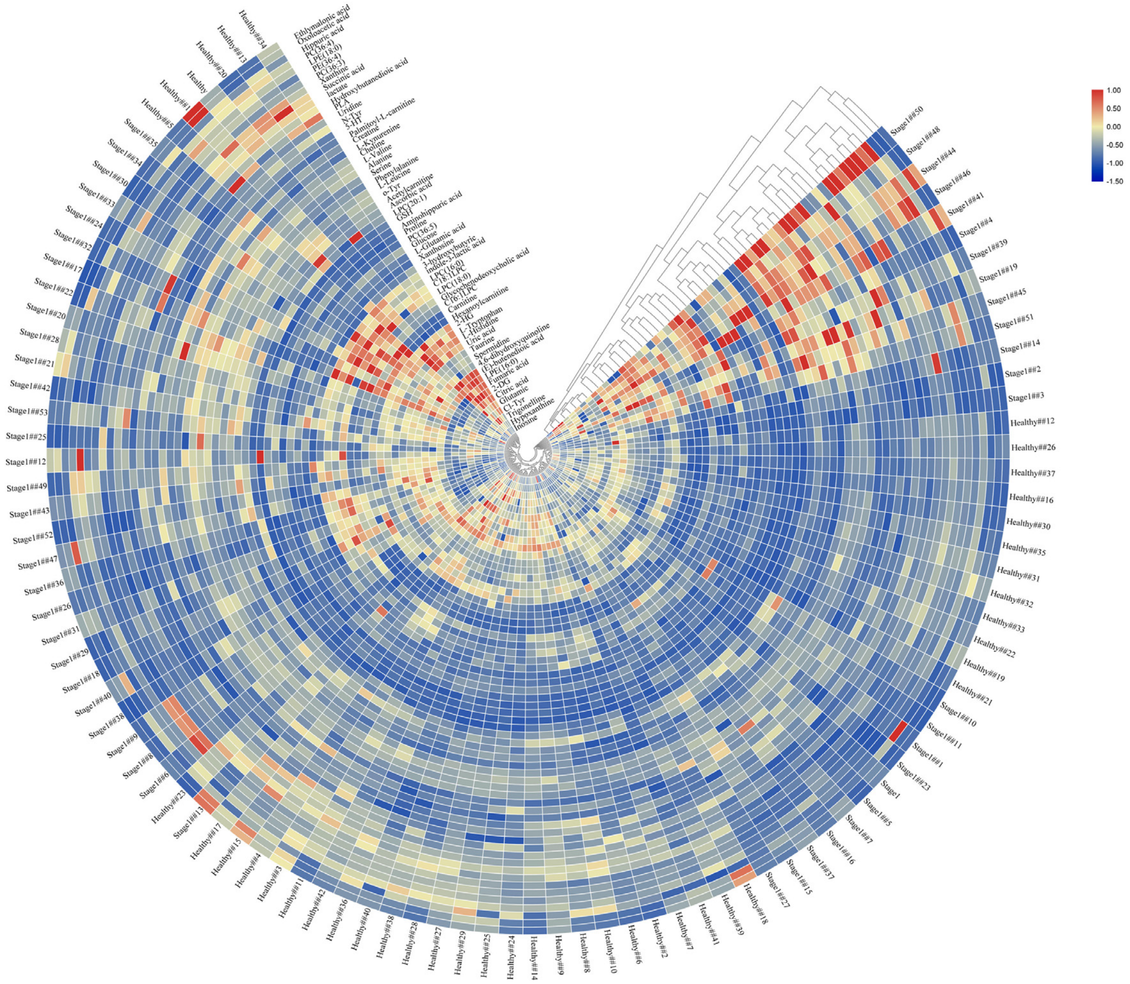

Fig. 1. Heatmap depicting the metabolomic biomarker levels of Stage I lung tumor patients ($n = 54$) and healthy people ($n = 43$). Stage I lung tumor patients and healthy people were grouped by hierarchical clustering of metabolomic biomarker levels.

uals, disease stages as well as pathological patterns with high sensitivity and specificity. Firstly, stage I lung cancer patients ($n = 54$) and healthy individuals ($n = 43$) were separated by using unsupervised hierarchical clustering with heat map shown in Fig. 1. Through Mann–Whitney U test, 46 influential metabolic biomarkers (Fig. 2A and Table S1) showed statistically significant difference (*p-value*<0.05) among 61 metabolites. Increased levels of L-Leucine, L-Valine, serine and other 25 metabolites were observed in stage I lung cancer patients compared to healthy individuals. Moreover, fumaric acid, citric acid, PC (36:4), and other 15 metabolites were downregulated in stage I lung cancer patients compared to healthy controls. These 46 influential metabolites are potential markers for pre-clinical screening of lung tumor.

Next, these 46 influential metabolic biomarkers were applied to construct ROC curves. Based on the AUC (area under the ROC curve) value, sensitivity and specificity, top 10 metabolic biomarkers with higher diagnostic value (AUC>0.800) were showed in Table 1 and Fig. S1. In addition, PCA (principal component analysis) with these 10 metabolites revealed a clear separation between stage I lung tumor patients and healthy individuals (Fig. 2B and Table S2). In the comparison between stage I lung tumor patients and healthy individuals, proline showed the best AUC value of 0.923 (95% CI: 0.871–0.975), with a sensitivity of 79.6% and specificity of 93.0% at the cut off value of 24.350.

Moreover, the potential combination schemes of metabolic biomarkers based on logistic regression analysis were carried out to enhance the sensitivity and accuracy of diagnostic of early stages of lung cancer. As shown in Table 1, Table S3 and Fig. 2C, the combination of six variables (metabolites) remarkably enhanced the AUC to 0.989 (95% CI: 0.967–1.000, Sensitivity = 98.1%, Specificity = 100.0%). The metabolites used included proline, L-kynurenine (AUC = 0.825, Sensitivity = 85.2%, Specificity = 72.1%), spermidine (AUC = 0.890, Sensitivity = 81.5%, Specificity = 90.7%), amino-hippuric acid (AUC = 0.811, Sensitivity = 68.5%, Specificity = 93.0%), palmitoyl-L-carnitine (AUC = 0.906, Sensitivity = 74.1%, Specificity = 100.0%) and taurine (AUC = 0.920, Sensitivity = 88.9%, Specificity = 95.3%). These results indicated that 6 metabolic biomarkers could act as a promising combination for early detection of lung tumor.

Metabolic biomarkers for disease progression

In addition, we were also interested in the alteration of metabolites in different stages. To identify the metabolites level changes with tumor stage progress, Kruskal–Wallis test was applied. Fig. 3A showed the 10 metabolites which showed significant difference in stage I ($n = 54$), stage II ($n = 31$), stage III ($n = 25$) lung tumor patients and healthy individuals ($n = 43$). Although there was statistically significant difference between lung tumor patients and healthy individuals (Fig. S2 and Table S4), the metabolic biomarkers showed poor performance for discrimination of stage I, II, and III patients with lung cancer (Fig. 3B–D, Table S5 and Table S6). Both of non-parametric test and receiver operating characteristic curves suggested the continuous abnormal expression in lung cancer patients compared with healthy individuals.

