

Original Paper

A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study

Jiangpeng Wu^{1,2*}, MD; Xiangyi Zan^{3*}, MD; Liping Gao³, MD; Jianhong Zhao⁴, MD; Jing Fan², MD; Hengxue Shi², MD; Yixin Wan³, MD; E Yu⁵, MD; Shuyan Li^{1,2}, PhD; Xiaodong Xie⁶, PhD

¹State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou, China

²College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China

³Department of Pneumology, Lanzhou University Second Hospital, Lanzhou, China

⁴Department of Radiology, Lanzhou University Second Hospital, Lanzhou, China

⁵National Demonstration Centre for Experimental Chemistry Education, Lanzhou University, Lanzhou, China

⁶School of Basic Medical Science, Lanzhou University, Lanzhou, China

*these authors contributed equally

Corresponding Author:

Shuyan Li, PhD

State Key Laboratory of Applied Organic Chemistry

Lanzhou University

222 South Tianshui Road

Lanzhou,

China

Phone: 86 931 8912578

Fax: 86 931 8912583

Email: lishuyan@lzu.edu.cn

Abstract

Background: Liquid biopsies based on blood samples have been widely accepted as a diagnostic and monitoring tool for cancers, but extremely high sensitivity is frequently needed due to the very low levels of the specially selected DNA, RNA, or protein biomarkers that are released into blood. However, routine blood indices tests are frequently ordered by physicians, as they are easy to perform and are cost effective. In addition, machine learning is broadly accepted for its ability to decipher complicated connections between multiple sets of test data and diseases.

Objective: The aim of this study is to discover the potential association between lung cancer and routine blood indices and thereby help clinicians and patients to identify lung cancer based on these routine tests.

Methods: The machine learning method known as Random Forest was adopted to build an identification model between routine blood indices and lung cancer that would determine if they were potentially linked. Ten-fold cross-validation and further tests were utilized to evaluate the reliability of the identification model.

Results: In total, 277 patients with 49 types of routine blood indices were included in this study, including 183 patients with lung cancer and 94 patients without lung cancer. Throughout the course of the study, there was correlation found between the combination of 19 types of routine blood indices and lung cancer. Lung cancer patients could be identified from other patients, especially those with tuberculosis (which usually has similar clinical symptoms to lung cancer), with a sensitivity, specificity and total accuracy of 96.3%, 94.97% and 95.7% for the cross-validation results, respectively. This identification method is called the routine blood indices model for lung cancer, and it promises to be of help as a tool for both clinicians and patients for the identification of lung cancer based on routine blood indices.

Conclusions: Lung cancer can be identified based on the combination of 19 types of routine blood indices, which implies that artificial intelligence can find the connections between a disease and the fundamental indices of blood, which could reduce the necessity of costly, elaborate blood test techniques for this purpose. It may also be possible that the combination of multiple indices obtained from routine blood tests may be connected to other diseases as well.

(*JMIR Med Inform* 2019;7(3):e13476) doi: [10.2196/13476](https://doi.org/10.2196/13476)

KEYWORDS

lung cancer identification; routine blood indices; Random Forest

Introduction

Using liquid biopsies based on blood tests is a promising method to achieve noninvasive diagnosis of cancers, but it is also currently a challenge in oncology [1-3]. The main approach for this technique involves the detection of circulating tumor DNAs (ctDNA) [4-6] or specific protein biomarkers [7,8] in plasma. Other cancer biomarkers, such as metabolites [9,10], autoantibodies [11,12], antigens [13,14], microRNAs [15-17], long noncoding RNAs [18,19], and methylated DNAs [3,20,21] were also used. The advantages of this approach include its convenience, and that it is both noninvasive and effective for helping physicians to decide or adjust the treatment schedule for a patient [5,22]. However, its proper usage is still being debated, in part because of its varied results among different patients but also due to its relatively low sensitivity and specificity [7,17,22,23].

Cancers that can be detected with liquid biopsy methods include breast [10], stomach [24], liver [18], pancreas [19], esophagus [14], prostate [17], colorectum [25], laryngeal [9], ovary [26] and lung [27] cancers. Cohen et al even demonstrated the possibility of identifying eight common cancer types simultaneously using blood biopsy, including lung, ovary, liver, stomach, pancreas, esophagus, colorectum and breast cancer, based on a multi-analyte blood test [1]. Among these cancers, lung cancer has a consistently high morbidity and mortality rate compared to all other types of cancers [28], and it has become the leading cause of cancer death worldwide [29]. Therefore, liquid biopsy studies on lung cancer, especially using multiple biomarkers, have attracted a lot of attention [16]. For instance, Leng et al used the integrity of cell-free DNAs to distinguish lung cancer patients from healthy ones with a sensitivity of 79.2% and a specificity of 67.3% [30]. Li et al used a combination of 13 protein biomarkers as a classifier to distinguish lung cancer and reached a sensitivity of 93% [31]. Chen et al utilized 10 serum microRNAs as biomarkers to identify lung cancer and achieved a sensitivity of 93% as well as a specificity of 90% [32]. These results suggest that a combination of multiple biomarkers performs better than testing for a single marker.

Meanwhile, misdiagnosis of lung cancer and tuberculosis occurs frequently in clinical situations [33] due to some misleading images obtained by computed tomography (CT) scans. This is one of the most common detection approaches for lung cancer in the clinic, along with tissue biopsies, as CT scans can detect a smaller nodule and find hidden areas when detecting lung cancer. However, they aren't specific enough to identify lung cancer from benign nodules and tuberculosis [34]. Therefore, patients who are not immediately found to have lung cancer usually undergo unnecessary tissue biopsies, such as needle biopsy, bronchoscopy, thoracoscopy, mediastinoscopy or thoracotomy [35]. Aiming at this problem, Leng et al tried to

use DNA biomarkers to distinguish lung cancer from tuberculosis and got an 82.9% specificity and a barely satisfactory 55.7% sensitivity [30].

In this work, inspired both by the fact that multi-analyte blood tests can reveal greater correlation between complicated connections, and that comprehensive consideration of multiple factors may also mitigate the effects of variation between individual patients, we tried to find the connection between the results of routine blood examinations and serious diseases. Although none of the blood test data for a single factor was proven to be the sole indicator of lung cancer, it was found that a combination of 19 routine blood biochemical indices were highly related as indicators of lung cancer, based on the Random Forest method [36]. This approach presented a chance to classify lung cancer through the use of a cross-validation set and a test set, with tuberculosis samples included. To the best of our knowledge, this is the first time that a combination of routine blood biochemical indices is presented for its capability to well distinguish lung cancer, especially from tuberculosis.