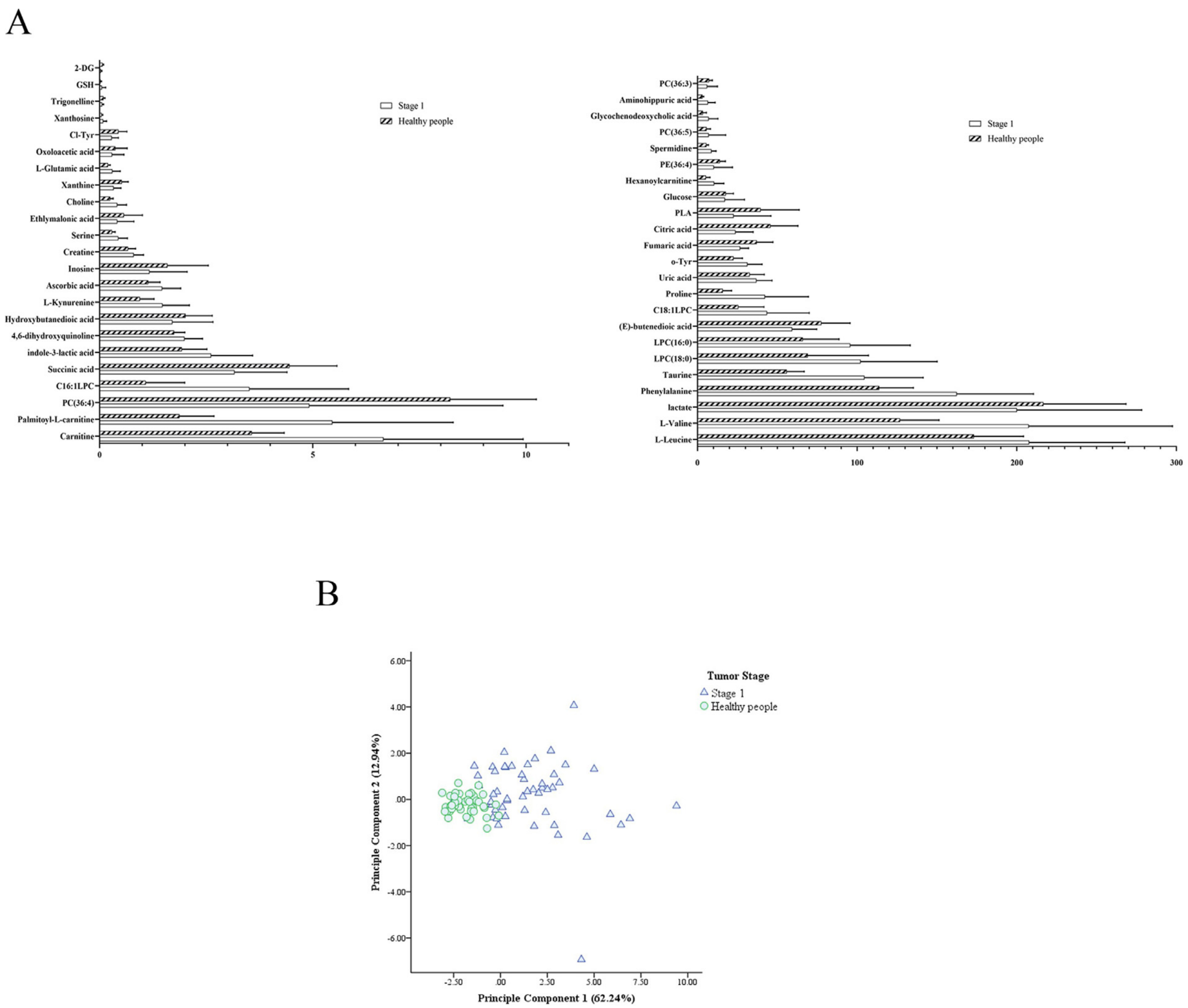

Fig. 2. Metabolic biomarkers for detection of early lung tumor and evaluation of different histological types. **(A)** 46 influential metabolomic biomarkers with statistical significance of Stage I lung tumor patients (mean value with SD). Through Mann–Whitney U test, 46 influential metabolic biomarkers showed statistically significant difference (*p*-value<0.05) among 61 metabolites. **(B)** PCA of 10 metabolomic biomarkers in early lung tumor detection. It revealed a clear separation between stage I lung tumor patients and healthy individuals. **(C)** ROC curve of metabolomic biomarkers and combined variates in early lung tumor detection. The combination of six variates included proline, L-kynurenine, spermidine, amino-hippuric acid, palmitoyl-L-carnitine and taurine. **(D)** ROC analysis of metabolomic biomarkers and combined variates of adenocarcinoma ($n = 63$) and squamous carcinoma ($n = 41$) patients. The combination of four variates included hypoxanthine, L-Kynurenine, proline and Carnitine. SD, standard deviation. PCA, principal component analysis. ROC, receiver operating characteristic.

Metabolic biomarkers for evaluation of different histological types

For lung cancer histological type prediction, particularly the distinction between squamous carcinoma and adenocarcinoma, it is a significant diagnostic requirement in clinical practice. Several clinical studies have demonstrated that tumor histological type differing toxicity and efficacy of treatment [33]. Tumor histological types identification will be helpful for improving the treatment efficiency. We applied the metabolites from the adenocarcinoma ($n = 63$) and squamous carcinoma ($n = 41$) patients to construct ROC curves and performed Mann–Whitney U test. There were only 2 influential metabolites identified between adenocarcinoma and squamous carcinoma, including hippuric acid (*p-value*=0.029) and hypoxanthine (*p-value*=0.017). In the ROC analysis (Fig. S3), hypoxanthine showed the AUC value of 0.639 (95% CI: 0.531–0.746), with a sensitivity of 69.8% and specificity of 56.1% at the cut off value of 0.092. And the hippuric acid showed the AUC value of 0.628 (95% CI: 0.519–0.737), with a sensitivity of 49.2% and specificity of 77.5% at the cut off value of 2.620. As shown in Fig. 2D and Table 1, the combination of four variates enhanced the AUC to 0.740 (95% CI: 0.644–0.837, sensitivity = 58.7%, specificity = 78.0%), including hypoxanthine, L-Kynurenine (AUC = 0.423, sensitivity = 77.8%, specificity = 24.4%), proline (AUC =0.580, sensitivity = 54.0%, specificity = 65.9%) and Carnitine (AUC = 0.536, sensitivity = 38.1%, specificity = 75.6%). These results indicated metabolic biomarkers showed poor performance on distinguishing different lung tumor histological types in our study, since all AUC values < 0.800 with poor sensitivity and specificity.

Utilization of machine learning methods

To develop the early lung tumor prediction model, we considered six machine learning techniques: K-nearest-neighbor (KNN), Naïve Bayes, AdaBoost, Support Vector Machine (SVM), Random Forest, and Neural Network (Fig. 4A). The training set of stage I lung tumor patients ($n = 43$) and healthy individuals ($n = 35$) was used to develop machine learning models based on the metabolic biomarker features. Using Orange 3.23 platform, there were 61 kinds of metabolites used as features to develop the machine learning models. To determine which model would provide the most precise predictions on the lung tumor metabolomics data, the sensitivity,

Table 1
ROC analysis of metabolomic biomarkers and combined variates.

	AUC	Std. error	Asymptotic 95% confidence interval		Optimal cut off	Sensitivity	Specificity	Youden index
			Lower bound	Upper bound				
ROC analysis of metabolomic biomarkers and combined variates in early lung tumor detection.								
L-Kynurenine	0.825	0.043	0.740	0.909	0.975	85.2%	72.1%	0.573
Proline	0.923	0.026	0.871	0.975	24.350	79.6%	93.0%	0.727
Spermidine	0.890	0.035	0.821	0.958	7.195	81.5%	90.7%	0.722
Amino-hippuric acid	0.811	0.045	0.722	0.900	4.035	68.5%	93.0%	0.615
Palmitoyl-L-carnitine	0.906	0.032	0.843	0.969	3.655	74.1%	100.0%	0.741
Taurine	0.920	0.032	0.856	0.983	71.300	88.9%	95.3%	0.842
Phenylalanine	0.848	0.038	0.774	0.922	125.500	79.6%	76.7%	0.564
L-Valine	0.876	0.036	0.806	0.946	167.000	68.5%	95.3%	0.639
o-Tyr	0.822	0.043	0.738	0.906	24.650	83.3%	72.1%	0.554
Carnitine	0.848	0.040	0.769	0.926	4.680	72.2%	93.0%	0.652
Combination of two	0.933	0.028	0.878	0.978	0.337	85.2%	93.0%	0.782
Combination of three	0.968	0.019	0.931	1.000	−0.147	94.4%	97.7%	0.921
Combination of six	0.989	0.011	0.967	1.000	−0.102	98.1%	100.0%	0.981
ROC analysis of metabolomic biomarkers and combined variates of adenocarcinoma and squamous carcinoma patients.								
L-Kynurenine	0.423	0.060	0.306	0.540	1.050	77.8%	24.4%	0.022
Proline	0.580	0.057	0.469	0.692	35.150	54.0%	65.9%	0.198
Carnitine	0.536	0.058	0.422	0.650	6.835	38.1%	75.6%	0.137
Hypoxanthine	0.639	0.055	0.531	0.746	0.092	69.8%	56.1%	0.259
Hippuric acid	0.628	0.056	0.519	0.737	2.620	49.2%	77.5%	0.267
Combination of four	0.740	0.049	0.644	0.837	0.556	58.7%	78.0%	0.368

Abbreviations: ROC,receiver operating characteristic; AUC, area under the curve.

Table 2
Machine learning models used for early lung tumor detection based on the metabolomic biomarker features.

		TP	FP	TN	FN	Classification accuracy	Sensitivity	Specificity	AUC	Precision
Training set	KNN	38	0	35	5	0.936	0.884	**1.000**	**1.000**	0.944
	SVM	43	2	33	0	0.974	**1.000**	0.943	**1.000**	0.975
	Random Forest	41	0	35	2	0.974	0.953	**1.000**	**1.000**	0.976
	Neural Network	43	0	35	0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Naïve Bayes	43	0	35	0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	AdaBoost	38	4	31	5	0.885	0.884	0.886	0.885	0.885
Test set	KNN	9	0	8	2	0.895	0.818	**1.000**	**1.000**	0.916
	SVM	10	0	8	1	0.947	0.909	**1.000**	**1.000**	0.953
	Random Forest	11	2	6	0	0.895	**1.000**	0.750	**1.000**	0.911
	Neural Network	10	0	8	1	0.947	0.909	**1.000**	**1.000**	0.953
	Naïve Bayes	11	0	8	0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	AdaBoost	4	0	8	7	0.632	0.364	**1.000**	0.682	0.804

Abbreviations: AdaBoost, Adaptive Boosting; SVM, support vector machines; KNN, k-nearest neighbor; TN, true negative; FN, false negative; TP, true positive; FP, false positive; AUC, area under the curve.

specificity, precision, classification accuracy, and AUC value of six machine learning models were assessed (Supplementary methods). With the best value in each evaluation is highlighted in Table 2. In training set, Naïve Bayes and Neural Network indicated better results in comparison with other techniques (KNN, AdaBoost, SVM, and Random Forest). The precision, classification accuracy, specificity, sensitivity and the AUC value of Naïve Bayes and Neural Network are 100.0%. AdaBoost machine learning technique showed poor performance (precision =0.885, classification accuracy=0.885, specificity=0.886, sensitivity=0.884, and AUC=0.885).

Subsequently, we used our trained diagnostic machine learning models to classify the test set that consisted of stage I lung cancer patients ($n = 11$) and healthy individuals ($n = 8$) to evaluate its performance. As shown in Table 2, the Naïve Bayes model showed the best performance on all evaluation parameters. The specificity of Naïve Bayes, Neural Network, KNN, AdaBoost, and SVM was 1.000, which means good prediction power against healthy individuals. For overall quality of prediction, AUC of Naïve Bayes, Neural Network, KNN, Random Forest, and SVM were 1.000. The sensitivity of Naïve Bayes and Random Forest was 1.000, SVM and Neural Network was 0.909, which means little FN (false negative) scale. In medical situation, FN scale is more important than FP (false positive) [34]. Consequently, these four models (Naïve Bayes, Random Forest, SVM and Neural Network) were appropriate for diagnosis of early lung tumor. In classification accuracy, Naïve Bayes model has a 1.000 rate with SVM and Neural Network models of 0.947. All of three models have enough accuracy that could be used for medical application. Naïve Bayes is a simple probabilistic classifier based on applying the Bayes' theorem with strong independence and normality assumptions between the variables [35]. It is one of the most valid machine learning algorithms with strong independence and normality assumptions between features, which has been widely employed for the prediction [36]. Given the above, our study testified that Naïve Bayes was the best model with the highest level of sensitivity, specificity, and accuracy. Therefore, Naïve Bayes is recommended as an exploitable tool for early lung tumor prediction.

The Fast Correlation-Based Filter (FCBF) algorithm is a supervised method, which is based on information theory [37]. It takes both identifying correlated features for classification and eliminating redundant features into account [38]. Based on the symmetrical uncertainty (SU), FCBF ranks features in descending order of correlation and casts off those redundant features that are less correlated [39]. Consequently, the optimal correlated and non-redundant features subset is acquired. It could maximize the diagnostic potential of the extractable information. The relative importance of a metabolic biomarker feature is de-