Methods**Source of Materials**

Data from routine blood tests were collected from the Second Hospital of Lanzhou University. A total of 277 patients with 49 types of routine blood indices were included in this study, including 183 patients whose lung cancer was diagnosed by tissue biopsies as positive samples and another 94 patients, without lung cancer, as negative samples. These patients ranged from 20 to 81 years of age, and general information about their data sets can be accessed in [Table 1](#) (for detailed information about these patients, including sex, age, smoking status, cancer stage and blood indices, see [Multimedia Appendix 1](#)). It should be noted that among the 94 negative patients, 51 with tuberculosis were specifically included since there is a high false positive rate in using CT scans to distinguish lung cancer from tuberculosis. Tuberculosis patients were carefully diagnosed with a combination of CT images and clinical symptoms by an experienced clinician. The other patients in the negative group just went to the hospital for routine physical examinations and were not diagnosed with any lung tissue-related diseases. All of the samples were collected from unrelated patients. The Lanzhou University Ethics Committee granted approval of this study and each participant signed an informed consent form after receiving a verbal explanation of the study.

After collection, the data were randomly split into a training set and a test set with a ratio of about 4 to 1. The training set included 222 patients and was constructed with 149 lung cancer samples, 37 tuberculosis samples, and 36 other samples, and then the remaining 55 samples were assigned to the test set.

Table 1. General demographic information on the test set and the training set (N=277).

Characteristic	Training set			Test set		
	Lung cancer (n=149)	Tuberculosis (n=37)	Other (n=36)	Lung cancer (n=34)	Tuberculosis (n=14)	Other (n=7)
Gender, n						
Male	110	37	12	22	5	5
Female	39	20	24	12	9	2
Median age (range)	60 (27-81)	46 (20-79)	55 (30-78)	58 (38-79)	52 (20-78)	62 (49-68)
Smokers, n	44	2	2	5	0	1

Machine Learning Method

The Random Forest method (RF) [36] was adopted here to build the final classification model. RF is a very powerful and practical classifier that can use multiple trees to train an AI to predict samples, and it has been extensively employed in the fields of chemometrics and bioinformatics [37]. There are two main advantages to the RF method which are that, first, it can use an out of the bag set to monitor errors, strengths, and correlation [38], and second, it can measure variable importance through permutation. The RF method can handle high-dimensional data and approach the best predictor for them by further decreasing the dimensions of feature space and discovering rigorous feature numbers. For this algorithm, the two most important parameters were the tree number (ntree) and the number of randomly selected features to split at each node (mtry), which needed to be adjusted to get the best classification model. In this work, we at first made use of the entire set of indices to establish an RF classification prediction model on the basis of the 10-fold cross-validation. For each index, the importance of its association with the prediction target was demonstrated in this procedure. Then, based on increasing the number of top-ranking indices, the RF model was built with adjusted parameters. The initial value of ntree was 100, which increased by 100 until it reached 1500. The value of mtry was set to 2-10 with a step of 1. Finally, we chose the most suitable model with the fewest number of top-ranking indices but with a similar prediction performance compared to the entire index space. Then, the 19 top-ranking indices with ntree and mtry values of 1300 and 9, respectively, were selected for the final model. This selection process also helped us to locate the key indices for predicting lung cancer. RF was executed by applying the Random Forest package of R.

Validation Method

Both internal cross-validation and further tests were adopted to obtain a reliable classifier for lung cancer. The entire modelling process, including feature ranking, RF parameter adjusting, and final model selection, was performed based only on the training set using 10-fold cross-validation. The presplitting test set for further testing of the built model was not involved in any of these model-building processes, as emphasized by Smialowski et al [39]. Ten-fold cross-validation is employed to randomly divide the training set into 10 nonoverlapping parts, one of which is used as an internal test set while the rest are used as the training set. This process is repeated 10 times so that all samples can be used as an internal test set once. The circular work thus facilitates the potential establishment of a stable

classification model for predicting lung cancer. The average results were obtained after 10 runs of the circular process as the final 10-fold cross-validation result.

Five frequently used indicators were adopted here to evaluate the final performance of the routine blood indices model for lung cancer (RBLC) method, including sensitivity (Sens), specificity (Spec), accuracy (ACC), Matthews correlation coefficient (MCC), and the area under the curve (AUC), where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

$$Sens = \frac{TP}{TP+FN}$$

$$Spec = \frac{TN}{TN+FP}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The receiver operating characteristic (ROC) curve is a composite indicator and a graphical plot for the continuous variables of Sens and Spec, with Sens as the y-axis and 1-Spec as the x-axis. One characteristic of the ROC curve is that it could remain unchanged if the positive and negative samples are out of balance in the test set.

AUC is the area under the ROC curve, and it can range from a value of 0 to a value of 1. The closer the AUC is to 1, the better the prediction performance of lung cancer. It is one of the main evaluation indices for a binary classifier system.

Results

Model Selection

Routine blood tests listed in [Multimedia Appendix 1](#) are easy to perform and low cost, but no direct connection between these routine blood tests and the diagnosis of cancers has been found and used in clinical trials yet. This is one of the most important reasons for a surge in interest in finding new biomarkers for cancers. Recent research has indicated some comprehensive connections between certain symptoms and some disorders, such as Axelsson et al demonstrating the facial cues of sick people [40]. However, these studies left unanswered the question of if it was possible to use machine learning methods to find any connection between cancer and these routine blood indices.

To answer that question in this study, we used routine blood and biochemical test data that can be measured by common chemistry analyzers, with a cost of approximately \$10-20 for each sample, to determine their correlation with lung cancer. Surprisingly, positive correlation was found with a simple Random Forest (RF) test method, with 19 blood indices enough to prove correlation. With the data set we used, an MCC of 91.36%, ACC of 95.7% (Figure 1A) and AUC of 99.01% (Figure 1B) were attained. The detailed information about these 19 indices, such as their typical values, units and biological meanings, can be found in in Table 2. The model that was constructed is referred to as RBLC.

In fact, 19 indices are equivalent to a critical point (Figure 1A). The principle of selecting the number of features was to use the minimum features possible to achieve a comparable prediction performance as the entire feature space. The fewer features that a model consists of, the less probability it gets an overfitting problem. If the number of features was increased from 19 to 38, many features would be unnecessary because its results would be comparable to the previous predictive performance. Therefore, in our opinion it is a better choice that the final model

has only 19 features, to not only establish a simple, efficient and robust classification model, but also to avoid excessive waste of blood test procedures and save diagnosis time.

The detailed forest structure for the RBLC model is illustrated in Figure 2. Each tree in the forest votes for the major classification based on different combinations of blood indices, and the majority of votes results in the final classification of the RBLC model (Figure 2A). In addition, each node in each tree votes for the classification, upon independent decision rule, for each different blood index, and hence deduces a final vote for a single tree (Figure 2B). This model achieved not only a great improvement in sensitivity and specificity but also high precision prediction performance, such that the sensitivity, specificity, and accuracy scores were all greater than 85% in the test set, with values of 85.71%, 90%, 88.24%, respectively. The MCC value and AUC for the test set also got 75.71% and 90.16%, respectively. These results indicate that this RBLC method has the optimum and stable prediction performance needed to distinguish lung cancer from tuberculosis and other samples.