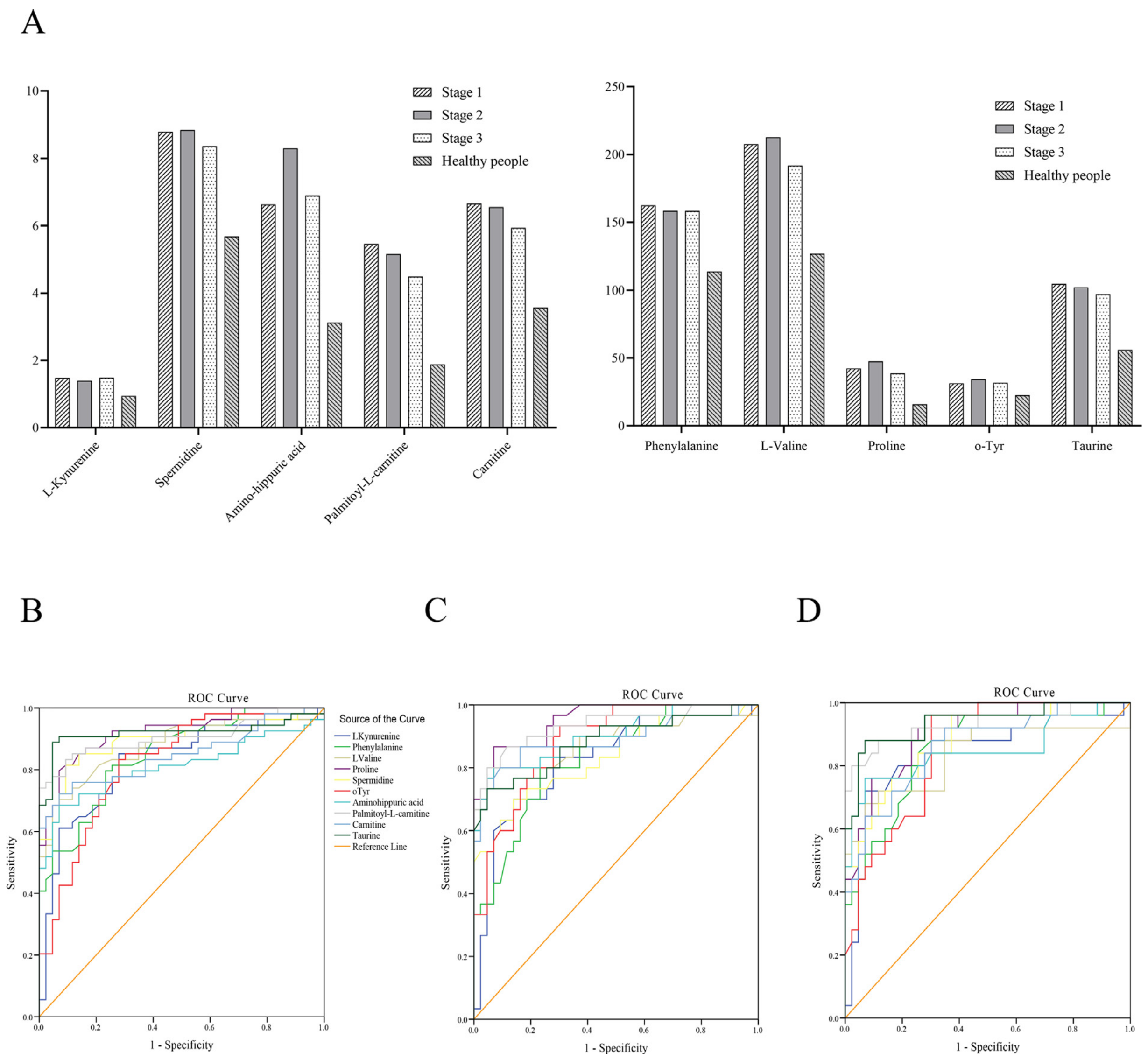

Fig. 3. Metabolomic biomarkers changes with tumor stage progress. **(A)** It showed the levels of 10 metabolites which showed significant difference in stage I ($n = 54$), stage II ($n = 31$), stage III ($n = 25$) lung tumor patients and healthy individuals ($n = 43$). **(B)** ROC curve of metabolomic biomarkers of stage I ($n = 54$) lung tumor patients. **(C)** ROC curve of metabolomic biomarkers of stage II ($n = 31$) lung tumor patients. **(D)** ROC curve of metabolomic biomarkers of stage III ($n = 25$) lung tumor patients. ROC, receiver operating characteristic.

veloped using the Fast Correlation-Based Filter (FCBF) algorithm. All metabolites features were ranked and scored according to their ability to discern the classification label of an object (Table S7). According to the ranking, top 8 metabolic biomarkers were used to develop the machine learning models, respectively. Table S8 and Fig. S4 showed that AUC values changed with the number of variates. When the top 5 variates including taurine, Palmitoyl-L-carnitine, proline, 2-DG, and PE (36:4) were used as metabolites features, prediction model showed the excellent performance which was similar with previous models. Therefore, these 5 metabolic biomarkers could be potential candidates for preclinical screening of lung cancer. Then based on the results of logistic regression analysis, the combination of top three variates including taurine, Palmitoyl-L-carnitine, and proline, showed the AUC value of 0.968 (95% CI: 0.931–1.000), with a sensitivity of 94.4% and specificity of 97.7% in classical analysis (Table 1).

To validate diagnosis performance of machine learning models and demonstrate the specificity of metabolic biomarker features of early lung cancer patients found in our study, we performed a control experiment [40]. We created a scrambled training set that correctly labeled training set was replaced with a training set where the labels were randomly assigned. As expected, the accuracy of our machine learning models markedly dropped, which is equivalent to randomly choosing, showing no predictive value and validating the specific predictive signature of metabolic biomarker features (Fig. 4B).

Discussion

Lung cancer is the worldwide leading cause of cancer-related mortality, which early diagnosis could improve survival rate. However, high-risk people are generally recommended annual radiologic screening by low-dose computed tomography (LDCT) [41], which is also the one and only way of the clinically lung cancer detection at present. Due to the significant cost and high false-discovery rate [42], fulfillment of CT screening is unsatisfactory. Therefore, the availability of blood-based screening could increase lung cancer patient uptake, including plasma metabolic biomarkers detection. Our present study attempted to discover Chinese patients' plasma metabolites as predictive biomarkers for lung cancer diagnosis. In this work, we use a pioneering interdisciplinary

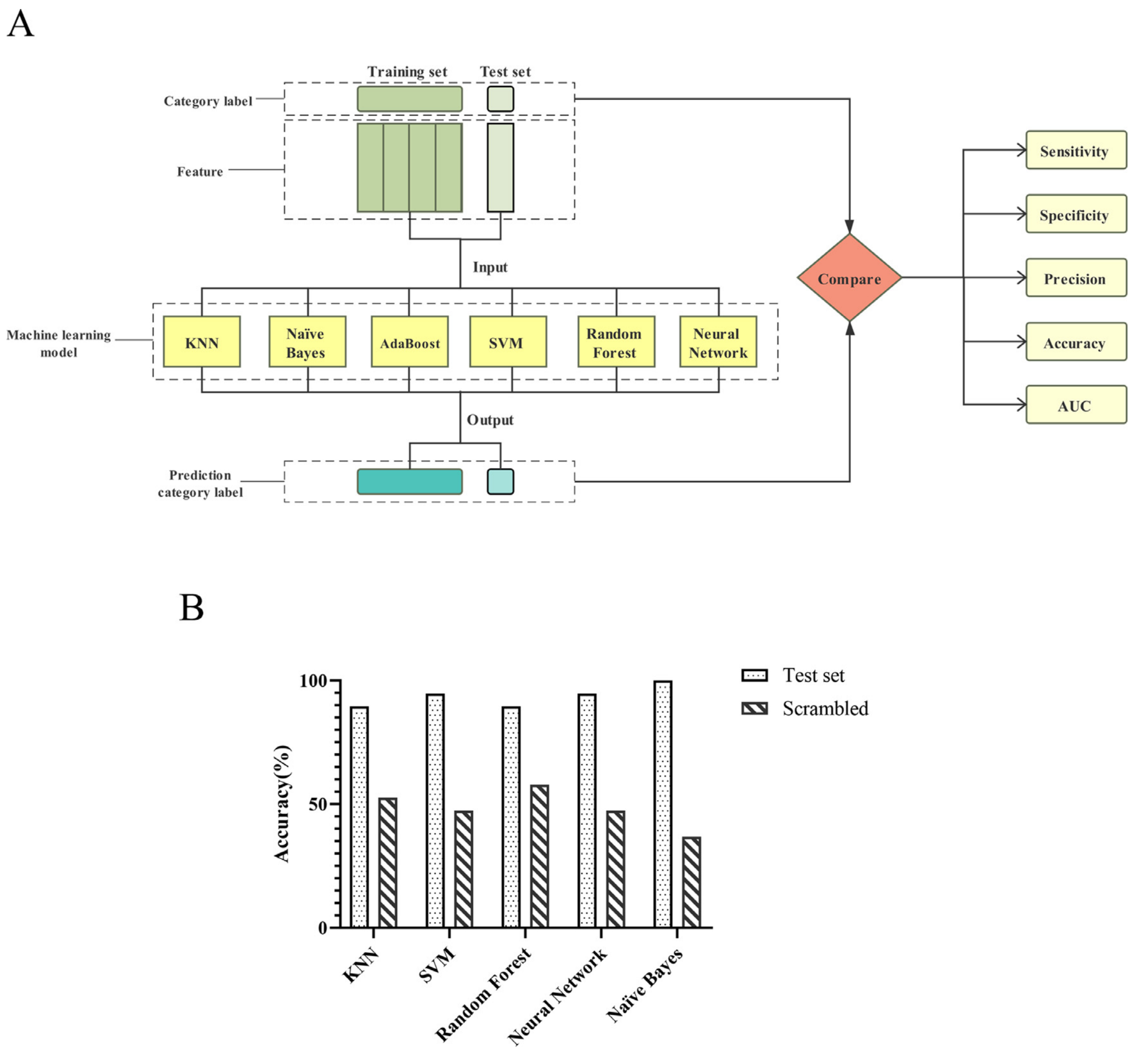

Fig. 4. Machine learning was applied to develop the diagnostic model for early stages of lung cancer. **(A)** Machine learning applications build early lung tumor prediction models. **(B)** To validate diagnosis performance of machine learning models and demonstrate the specificity of metabolic biomarker features of early lung cancer patients found in our study, we created a scrambled set that showed no predictive value. AUC, area under the curve. AdaBoost, Adaptive Boosting. SVM, support vector machines. KNN, k-nearest neighbor.

mechanism, which is firstly applied to lung cancer, to detect early lung cancer diagnostic biomarkers by combining metabolomics and machine learning methods. The highly sensitive and accurate metabolomics technology and machine learning methods bring novelty to our work compared with previous study that used miRNAs as biomarkers for early lung cancer diagnostic.

Main previous studies of breast cancer, prostate cancer and other cancers have screened miRNAs as biomarkers for distinguishing cancer patients and different tumor subtypes [43]. Nevertheless, miRNA screening has its limitation, such as high cost and technical monopolization [44]. Then, clinical usual tests of blood antigen CEA, CA125, SCC, etc. still meets the problem of low sensitivity and low accuracy [45]. At the same time, small biopsy specimen means trauma and false negative as the limited section site [46]. In our present study, we focused on seeking out metabolic biomarkers as early stage lung cancer diagnostic biomarkers. Herein, we identified a new range of metabolites by determining their plasma profiles in patients prior to lung tumor clinical diagnosis. From 61 kinds of metabolites, we found 10 metabolic biomarkers could act as the promising biomarker for early detection of lung tumor. Particularly, the combination of six variates remarkably enhanced the AUC to 0.989 with sensitivity of 98.1% and specificity of 100.0%, which has not been reported so far. According to these results, we recommended the specific combination of these six metabolites as screening biomarker for early detection of lung tumor. We believe that this finding could allow us to develop a specific, sensitive, and minimally invasive implement for early lung tumor prevention and prediction.

To date, machine learning approaches, which provide a promising alternative to classical data analysis methods, have been utilized in varied biomedical applications, including drug discovery [47], and biomarker development [48]. Rather than conventional types of data analysis requiring prior awareness of biological dependencies, machine learning could improve diagnostic capability through abundant and high-quality data. One team collected 23 items of demographic data and tumor-related parameters of 102 cervical cancer patients who had undergone radical hysterectomy for treatment, and investigated diverse machine learning models to predict 5-year survival rate in patients [49]. Another group made use of 2267 women colposcopy findings and human papillomavirus (HPV) biomarkers to develop a clinical decision support scoring system using artificial neural networks for cervical intraepithelial neoplasia patients, in which showing the ANN predicted with higher accuracy compared to cytology with or without HPV test [50]. Moreover, it has been proved the possibility to excavate serum microRNA panel as a potential biomarker for the detection of gastric cancer by machine learning [51]. Six types of machine learning techniques were used to select three biomarkers (miR-21-5p, miR-29c-3p, and miR-22-3p) from the published miRNA profiling study (GSE23739). Herein, we firstly used metabolic biomarkers as machine-learning features for lung cancer diagnosis, and the obtained results were analyzed and discussed. The top 5 relative importance metabolic biomarkers (taurine, Palmitoyl-L-carnitine, proline, PE (36:4) and 2-DG), which were developed by FCBF algorithm, could be potential candidates for pre-clinical screening of lung cancer. In this study, several machine learning models were applied and compared, which were evaluated by test set and control experiment (Fig. 4B and Table 2). It leads to greatest assessment values on Naïve Bayes, but it also leads to decent assessment values on Neural Network, and SVM. As standalone screening models, high sensitivity would be desirable to minimize false positives. As shown in Table 2, Naïve Bayes, Random Forest, Neural Network, and SVM models with high sensitivity have dependable and stable potential for early lung tumor prediction. Furthermore, specificity thresholds set in the training set performed similarly when applied to the test set, indicating that models are well calibrated. The specificity threshold of Naïve Bayes, Neural Network, and SVM models defined in the training set achieved a similar specificity in the test set. Given the above of our study, Naïve Bayes, Neural Network, and SVM, based on the metabolic biomarker features, may be conducive for the diagnosis of early lung tumor. These results provided strong support for the feasibility of blood-based screening basing on metabolomics technology and machine learning for early lung cancer diagnostic.

Cancer progression is strongly related to cellular metabolism, and the dysregulated metabolism is conducive to tumor progression and initiation [52]. Cancer cells change their metabolic pathways, which request specific enzymes to catalyze biochemical reactions, to meet the growing need of the rapid cell reproduction and division. Metabolites, including metabolic biomarkers for lung cancer diagnosis found in our study, have functions related to tumorigenesis and tumor progression. Proline dehydrogenase (PRODH), which catalyzes the first step of proline degradation, is activated by lymphoid-specific helicase (LSH) to decrease proline levels [53]. Proline catabolism relating PRODH has been shown could either promote tumor survival through ROS-induced autophagy or ATP production, or as tumor suppressor to initiate ROS-mediated apoptosis depending on the tumor microenvironment [54,55]. One team recent study found that PRODH promotes lung cancer tumorigenesis by eliciting the expression of IKKα-dependent inflammatory genes and epithelial to mesenchymal transition (EMT) [56]. High levels of L- kynurenine could provide a microenvironment for lung tumor growing through initiating T-cell apoptosis, inhibiting T-cell proliferation and leading to immune tolerance [57,58]. Spermine N1-acetyltransferase (SSAT) is the pivotal protein involved in the homeostasis and synthesis of the spermidine [59]. Spermidine/SSAT is the rate-limiting step in the catabolism of polyamines, which play particular role in maintaining the membrane potential and regulating cell volume [60, 61]. Recent studies have reported that SSAT is upregulated in lung cancer [62]. 2-Deoxy-D-glucose (2DG), a glucose analogue, is converted to 2-DG-P by hexokinase. 2-DG-P cannot be metabolized but it could allosterically suppress hexokinase, which is the rate-limiting enzyme of glycolysis [63]. On account of blocking glycolysis, 2-DG influences in various biological processes. It could inhibit N-linked glycosylation, increases oxidative stress, and efficiently suppresses cell growth and invasion [64].The amino acid 2-aminoethanesulfonic acid, commonly known as taurine, has widespread physiological effects and was confirmed as the endogenous anti-injury material [65]. It could up-regulate the expression of N-acetyl galactosaminyl transferase 2, down-regulate the expression of matrix metalloproteinase-2, and inhibit the potential invasion and metastasis [66]. Previous studies have proposed that changes of taurine levels could be used to predict the malignant transformation and formation of breast, bladder and colorectal tumors [67–69].