Figure 1. Classification performance of the RBLC model. (A) Cross-validation results of models which were built on top ranking features. (B) ROC curves and the corresponding AUCs for the cross-validation on the training set and for the test set. RBLC: routine blood indices model for lung cancer; ROC: receiver operating characteristic; AUC: area under the curve; ACC: accuracy; MCC: Matthews correlation coefficient.

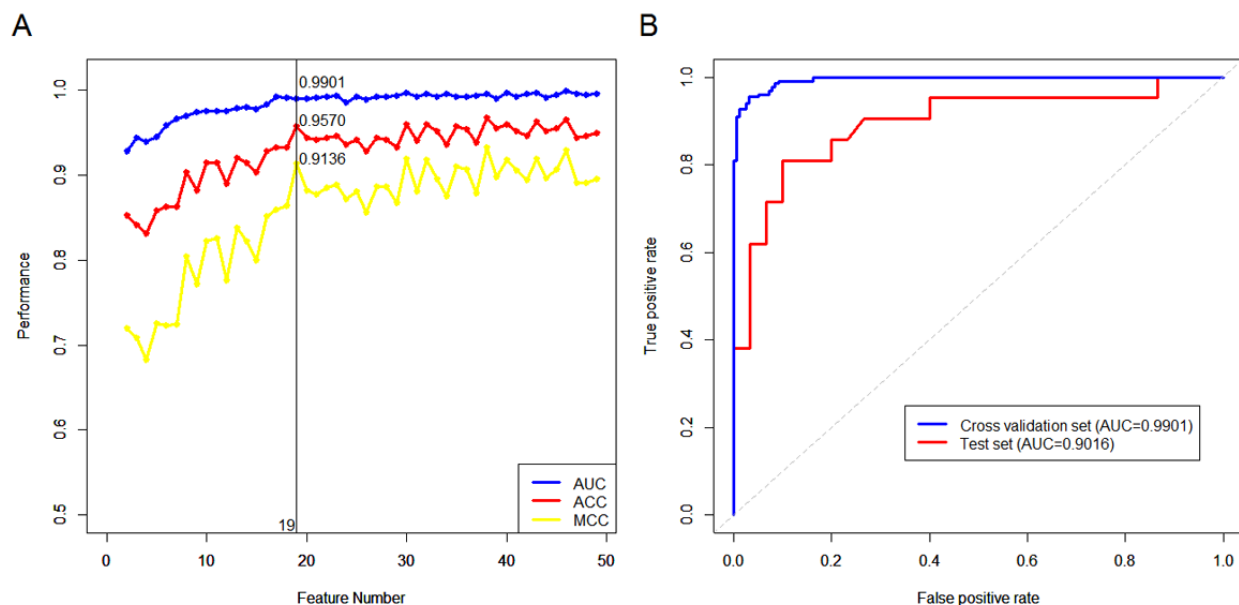
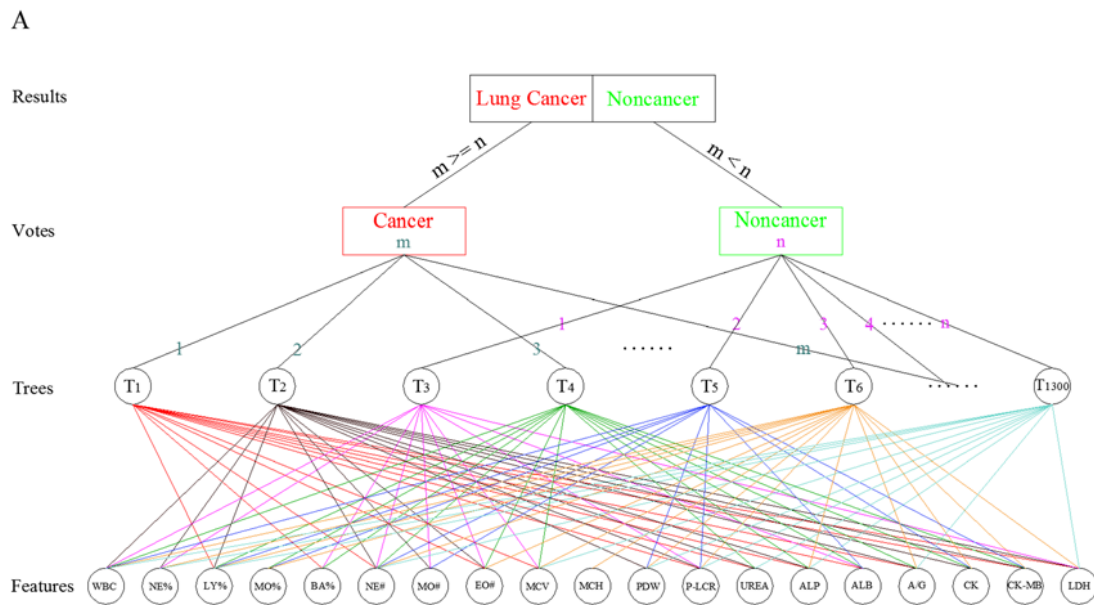


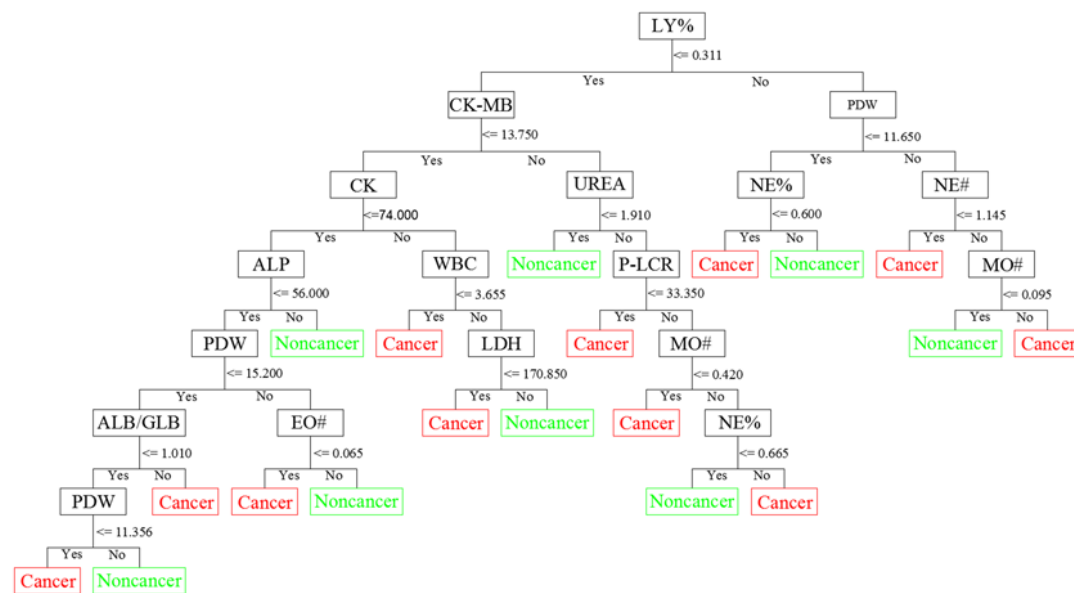
Table 2. Top-ranking blood indices for the identification of lung cancer.