Recently, on March 2020, Chabon et al. used integrating genomic features for non-invasive early lung cancer detection [70], which initially demonstrated machine learning method could be used for lung cancer detection. Based on cell-free DNA (cfDNA) features, researchers developed and prospectively validated a machine-learning method termed 'lung cancer likelihood in plasma' (Lung-CLiP), which could discriminate early lung cancer patients from controls. During screening test, they observed sensitivities of 63% stage I lung cancer patients with 80% specificity. Compared with Naïve Bayes, Neural Network, and SVM models basing on the metabolic biomarkers as features developed in our study, the sensitivity and specificity are > 85%. Although our machine learning models is less accurate than LDCT, this strategy could potentially increase the total number of patients screened. As this strategy progresses into clinical trial, abundant sample data will allow for improvement of performance by using more progressive machine-learning algorithms.

On the other hand, our current study still has several limitations. First, our analysis was built on data from single healthcare institution of finite geographic region. And non-small cell lung cancer excepted, the number of patients with other types of lung cancer was inadequate. Therefore, further confirmatory studies at other institutions are necessary prior to implementation. Second, our data only included the metabolites level. More information of lung tumor patients and healthy individuals, such as age, history of smoking, the concurrent tumor diagnosis, and past medical history, would be helpful for further study. We propose that integration of plasma metabolic biomarkers with CT screening or other lung tumor features could further improve performance.

In any case, we need to figure out a proper method to apply this strategy in clinical practice, such as combined with electronic chips system in order to make the plasma tests and model application in an assembly line. One potential application of our study could be served as a premier screening for some of the lung cancer patients. In despite of being candidates for LDCT as high-risk people, these patients are not being screened due to the concerns with false positives, limited access and other limited reasons. Then patients who show positive tests would then be referred to LDCT screening. Additionally, by modifying the machine learning methods and incorporating features appropriate for other cancer types, we expect that it could be feasible to develop strategy combining metabolomics and machine learning for a diverse range of malignancies diagnosis.

Conclusions

A pioneering interdisciplinary method was proposed in this study to detect early lung cancer diagnostic biomarkers by combining metabolomics and machine learning methods. Metabolic biomarkers demonstrate significant diagnostic strength for early detection of lung tumor.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ying Xie: Methodology, Investigation, Writing - review & editing. **Wei-Yu Meng:** Writing - original draft, Methodology, Formal analysis. **Run-Ze Li:** Writing - original draft, Conceptualization, Investigation. **Yu-Wei Wang:** Software, Data curation. **Xin Qian:** Investigation, Resources. **Chang Chan:** Investigation, Resources. **Zhi-Fang Yu:** Investigation, Resources. **Xing-Xing Fan:** Resources. **Hu-Dan Pan:** Resources. **Chun Xie:** Resources, Project administration. **Qi-Biao Wu:** Resources. **Pei-Yu Yan:** Resources. **Liang Liu:** Resources, Supervision. **Yi-Jun Tang:** Resources, Supervision. **Xiao-Jun Yao:** Conceptualization, Writing - review & editing, Funding acquisition. **Mei-Fang Wang:** Conceptualization, Resources, Writing - review & editing. **Elaine Lai-Han Leung:** Conceptualization, Writing - review & editing, Funding acquisition.

Funding

This work was funded by The Science and Technology Development Fund, Macau SAR (File no: 0096/2018/A3, 0003/2019/AKP, 0055/2020/A) and the NSFC overseas and Hong Kong and Macao Scholars Cooperative Research Fund Project (Project no: 81828013).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tranon.2020.100907.

References

[1] D.S. Ettinger, D.E. Wood, D.L. Aisner, W. Akerley, J. Bauman, L.R. Chirieac, T.A. D'Amico, M.M. DeCamp, T.J. Dilling, M. Dobelbower, R.C. Doebele, R. Govindan, M.A. Gubens, M. Hennon, L. Horn, R. Komaki, R.P. Lackner, M. Lanuti, T.A. Leal, L.J. Leisch, R. Lilenbaum, J. Lin, B.W. Loo Jr., R. Martins, G.A. Otterson, K. Reckamp, G.J. Riely, S.E. Schild, T.A. Shapiro, J. Stevenson, S.J. Swanson, K. Tauer, S.C. Yang, K. Gregory, M. Hughes, Non-small cell lung cancer, version 5.2017, NCCN clinical practice guidelines in oncology, J. Natl. Compr. Cancer Netw. 15 (2017) 504–535.

[2] Z. Chen, C.M. Fillmore, P.S. Hammerman, C.F. Kim, K.K. Wong, Non-small-cell lung cancers: a heterogeneous set of diseases, Nat. Rev. Cancer 14 (2014) 535–546.

[3] X. Jin, S.J. Yun, P. Jeong, I.Y. Kim, W.J. Kim, S. Park, Diagnosis of bladder cancer and prediction of survival by urinary metabolomics, Oncotarget 5 (2014) 1635–1645.

[4] H. Wang, J. Chen, Y. Feng, W. Zhou, J. Zhang, Y.U. Yu, X. Wang, P. Zhang, (1)H nuclear magnetic resonance-based extracellular metabolomic analysis of multidrug resistant Tca8113 oral squamous carcinoma cells, Oncol. Lett. 9 (2015) 2551–2559.

[5] C.W. Lam, C.Y. Law, Untargeted mass spectrometry-based metabolomic profiling of pleural effusions: fatty acids as novel cancer biomarkers for malignant pleural effusions, J. Proteome Res. 13 (2014) 4040–4046.

[6] S. Bamji-Stocke, V. van Berkel, D.M. Miller, H.B. Frieboes, A review of metabolism-associated biomarkers in lung cancer diagnosis and treatment, Metabolomics 14 (2018) 81.

[7] J. Zhu, D. Djukovic, L. Deng, H. Gu, F. Himmati, E.G. Chiorean, D. Raftery, Colorectal cancer detection using targeted serum metabolic profiling, J. Proteome Res. 13 (2014) 4120–4130.

[8] W. Guan, M. Zhou, C.Y. Hampton, B.B. Benigno, L.D. Walker, A. Gray, J.F. McDonald, F.M. Fernández, Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines, BMC Bioinform. 10 (2009) 259.

[9] K. Kim, P. Aronov, S.O. Zakharkin, D. Anderson, B. Perroud, I.M. Thompson, R.H. Weiss, Urine metabolomics analysis for kidney cancer detection and biomarker discovery, Mol. Cell Proteom. 8 (2009) 558–570.

[10] S. Tiziani, V. Lopes, U.L. Günther, Early stage diagnosis of oral cancer using 1H NMR-based metabolomics, Neoplasia 11 (2009) 269–276 264p following 269.

[11] S. Urayama, W. Zou, K. Brooks, V. Tolstikov, Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer, Rapid Commun. Mass Spectrom. 24 (2010) 613–620.

[12] R.D. Nindrea, T. Aryandono, L. Lazuardi, I. Dwiprahasto, Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta–analysis, Asian Pac. J. Cancer Prev. 19 (2018) 1747–1752.

[13] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (2019) 463–477.