Rank	Index	Reference range
1	Basophil ratio	0.00-0.01
2	Creatine kinase isoenzymes (U/L)	0.0-25.0
3	Platelet large cell ratio (%)	17.0-45.0
4	Albumin (g/L)	30.0-55.0
5	Platelet distribution width (fl)	9.0-17.0
6	Neutrophilic granulocytes ($10^9/L$)	2.00-7.00
7	White blood cell count ($10^9/L$)	4.00-10.00
8	Albumin/Globulin ratio	1.10-2.50
9	Monocytes ($10^9/L$)	0.12-1.20
10	Monocyte ratio	0.03-0.08
11	Lymphocyte ratio	0.20-0.40
12	Neutrophil granulocyte ratio	0.50-0.70
13	Lactate dehydrogenase (U/L)	0.0-240.0
14	Carbamide (mmol/L)	1.80-8.00
15	Eosinophil cells ($10^9/L$)	0.02-0.50
16	Mean corpuscular volume (fl)	80.0-100.0
17	Alkaline phosphatase (U/L)	0.0-120.0
18	Mean corpuscular hemoglobin (pg)	27.0-34.0
19	Creatine kinase (U/L)	0-195

Figure 2. The detailed forest structure for the RBLC model. (A) The general structure of the voting strategy of the RBLC model. (B) The independent decision rulings for different blood indices for the first tree (T1) in (A). T: tree; WBC: white blood cell count; NE%: neutrophil granulocyte ratio; LY%: lymphocyte ratio; MO%: monocyte ratio; BA%: basophil ratio; NE#: neutrophilic granulocytes; MO#: monocytes; EO#: eosinophil cells; MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; PDW: platelet distribution width; P-LCR: platelet large cell ratio; UREA: carbamide; ALP: alkaline phosphatase; ALB: albumin; A/G: albumin/globulin; CK: creatine kinase; CK-MB: creatine kinase isoenzymes; LDH: lactate dehydrogenase.



B



Clinical Relevance

To confirm the efficiency, reliability, and repeatability of the RBLC model, 34 serial blood samples from 15 additional patients were also included in the study (detailed information, including the patients' sex, age, smoking status, cancer stage and blood data, is listed in [Multimedia Appendix 2](#)). Five of these patients were diagnosed with lung cancer by lung tissue biopsy when they got their first blood examination, and then serial blood tests were performed afterward either weekly or

monthly (for 13 samples in all). Of the blood samples collected, 11 were from 5 patients who were diagnosed with tuberculosis (without lung cancer) and 10 were from 5 patients who were diagnosed with neither lung cancer nor tuberculosis. These samples were used as the negative controls. Among these samples, 12/13 with lung cancer, 8/11 with tuberculosis and 9/10 healthy samples were accurately identified. Overall, the sensitivity reached 92.31%, the specificity reached 80.95%, and the total accuracy reached 85.29%. This result for the additional serial data is fairly consistent with the results of the

single-sample test in the test set, which further proves the reliability and stability of the RBLC model. More importantly, it appears to be able to distinguish tuberculosis and lung cancer.

Web Server of Routine Blood Indices Model for Lung Cancer Method

A user-friendly web server is available online to use the RBLC method [41]. Users can input the 19 key features from a routine

blood examination and blood biochemical examination into the corresponding text boxes on the web page (Figure 3) and then press the Submit button. After calculation and analysis of the outputs of the sample, the results page will display whether the input is considered a sample with lung cancer or not.

Figure 3. Web page of the RBLC tool for convenient usage online. RBLC: routine blood indices model for lung cancer; ALB/GLB: albumin/globulin.

Discussion

Overview

The performance of the RBLC method was compared to other commonly used identification methods of lung cancer and ended up showing a favorable result, and then, the association of these selected key routine blood indices with lung cancer was analyzed and further confirmed.

Performance Comparison

With regard to other identification methods, CT scans are a common tool for the detection of lung cancer. For instance, the National Lung Screening Trial (NLST) recommends the use of CT scans to help diagnose patients at high risk for lung cancer. The NLST also demonstrated that mortality could be reduced by 20% using CT screening, with a specificity of 72.6% [42]. However, the low specificity of CT may expose patients to anxiety and unnecessary further examinations.

Table 3. Comparison of the performance of different methods for predicting lung cancer on cross-validation.

Prediction method	Sample size	Sensitivity, %	Specificity, %	Area under the curve
RBLC ^a	226	96.30	94.97	0.99
Protein biomarker [31]	143	93.00	45.00	N/A ^b
RNA biomarker [32]	310	93.00	90.00	0.97
DNA biomarker [30]	318	79.20	67.30	0.75
Computed tomography scans [43]	N/A	94.40	72.60	N/A

^aRBLC: routine blood indices model for lung cancer.

^bN/A: not applicable.

Currently, biomarker analysis is another prevalent technique for detecting lung cancer in high-risk populations. Different lung cancer-related components are ideal biomarkers for the detection of lung cancer. The protein, DNA, and RNA referenced in Table 3 are the latest biomarkers to be developed. Compared to these other methods, the RBLC model demonstrates satisfactory performance in terms of sensitivity, specificity, and AUC, and it is much easier to perform. It is noteworthy that 94.74% of early stage (stage I/II) patients were distinguished by RBLC (see Multimedia Appendix 1), which implies it has further potential for application for identification of early-stage lung cancer.

Key Blood Indices Analysis

Detailed information for the selected key indices for the RBLC model was shown in Table 2, and these indices were listed in

decreasing order of importance. Afterward, all the values of these indices were normalized on a scale going from 0 to 1, and then the average values for both positive and negative samples were shown in Table 4. The *P* values within the table were determined using two-tailed *t* tests.

Among these key indices, the relationship between lactate dehydrogenase (LDH) and lung cancer has been discussed extensively [44]. The expression of LDH not only increases points in glucose metabolism progression, but research has also shown it has a strong association with lung cancer [45]. In this work, the LDH levels of blood samples from lung cancer patients was significantly different from that of negative samples ($P < .001$), which is consistent with previous studies as well.

Table 4. Feature comparison of lung cancer and other samples.

Feature	Negative sample	Positive sample (lung cancer)	<i>P</i> value
White blood cell count	0.1986	0.3088	<.001
Neutrophil-granulocyte ratio	0.4257	0.6502	<.001
Lymphocyte ratio	0.5298	0.3232	<.001
Monocyte ratio	0.4319	0.3970	.20
Basophil ratio	0.2555	0.1242	<.001
Neutrophilic granulocytes	0.1839	0.2808	<.001
Monocytes	0.2795	0.384	<.001
Eosinophil cells	0.3236	0.0833	<.001
Mean corpuscular volume	0.6808	0.5453	<.001
Mean corpuscular hemoglobin	0.6545	0.5983	.008
Platelet distribution width	0.5765	0.6337	.03
Platelet large cell ratio	0.5081	0.4010	<.001
Carbamide	0.4181	0.3197	<.001
Alkaline phosphatase	0.4138	0.1366	<.001
Albumin	0.5757	0.5574	.52
Albumin/globulin	0.3917	0.4155	.46
Creatine kinase	0.1103	0.0867	.19
Creatine kinase Isoenzymes	0.3557	0.2014	<.001
Lactate dehydrogenase	0.5441	0.1462	<.001

In addition, white blood cell count (WBC) is one of the most commonly used, nonspecific markers of inflammation [46]. Chronic bronchitis in a patient would be accompanied by an increase in their WBC, but the association between lung cancer risk and elevated WBC goes beyond preexisting, increased levels [47]. In addition, most tumors are surrounded by inflammatory cells which play an important role in the pathogenesis of cancer by recruiting immune cells that promote survival of the tumor [48]. Our results, like other studies, show a positive association between WBC and lung cancer, in which lung cancer patients have a relatively higher average WBC than negative samples ($P < .001$), although most of the indicators are in the normal clinical range. In previous studies, researchers mainly focused on the value of the neutrophil to lymphocyte ratio as a predictor of lung cancer [49], while neutrophil-granulocyte ratio (NE%) wasn't really considered to be an independent index. The NE% of lung cancer has an obvious difference compared with negative samples ($P < .001$) in our work, which may be of practical importance.

Research on eosinophil cells (EO#) associated with lung cancer is rarely reported. The significant difference in the EO# between lung cancer samples ($P < .001$) and negative samples is indicated in this study as well. There is a common view that paraneoplastic processes and distant metastases (to the bone marrow) will

increase EO# to some extent [50]. Alkaline phosphatase (ALP) is reported to be associated with cancer metastasis in the literature [51], and it was also a critical index for identifying lung cancer and negative samples in our analysis.

Although creatine kinase isoenzymes (CK-MB) have a good specificity for diagnosis of myocardial infarction, related reports have indicated that the presence of malignant tumors can cause a significant distinction in CK-MB levels [52]. Our study also suggested that CK-MB ($P < .001$) has a significantly different average value in lung cancer compared to negative samples.

Conclusion

All of above the results demonstrate that the blood indices we selected were related to lung cancer to some extent, but none of them solely exhibits a clear connection and can be used for diagnostic purposes. With the aid of machine learning, through a combination of multiple test items and connections between the complicated patterns of these blood indices, specific diseases may be distinguished. The identification performance of the RBLC model for lung cancer is rather encouraging, as shown in Table 3. We thus believe that machine learning can reveal the complicated correlation between routine blood test data and other serious diseases, which is currently a case of ongoing research in our group.

Acknowledgments

We thank Professor Qiaosheng Pu for his critical advice on the composition and improvement of this paper. We also thank Professor Jianxi Xiao at Lanzhou University, Professor Chongge You at Lanzhou University Second Hospital, Deputy Chief Examiner Juan Li at Lanzhou University First Hospital, and Deputy Chief Examiner Yonghong Li at Gansu Provincial Hospital for their valuable advice on this work. This work is supported by National Natural Science Foundation of China (#21405068 to SL), the Fundamental Research Funds for the Central Universities of China (#lzujbky-2017-104 to SL).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed information of the samples for RBLC modeling and validation.
[\[XLSX File \(Microsoft Excel File\), 104KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Detailed information of the samples for further clinical relevance evaluation.
[\[XLSX File \(Microsoft Excel File\), 15KB-Multimedia Appendix 2\]](#)

References

1. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018 Feb 23;359(6378):926-930 [FREE Full text] [doi: [10.1126/science.aar3247](https://doi.org/10.1126/science.aar3247)] [Medline: [29348365](https://pubmed.ncbi.nlm.nih.gov/29348365/)]
2. Voora D. A Liquid Solution for Solid Tumors. *Science Translational Medicine* 2013 Apr 10;5(180):180ec62-180ec62. [doi: [10.1126/scitranslmed.3006268](https://doi.org/10.1126/scitranslmed.3006268)]
3. Shen SY, Singhanian R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 2018 Dec;563(7732):579-583. [doi: [10.1038/s41586-018-0703-0](https://doi.org/10.1038/s41586-018-0703-0)] [Medline: [30429608](https://pubmed.ncbi.nlm.nih.gov/30429608/)]
4. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* 2017 Aug 16;9(403). [doi: [10.1126/scitranslmed.aan2415](https://doi.org/10.1126/scitranslmed.aan2415)] [Medline: [28814544](https://pubmed.ncbi.nlm.nih.gov/28814544/)]