[14] V. Dalal, J. Carmicheal, A. Dhaliwal, M. Jain, S. Kaur, S.K. Batra, Radiomics in stratification of pancreatic cystic lesions: machine learning in action, Cancer Lett. 469 (2020) 228–237.

[15] B. Zhang, X. He, F. Ouyang, D. Gu, Y. Dong, L. Zhang, X. Mo, W. Huang, J. Tian, S. Zhang, Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma, Cancer Lett. 403 (2017) 21–27.

[16] E.J. Mucaki, J.Z.L. Zhao, D.J. Lizotte, P.K. Rogan, Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning, Signal Transduct. Target Ther. 4 (2019) 1.

[17] W. Xu, M. Xu, L. Wang, W. Zhou, R. Xiang, Y. Shi, Y. Zhang, Y. Piao, Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers, Signal Transduct. Target. Ther. 4 (2019) 55.

[18] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges, Cancer Lett. 471 (2020) 61–71.

[19] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inform. 108 (2017) 1–8.

[20] S. Luo, J. Xu, Z. Jiang, L. Liu, Q. Wu, E.L. Leung, A.P. Leung, Artificial intelligence-based collaborative filtering method with ensemble learning for personalized lung cancer medicine without genetic sequencing, Pharmacol. Res. 160 (2020) 105037.

[21] E.I. Emin, E. Emin, A. Papalois, F. Willmott, S. Clarke, M. Sideris, Artificial intelligence in obstetrics and gynaecology: is this the way forward? In Vivo 33 (2019) 1547–1551.

[22] W.D. Travis, E. Brambilla, A.P. Burke, A. Marx, A.G. Nicholson, Introduction to the 2015 World Health Organization classification of tumors of the lung, pleura, thymus, and heart, J. Thorac. Oncol. 10 (2015) 1240–1242.

[23] R. Suguro, X.C. Pang, Z.W. Yuan, S.Y. Chen, Y.Z. Zhu, Y. Xie, Combinational applicaton of silybin and tangeretin attenuates the progression of non-alcoholic steatohepatitis (NASH) in mice via modulating lipid metabolism, Pharmacol. Res. 151 (2020) 104519.

[24] H. Pan, Y. Zheng, Z. Liu, Z. Yuan, R. Ren, H. Zhou, Y. Xie, L. Liu, Deciphering the pharmacological mechanism of Guan-Jie-Kang in treating rat adjuvant-induced arthritis using omics analysis, Front. Med. 13 (2019) 564–574.

[25] C. Chen, R. Xia, H. Chen, Y. He, TBtools: A Toolkit for Biologists Integrating Various HTS-data Handling Tools with a User-Friendly Interface, bioRxiv, (2018) 289660.

[26] J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, et al., Orange: data mining toolbox in Python, J. Mach. Learn. Res. 14 (2013) 2349–2353.

[27] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, Cancer Genom. Proteom. 15 (2018) 41–51.

[28] P. Shi, S. Ray, Q. Zhu, M.A. Kon, Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction, BMC Bioinform. 12 (2011) 375.

[29] B. Ambale-Venkatesh, X. Yang, C.O. Wu, K. Liu, W.G. Hundley, R. McClelland, A.S. Gomes, A.R. Folsom, S. Shea, E. Guallar, D.A. Bluemke, J.A.C. Lima, Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis, Circ. Res. 121 (2017) 1092–1101.

[30] N.A. Zaidi, J. Cerquides, M.J. Carman, G.I. Webb, Alleviating Naive Bayes attribute independence assumption by attribute weighting, J. Mach. Learn. Res. 14 (2013) 1947–1988.

[31] A.B. Banu, P. Thirumalaikolundusubramanian, Comparison of Bayes classifiers for breast cancer classification, Asian Pac. J. Cancer Prev. 19 (2018) 2917–2920.

[32] H.H. Rashidi, N.K. Tran, E.V. Betts, L.P. Howell, R. Green, Artificial intelligence and machine learning in pathology: the present landscape of supervised methods, Acad. Pathol. 6 (2019) 2374289519873088.

[33] W.A. Cooper, S. O'Toole, M. Boyer, L. Horvath, A. Mahar, What's new in non-small cell lung cancer for pathologists: the importance of accurate subtyping, EGFR mutations and ALK rearrangements, Pathology 43 (2011) 103–115.

[34] B. Liu, S.J. Kim, K.J. Cho, S. Oh, Development of machine learning models for diagnosis of glaucoma, PLoS One 12 (2017) e0177726.

[35] M.S. Ahmed, M. Shahjaman, M.M. Rana, M.N.H. Mollah, Robustification of Naïve Bayes classifier and its application for microarray gene expression data analysis, Biomed. Res. Int. 2017 (2017) 3020627.

[36] H. Zhang, J.X. Ren, J.X. Ma, L. Ding, Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier, Mol. Divers 23 (2019) 381–392.

[37] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, M. Clamp, The Ensembl genome database project, Nucleic Acids Res. 30 (2002) 38–41.

[38] F. Li, M. Yang, Y. Li, M. Zhang, W. Wang, D. Yuan, D. Tang, An improved clear cell renal cell carcinoma stage prediction model based on gene sets, BMC Bioinform. 21 (2020) 232.

[39] V. Barroso-García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, D. Álvarez, F. Vaquerizo-Villar, P. Núñez, F. Del Campo, D. Gozal, R. Hornero, Usefulness of recurrence plots from airflow recordings to aid in paediatric sleep apnoea diagnosis, Comput. Methods Programs Biomed. 183 (2020) 105083.

[40] J. Ko, N. Bhagwat, T. Black, S.S. Yee, Y.J. Na, S. Fisher, J. Kim, E.L. Carpenter, B.Z. Stanger, D. Issadore, miRNA profiling of magnetic nanopore-isolated extracellular vesicles for the diagnosis of pancreatic cancer, Cancer Res. 78 (2018) 3688–3697.

[41] W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, B. Ni, P. Gao, P. Wang, Y. Hua, M. Li, 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas, Cancer Res. 78 (2018) 6881–6889.

[42] P.F. Pinsky, D.S. Gierada, W. Black, R. Munden, H. Nath, D. Aberle, E. Kazerooni, Performance of lung-RADS in the National Lung Screening Trial: a retrospective assessment, Ann. Intern. Med. 162 (2015) 485–491.

[43] L. Li, J. Chen, X. Chen, J. Tang, H. Guo, X. Wang, J. Qian, G. Luo, F. He, X. Lu, Y. Ding, Y. Yang, W. Huang, G. Hou, X. Lin, Q. Ouyang, H. Li, R. Wang, F. Jiang, R. Pu, J. Lu, M. Jin, Y. Tan, F.J. Gonzalez, G. Cao, M. Wu, H. Wen, T. Wu, L. Jin, L. Chen, H. Wang, Serum miRNAs as predictive and preventive biomarker for pre-clinical hepatocellular carcinoma, Cancer Lett. 373 (2016) 234–240.

[44] J.B. Poell, R.J. van Haastert, F. Cerisoli, A.S. Bolijn, L.M. Timmer, B. Diosdado–Calvo, G.A. Meijer, A.A. van Puijenbroek, E. Berezikov, R.Q. Schaapveld, E. Cuppen, Functional microRNA screening using a comprehensive lentiviral human microRNA expression library, BMC Genom. 12 (2011) 546.

[45] Q. Yang, P. Zhang, R. Wu, K. Lu, H. Zhou, Identifying the best marker combination in CEA, CA125, CY211, NSE, and SCC for lung cancer screening by combining ROC curve and logistic regression analyses: is it feasible? Dis. Markers 2018 (2018) 1–12.