5. Almufti R, Wilboux M, Oza A, Henin E, Freyer G, Tod M, et al. A critical review of the analytical approaches for circulating tumor biomarker kinetics during treatment. *Ann Oncol* 2014 Jan;25(1):41-56. [doi: [10.1093/annonc/mdt382](https://doi.org/10.1093/annonc/mdt382)] [Medline: [24356619](https://pubmed.ncbi.nlm.nih.gov/24356619/)]
6. Bettgowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 2014 Feb 19;6(224):224ra24 [FREE Full text] [doi: [10.1126/scitranslmed.3007094](https://doi.org/10.1126/scitranslmed.3007094)] [Medline: [24553385](https://pubmed.ncbi.nlm.nih.gov/24553385/)]
7. Zamay TN, Zamay GS, Kolovskaya OS, Zukov RA, Petrova MM, Gargaun A, et al. Current and Prospective Protein Biomarkers of Lung Cancer. *Cancers (Basel)* 2017 Nov 13;9(11) [FREE Full text] [doi: [10.3390/cancers9110155](https://doi.org/10.3390/cancers9110155)] [Medline: [29137182](https://pubmed.ncbi.nlm.nih.gov/29137182/)]
8. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proc Natl Acad Sci U S A* 2017 Dec 19;114(38):10202-10207 [FREE Full text] [doi: [10.1073/pnas.1704961114](https://doi.org/10.1073/pnas.1704961114)] [Medline: [28874546](https://pubmed.ncbi.nlm.nih.gov/28874546/)]
9. Zhang X, Hou H, Chen H, Liu Y, Wang A, Hu Q. Serum metabolomics of laryngeal cancer based on liquid chromatography coupled with quadrupole time-of-flight mass spectrometry. *Biomed Chromatogr* 2018 May;32(5):e4181. [doi: [10.1002/bmc.4181](https://doi.org/10.1002/bmc.4181)] [Medline: [29272549](https://pubmed.ncbi.nlm.nih.gov/29272549/)]
10. Jové M, Collado R, Quiles JL, Ramírez-Tortosa M, Sol J, Ruiz-Sanjuan M, et al. A plasma metabolomic signature discloses human breast cancer. *Oncotarget* 2017 Mar 21;8(12):19522-19533 [FREE Full text] [doi: [10.18632/oncotarget.14521](https://doi.org/10.18632/oncotarget.14521)] [Medline: [28076849](https://pubmed.ncbi.nlm.nih.gov/28076849/)]
11. Lacombe J, Mangé A, Jarlier M, Bascoul-Molleivi C, Rouanet P, Lamy P, et al. Identification and validation of new autoantibodies for the diagnosis of DCIS and node negative early-stage breast cancers. *Int J Cancer* 2013 Mar 01;132(5):1105-1113 [FREE Full text] [doi: [10.1002/ijc.27766](https://doi.org/10.1002/ijc.27766)] [Medline: [22886747](https://pubmed.ncbi.nlm.nih.gov/22886747/)]
12. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat Rev Cancer* 2016 Dec;16(5):275-287 [FREE Full text] [doi: [10.1038/nrc.2016.36](https://doi.org/10.1038/nrc.2016.36)] [Medline: [27079802](https://pubmed.ncbi.nlm.nih.gov/27079802/)]
13. Hannoun-Levi J, Ginot A, Thariat J. [Prostate specific antigen: utilization modalities and interpretation]. *Cancer Radiother* 2008 Dec;12(8):848-855. [doi: [10.1016/j.canrad.2008.04.007](https://doi.org/10.1016/j.canrad.2008.04.007)] [Medline: [18539498](https://pubmed.ncbi.nlm.nih.gov/18539498/)]
14. Maddalo G, Fassan M, Cardin R, Piciocchi M, Marafatto F, Rugge M, et al. Squamous Cellular Carcinoma Antigen Serum Determination as a Biomarker of Barrett Esophagus and Esophageal Cancer: A Phase III Study. *J Clin Gastroenterol* 2018;52(5):401-406. [doi: [10.1097/MCG.0000000000000790](https://doi.org/10.1097/MCG.0000000000000790)] [Medline: [28422774](https://pubmed.ncbi.nlm.nih.gov/28422774/)]
15. Hannafon BN, Trigos YD, Calloway CL, Zhao YD, Lum DH, Welm AL, et al. Plasma exosome microRNAs are indicative of breast cancer. *Breast Cancer Res* 2016 Dec 08;18(1):90 [FREE Full text] [doi: [10.1186/s13058-016-0753-x](https://doi.org/10.1186/s13058-016-0753-x)] [Medline: [27608715](https://pubmed.ncbi.nlm.nih.gov/27608715/)]
16. Gyoba J, Shan S, Roa W, Bédard ELR. Diagnosing Lung Cancers through Examination of Micro-RNA Biomarkers in Blood, Plasma, Serum and Sputum: A Review and Summary of Current Literature. *Int J Mol Sci* 2016 Apr 01;17(4):494 [FREE Full text] [doi: [10.3390/ijms17040494](https://doi.org/10.3390/ijms17040494)] [Medline: [27043555](https://pubmed.ncbi.nlm.nih.gov/27043555/)]
17. Pinsky PF, Prorok PC, Kramer BS. Prostate Cancer Screening - A Perspective on the Current State of the Evidence. *N Engl J Med* 2017 Dec 30;376(13):1285-1289. [doi: [10.1056/NEJMs1616281](https://doi.org/10.1056/NEJMs1616281)] [Medline: [28355509](https://pubmed.ncbi.nlm.nih.gov/28355509/)]
18. Kamel MM, Matboli M, Sallam M, Montasser IF, Saad AS, El-Tawdi AHF. Investigation of long noncoding RNAs expression profile as potential serum biomarkers in patients with hepatocellular carcinoma. *Transl Res* 2016 Feb;168:134-145. [doi: [10.1016/j.trsl.2015.10.002](https://doi.org/10.1016/j.trsl.2015.10.002)] [Medline: [26551349](https://pubmed.ncbi.nlm.nih.gov/26551349/)]
19. Pang E, Yang R, Fu X, Liu Y. Overexpression of long non-coding RNA MALAT1 is correlated with clinical progression and unfavorable prognosis in pancreatic cancer. *Tumour Biol* 2015 Apr;36(4):2403-2407. [doi: [10.1007/s13277-014-2850-8](https://doi.org/10.1007/s13277-014-2850-8)] [Medline: [25481511](https://pubmed.ncbi.nlm.nih.gov/25481511/)]
20. deVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, et al. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clin Chem* 2009 Jul;55(7):1337-1346 [FREE Full text] [doi: [10.1373/clinchem.2008.115808](https://doi.org/10.1373/clinchem.2008.115808)] [Medline: [19406918](https://pubmed.ncbi.nlm.nih.gov/19406918/)]
21. Shalaby SM, El-Shal AS, Abdelaziz LA, Abd-Elbary E, Khairy MM. Promoter methylation and expression of DNA repair genes MGMT and ERCC1 in tissue and blood of rectal cancer patients. *Gene* 2018 Feb 20;644:66-73. [doi: [10.1016/j.gene.2017.10.056](https://doi.org/10.1016/j.gene.2017.10.056)] [Medline: [29080834](https://pubmed.ncbi.nlm.nih.gov/29080834/)]
22. Bardelli A, Pantel K. Liquid Biopsies, What We Do Not Know (Yet). *Cancer Cell* 2017 Dec 13;31(2):172-179 [FREE Full text] [doi: [10.1016/j.ccell.2017.01.002](https://doi.org/10.1016/j.ccell.2017.01.002)] [Medline: [28196593](https://pubmed.ncbi.nlm.nih.gov/28196593/)]
23. Cree IA, Uttley L, Buckley Woods H, Kikuchi H, Reiman A, Harnan S, UK Early Cancer Detection Consortium. The evidence base for circulating tumour DNA blood-based biomarkers for the early detection of cancer: a systematic mapping review. *BMC Cancer* 2017 Oct 23;17(1):697 [FREE Full text] [doi: [10.1186/s12885-017-3693-7](https://doi.org/10.1186/s12885-017-3693-7)] [Medline: [29061138](https://pubmed.ncbi.nlm.nih.gov/29061138/)]
24. Zhang K, Shi H, Xi H, Wu X, Cui J, Gao Y, et al. Genome-Wide lncRNA Microarray Profiling Identifies Novel Circulating lncRNAs for Detection of Gastric Cancer. *Theranostics* 2017;7(1):213-227 [FREE Full text] [doi: [10.7150/thno.16044](https://doi.org/10.7150/thno.16044)] [Medline: [28042329](https://pubmed.ncbi.nlm.nih.gov/28042329/)]
25. Toiyama Y, Hur K, Tanaka K, Inoue Y, Kusunoki M, Boland CR, et al. Serum miR-200c is a novel prognostic and metastasis-predictive biomarker in patients with colorectal cancer. *Ann Surg* 2014 Apr;259(4):735-743 [FREE Full text] [doi: [10.1097/SLA.0b013e3182a6909d](https://doi.org/10.1097/SLA.0b013e3182a6909d)] [Medline: [23982750](https://pubmed.ncbi.nlm.nih.gov/23982750/)]

26. Laloglu E, Kumtepe Y, Aksoy H, Topdagi Yilmaz EP. Serum endocan levels in endometrial and ovarian cancers. *J Clin Lab Anal* 2017 Sep;31(5). [doi: [10.1002/jcla.22079](https://doi.org/10.1002/jcla.22079)] [Medline: [27734523](https://pubmed.ncbi.nlm.nih.gov/27734523/)]
27. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, TRACERx consortium, PEACE consortium, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 2017 Dec 26;545(7655):446-451 [FREE Full text] [doi: [10.1038/nature22364](https://doi.org/10.1038/nature22364)] [Medline: [28445469](https://pubmed.ncbi.nlm.nih.gov/28445469/)]
28. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015 Mar;65(2):87-108 [FREE Full text] [doi: [10.3322/caac.21262](https://doi.org/10.3322/caac.21262)] [Medline: [25651787](https://pubmed.ncbi.nlm.nih.gov/25651787/)]
29. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015 Mar 1;136(5):E359-E386. [doi: [10.1002/ijc.29210](https://doi.org/10.1002/ijc.29210)] [Medline: [25220842](https://pubmed.ncbi.nlm.nih.gov/25220842/)]
30. Leng S, Zheng J, Jin Y, Zhang H, Zhu Y, Wu J, et al. Plasma cell-free DNA level and its integrity as biomarkers to distinguish non-small cell lung cancer from tuberculosis. *Clin Chim Acta* 2018 Feb;477:160-165. [doi: [10.1016/j.cca.2017.11.003](https://doi.org/10.1016/j.cca.2017.11.003)] [Medline: [29113814](https://pubmed.ncbi.nlm.nih.gov/29113814/)]
31. Li X, Hayward C, Fong P, Dominguez M, Hunsucker SW, Lee LW, et al. A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. *Sci Transl Med* 2013 Oct 16;5(207):207ra142 [FREE Full text] [doi: [10.1126/scitranslmed.3007013](https://doi.org/10.1126/scitranslmed.3007013)] [Medline: [24132637](https://pubmed.ncbi.nlm.nih.gov/24132637/)]
32. Chen X, Hu Z, Wang W, Ba Y, Ma L, Zhang C, et al. Identification of ten serum microRNAs from a genome-wide serum microRNA expression profile as novel noninvasive biomarkers for nonsmall cell lung cancer diagnosis. *Int J Cancer* 2012 Apr 01;130(7):1620-1628 [FREE Full text] [doi: [10.1002/ijc.26177](https://doi.org/10.1002/ijc.26177)] [Medline: [21557218](https://pubmed.ncbi.nlm.nih.gov/21557218/)]
33. Singh VK, Chandra S, Kumar S, Pangtey G, Mohan A, Guleria R. A common medical error: lung cancer misdiagnosed as sputum negative tuberculosis. *Asian Pac J Cancer Prev* 2009;10(3):335-338 [FREE Full text] [Medline: [19640168](https://pubmed.ncbi.nlm.nih.gov/19640168/)]
34. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, et al. Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA* 2012 Jun 13;307(22):2418-2429 [FREE Full text] [doi: [10.1001/jama.2012.5521](https://doi.org/10.1001/jama.2012.5521)] [Medline: [22610500](https://pubmed.ncbi.nlm.nih.gov/22610500/)]
35. National LSTRT, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011 Aug 4;365(5):395-409 [FREE Full text] [doi: [10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873)] [Medline: [21714641](https://pubmed.ncbi.nlm.nih.gov/21714641/)]
36. Breiman L. Random forests. *Machine Learning* 2001 Oct;45(1):5-32. [doi: [10.1023/a:1010933404324](https://doi.org/10.1023/a:1010933404324)]
37. Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. *Bioinformatics* 2015 Jun 15;31(12):i197-i205. [doi: [10.1093/bioinformatics/btv268](https://doi.org/10.1093/bioinformatics/btv268)] [Medline: [26072483](https://pubmed.ncbi.nlm.nih.gov/26072483/)]
38. Bylander T, Hanzlik D. Estimating generalization error using out-of-bag estimates. In: *AAAI-99 Proceedings*. 1999 Presented at: National Conference on Artificial Intelligence; 1999; Orlando, FL.
39. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. *Bioinformatics* 2010 Feb 01;26(3):440-443 [FREE Full text] [doi: [10.1093/bioinformatics/btp621](https://doi.org/10.1093/bioinformatics/btp621)] [Medline: [19880370](https://pubmed.ncbi.nlm.nih.gov/19880370/)]
40. Axelsson J, Sundelin T, Olsson MJ, Sorjonen K, Axelsson C, Lasselin J, et al. Identification of acutely sick people and facial cues of sickness. *Proc Biol Sci* 2018 Jan 10;285(1870) [FREE Full text] [doi: [10.1098/rspb.2017.2430](https://doi.org/10.1098/rspb.2017.2430)] [Medline: [29298938](https://pubmed.ncbi.nlm.nih.gov/29298938/)]
41. Wu J, Li S. ATB Discrimination. 2019 Jul 01. URL: <http://lishuyan.lzu.edu.cn/ATB/ATBdiscrimination.html> [accessed 2019-08-07]
42. Nanavaty P, Alvarez M, Alberts W. Lung cancer screening: advantages, controversies, and applications. *Cancer Control* 2014 Jan;21(1):9-14. [doi: [10.1177/107327481402100102](https://doi.org/10.1177/107327481402100102)] [Medline: [24357736](https://pubmed.ncbi.nlm.nih.gov/24357736/)]
43. Aberle D, DeMello S, Berg C, Black W, Brewer B, Church T, National Lung Screening Trial Research Team. Results of the two incidence screenings in the National Lung Screening Trial. *N Engl J Med* 2013 Sep 05;369(10):920-931 [FREE Full text] [doi: [10.1056/NEJMoa1208962](https://doi.org/10.1056/NEJMoa1208962)] [Medline: [24004119](https://pubmed.ncbi.nlm.nih.gov/24004119/)]
44. Ziaian B, Saberi A, Ghayyoumi M, Safaei A, Ghaderi A, Mojtahedi Z. Association of high LDH and low glucose levels in pleural space with HER2 expression in non-small cell lung cancer. *Asian Pac J Cancer Prev* 2014;15(4):1617-1620 [FREE Full text] [doi: [10.7314/apjcp.2014.15.4.1617](https://doi.org/10.7314/apjcp.2014.15.4.1617)] [Medline: [24641377](https://pubmed.ncbi.nlm.nih.gov/24641377/)]
45. Zhang X, Guo M, Fan J, Lv Z, Huang Q, Han J, et al. Prognostic significance of serum LDH in small cell lung cancer: A systematic review with meta-analysis. *Cancer Biomark* 2016;16(3):415-423. [doi: [10.3233/CBM-160580](https://doi.org/10.3233/CBM-160580)] [Medline: [27062698](https://pubmed.ncbi.nlm.nih.gov/27062698/)]
46. Sprague B, Trentham-Dietz A, Klein B, Klein R, Cruickshanks K, Lee K, et al. Physical activity, white blood cell count, and lung cancer risk in a prospective cohort study. *Cancer Epidemiol Biomarkers Prev* 2008;17(10):2714-2722 [FREE Full text] [doi: [10.1158/1055-9965.EPI-08-0042](https://doi.org/10.1158/1055-9965.EPI-08-0042)] [Medline: [18843014](https://pubmed.ncbi.nlm.nih.gov/18843014/)]
47. Phillips AN, Neaton JD, Cook DG, Grimm RH, Gerald Shaper A. The leukocyte count and risk of lung cancer. *Cancer* 1992 Feb 01;69(3):680-684. [doi: [10.1002/1097-0142\(19920201\)69:3<680::aid-cnrcr2820690314>3.0.co;2-d](https://doi.org/10.1002/1097-0142(19920201)69:3<680::aid-cnrcr2820690314>3.0.co;2-d)] [Medline: [1730118](https://pubmed.ncbi.nlm.nih.gov/1730118/)]
48. Margolis K, Rodabough R, Thomson C, Lopez A, McTiernan A, Women's Health Initiative Research Group. Prospective study of leukocyte count as a predictor of incident breast, colorectal, endometrial, and lung cancer and mortality in