[46] S. Kitazono, Y. Fujiwara, K. Tsuta, H. Utsumi, S. Kanda, H. Horinouchi, H. Nokihara, N. Yamamoto, S. Sasada, S. Watanabe, H. Asamura, T. Tamura, Y. Ohe, Reliability of small biopsy samples compared with resected specimens for the determination of programmed death-ligand 1 expression in non–small-cell lung cancer, Clin. Lung Cancer 16 (2015) 385–390.

[47] S. Riniker, Y. Wang, J.L. Jenkins, G.A. Landrum, Using information from historical high-throughput screens to predict active compounds, J. Chem. Inf. Model. 54 (2014) 1880–1891.

[48] J. Jeon, S. Nim, J. Teyra, A. Datti, J.L. Wrana, S.S. Sidhu, J. Moffat, P.M. Kim, A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening, Genome Med. 6 (2014) 57.

[49] B. Obrzut, M. Kusy, A. Semczuk, M. Obrzut, J. Kluska, Prediction of 5-year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods, BMC Cancer 17 (2017) 840.

[50] M. Kyrgiou, A. Pouliakis, J.G. Panayiotides, N. Margari, P. Bountris, G. Valasoulis, M. Paraskevaidi, E. Bilirakis, M. Nasioutziki, A. Loufopoulos, M. Haritou, D.D. Koutsouris, P. Karakitsos, E. Paraskevaidis, Personalised management of women with cervical abnormalities using a clinical decision support scoring system, Gynecol. Oncol. 141 (2016) 29–35.

[51] Y. Huang, J. Zhu, W. Li, Z. Zhang, P. Xiong, H. Wang, J. Zhang, Serum microRNA panel excavated by machine learning as a potential biomarker for the detection of gastric cancer, Oncol. Rep. 39 (2018) 1338–1346.

[52] Y. He, M. Gao, H. Tang, Y. Cao, S. Liu, Y. Tao, Metabolic intermediates in tumorigenesis and progression, Int. J. Biol. Sci. 15 (2019) 1187–1199.

[53] J.M. Phang, Proline metabolism in cell regulation and cancer biology: recent advances and hypotheses, Antioxid. Redox Signal. 30 (2019) 635–649.

[54] W. Liu, K. Glunde, Z.M. Bhujwalla, V. Raman, A. Sharma, J.M. Phang, Proline oxidase promotes tumor cell survival in hypoxic tumor microenvironments, Cancer Res. 72 (2012) 3677–3686.

[55] W. Liu, C.N. Hancock, J.W. Fischer, M. Harman, J.M. Phang, Proline biosynthesis augments tumor cell growth and aerobic glycolysis: involvement of pyridine nucleotides, Sci. Rep. 5 (2015) 17206.

[56] Y. Liu, C. Mao, M. Wang, N. Liu, L. Ouyang, S. Liu, H. Tang, Y. Cao, S. Liu, X. Wang, D. Xiao, C. Chen, Y. Shi, Q. Yan, Y. Tao, Cancer progression is mediated by proline catabolism in non-small cell lung cancer, Oncogene 39 (2020) 2358–2376.

[57] L.R. Kolodziej, E.M. Paleolog, R.O. Williams, Kynurenine metabolism in health and disease, Amino Acids 41 (2011) 1173–1183.

[58] S.C. Chuang, A. Fanidi, P.M. Ueland, C. Relton, O. Midttun, S.E. Vollset, M.J. Gunter, M.J. Seckl, R.C. Travis, N. Wareham, A. Trichopoulou, P. Lagiou, D. Trichopoulos, P.H. Peeters, H.B. Bueno-de-Mesquita, H. Boeing, A. Wientzek, T. Kuehn, R. Kaaks, R. Tumino, C. Agnoli, D. Palli, A. Naccarati, E.A. Aicua, M.J. Sánchez, J.R. Quirós, M.D. Chirlaque, A. Agudo, M. Johansson, K. Grankvist, M.C. Boutron-Ruault, F. Clavel-Chapelon, G. Fagherazzi, E. Weiderpass, E. Riboli, P.J. Brennan, P. Vineis, M. Johansson, Circulating biomarkers of tryptophan and the kynurenine pathway and lung cancer risk, Cancer Epidemiol. Biomark. Prev. 23 (2014) 461–468.

[59] A.E. Pegg, Spermidine/spermine-N(1)-acetyltransferase: a key metabolic regulator, Am. J. Physiol. Endocrinol. Metab. 294 (2008) E995–1010.

[60] N. Babbar, A. Hacker, Y. Huang, R.A. Casero Jr., Tumor necrosis factor alpha induces spermidine/spermine N1-acetyltransferase through nuclear factor kappaB in non-small cell lung cancer cells, J. Biol. Chem. 281 (2006) 24182–24192.

[61] A.N. Kingsnorth, H.M. Wallace, Elevation of monoacetylated polyamines in human breast cancers, Eur. J. Cancer Clin. Oncol. 21 (1985) 1057–1062.

[62] S. Singhal, C. Rolfo, A.W. Maksymiuk, P.S. Tappia, D.S. Sitar, A. Russo, P.S. Akhtar, N. Khatun, P. Rahnuma, A. Rashiduzzaman, R. Ahmed Bux, G. Huang, B. Ramjiawan, Liquid biopsy in lung cancer screening: the contribution of metabolomics. Results of a pilot study, Cancers 11 (2019) 1069.

[63] M. Parniak, N. Kalant, Incorporation of glucose into glycogen in primary cultures of rat hepatocytes, Can. J. Biochem. Cell Biol. 63 (1985) 333–340.

[64] D. Zhang, J. Li, F. Wang, J. Hu, S. Wang, Y. Sun, 2-Deoxy-D-glucose targeting of glucose metabolism in cancer cells as a potential therapy, Cancer Lett. 355 (2014) 176–183.

[65] X. Zhang, S. Tu, Y. Wang, B. Xu, F. Wan, Mechanism of taurine-induced apoptosis in human colon cancer cells, Acta Biochim. Biophys. Sin. 46 (2014) 261–272.

[66] P.M. Neary, P. Hallihan, J.H. Wang, R.W. Pfirrmann, D.J. Bouchier-Hayes, H.P. Redmond, The evolving role of taurolidine in cancer therapy, Ann. Surg. Oncol. 17 (2010) 1135–1143.

[67] I.M. El Agouza, S.S. Eissa, M.M. El Houseini, D.E. El-Nashar, O.M. Abd El Hameed, Taurine: a novel tumor marker for enhanced detection of breast cancer among female patients, Angiogenesis 14 (2011) 321–330.

[68] S. Srivastava, R. Roy, S. Singh, P. Kumar, D. Dalela, S.N. Sankhwar, A. Goel, A.A. Sonkar, Taurine – a possible fingerprint biomarker in non-muscle invasive bladder cancer: a pilot study by 1H NMR spectroscopy, Cancer Biomark. 6 (2010) 11–20.

[69] S. Tu, X.L. Zhang, H.F. Wan, Y.Q. Xia, Z.Q. Liu, X.H. Yang, F.S. Wan, Effect of taurine on cell proliferation and apoptosis human lung cancer A549 cells, Oncol. Lett. 15 (2018) 5473–5480.

[70] J.J. Chabon, E.G. Hamilton, D.M. Kurtz, M.S. Esfahani, M. Diehn, Integrating genomic features for non-invasive early lung cancer detection, Nature 580 (2020) 245–251.