- postmenopausal women. *Arch Intern Med* 2007 Sep 24;167(17):1837-1844. [doi: [10.1001/archinte.167.17.1837](https://doi.org/10.1001/archinte.167.17.1837)] [Medline: [17893304](https://pubmed.ncbi.nlm.nih.gov/17893304/)]
49. Cedrés S, Torrejon D, Martínez A, Martínez P, Navarro A, Zamora E, et al. Neutrophil to lymphocyte ratio (NLR) as an indicator of poor prognosis in stage IV non-small cell lung cancer. *Clin Transl Oncol* 2012;14(11):864-869. [doi: [10.1007/s12094-012-0872-5](https://doi.org/10.1007/s12094-012-0872-5)] [Medline: [22855161](https://pubmed.ncbi.nlm.nih.gov/22855161/)]
 50. Venkatesan R, Salam A, Alawin I, Willis M. Non-small cell lung cancer and elevated eosinophil count: A case report and literature review. *Cancer Treatment Communications* 2015;4:55-58. [doi: [10.1016/j.ctrc.2015.05.002](https://doi.org/10.1016/j.ctrc.2015.05.002)]
 51. Nishio H, Sakuma T, Nakamura S, Horai T, Ikegami H, Matsuda M. Diagnostic value of high molecular weight alkaline phosphatase in detection of hepatic metastasis in patients with lung cancer. *Cancer* 1986 May 01;57(9):1815-1819. [doi: [10.1002/1097-0142\(19860501\)57:9<1815::aid-cnrc2820570918>3.0.co;2-1](https://doi.org/10.1002/1097-0142(19860501)57:9<1815::aid-cnrc2820570918>3.0.co;2-1)] [Medline: [3006908](https://pubmed.ncbi.nlm.nih.gov/3006908/)]
 52. Lee B, Bach P, Horton J, Hickey T, Davis W. Elevated CK-MB and CK-BB in serum and tumor homogenate of a patient with lung cancer. *Clin Cardiol* 1985;8(4):233-236 [FREE Full text] [doi: [10.1002/clc.4960080409](https://doi.org/10.1002/clc.4960080409)] [Medline: [2985311](https://pubmed.ncbi.nlm.nih.gov/2985311/)]

Abbreviations

ACC: accuracy
ALP: alkaline phosphatase
AUC: area under the curve
CK-MB: creatine kinase isoenzymes
ctDNA: circulating tumor DNA
CT: computed tomography
EO#: eosinophil cells
FN: false negative
FP: false positive
LDH: lactate dehydrogenase
MCC: Matthews correlation coefficient
mtry: number of randomly selected features to split at each node
NE%: neutrophil granulocyte ratio
NLST: National Lung Screening Trial
ntree: tree number
RBLC: routine blood indices model for lung cancer
RF: Random Forest method
ROC: receiver operating characteristic
Sens: sensitivity
Spec: specificity
TN: true negative
TP: true positive
WBC: white blood cell count

Edited by G Eysenbach; submitted 25.01.19; peer-reviewed by F Zhu, H Liu, K Pradeep; comments to author 28.04.19; revised version received 12.05.19; accepted 19.07.19; published 15.08.19

Please cite as:

Wu J, Zan X, Gao L, Zhao J, Fan J, Shi H, Wan Y, Yu E, Li S, Xie X
A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study
JMIR Med Inform 2019;7(3):e13476
 URL: <http://medinform.jmir.org/2019/3/e13476/>
 doi: [10.2196/13476](https://doi.org/10.2196/13476)
 PMID: [31418423](https://pubmed.ncbi.nlm.nih.gov/31418423/)

©Jiangpeng Wu, Xiangyi Zan, Liping Gao, Jianhong Zhao, Jing Fan, Hengxue Shi, Yixin Wan, E Yu, Shuyan Li, Xiaodong Xie. Originally published in *JMIR Medical Informatics* (<http://medinform.jmir.org>), 15.08.2019. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <http://medinform.jmir.org/>, as well as this copyright and license information must be included.