frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fonc.2022.924144

Explainable Machine Learning Model to Prediction EGFR Mutation in Lung Cancer

Ruiyuan Yang[1†], *Xingyu Xiong*[1†], *Haoyu Wang*[1] *and Weimin Li*[1,2,3,4*]

[1] Department of Respiratory and Critical Care Medicine, West China Hospital, Sichuan University, Chengdu, China, [2] Institute of Respiratory Health Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, China, [3] Precision Medicine Center, Precision Medicine Key Laboratory of Sichuan Province, West China Hospital, Sichuan University, Chengdu, China, [4] The Research Units of West China, Chinses Academy of Medical Sciences, West China Hospital, Chengdu, China

OPEN ACCESS

Edited by:
Jiayuan Sun,
Shanghai Jiao Tong University, China

Reviewed by:
Yun Gu,
Shanghai Jiao Tong University, China
Chengzhi Zhou,
National Respiratory Medical Center, China

***Correspondence:**
Weimin Li
weimin003@163.com

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Thoracic Oncology, a section of the journal Frontiers in Oncology

Received: *20 April 2022*
Accepted: *16 May 2022*
Published: *23 June 2022*

Citation:
Yang R, Xiong X, Wang H and Li W (2022) Explainable Machine Learning Model to Prediction EGFR Mutation in Lung Cancer. Front. Oncol. 12:924144. doi: 10.3389/fonc.2022.924144

Objectives: The aim of this study is to determine whether the clinical features including blood markers can establish an explainable machine learning model to predict epidermal growth factor receptor (EGFR) mutation in lung cancer.

Methods: We retrospectively analyzed 7,413 patients with lung adenocarcinoma (LA) diagnosed by gene sequencing in West China Hospital of the Sichuan University from April 2015 to June 2019. The machine learning algorithms (MLAs) included logistic regression (LR), random forest (RF), LightGBM, support vector machine (SVM), multi-layer perceptron (MLP), extreme gradient boosting (XGBoost), and decision tree (DT). Demographic characteristics, personal history, and blood markers were taken into. The area under the receiver operating characteristic curve (AUC) and SHapley Additive exPlanation (SHAP) value were used to explain the prediction models.

Results: Of the 7,413 patients with LA (47.6%), 3,527 were identified with EGFR mutation; RF achieved greatest performance in predicting EGFR mutation AUC [0.771, 95% confidence interval (CI): 0.770, 0.772], which was like XGBoost with AUC (0.740, 95% CI: 0.739, 0.741). The five most influential features were smoking consumption, sex, cholesterol, age, and albumin globulin ratio. The SHAP summary and dependence plot have been used to explain the affection of the 12 features to this model and how a single feature influences the output, respectively.

Conclusion: We established EGFR mutation prediction models by MLAs and revealed that the RF was preferred, AUC (0.771, 95% CI: 0.770, 0.772), which was better than the traditional models. Therefore, the artificial intelligence–based MLA predicting model may become a practical tool to guide in diagnosis and therapy of LA.

Keywords: EGFR mutation, lung cancer, prediction, machine learning, SHAP value

INTRODUCTION

Lung cancer has become a commonly diagnosed tumor, which accounts for approximately 11.4% of all cancers diagnosed and 18% of cancer-related death (1–4), which induced a high economic burden and life loss. The first choice is still surgical resection when lung cancer occurs and stage II or III patients also receive adjuvant therapy followed by surgery (5, 6) for decreasing the possibility of progression and relapse and further increased the progression-free survival (7, 8). However, many patients have missed the chance of surgical therapy with a visit to doctors because of the symptoms. They usually cannot access surgery; usually, chemotherapy or radiotherapy is preferred for this part of unresectable or inoperable patients (7). The traditional plan is Cisplatin-based adjuvant chemotherapy. With the rapidly evolving treatment landscape, tyrosine kinase inhibitors (TKIs) have an expanding place in lung cancer therapy (6), as first-line or adjuvant treatment plans. Patients get greater clinical benefit compared with traditional chemotherapy, for those who were confirmed with targetable gene mutations, like Epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK) (7, 8).

EGFR, a tyrosine kinase receptor, has a mutation rate of about 10% in lung adenocarcinoma (LA) and an unignorable higher rate in non-smoking patients. It is the earliest gene to be uncovered, and EGFR-TKIs were adopted for clinical work, which markedly changed the therapy strategy and yielded better therapeutic prospects in lung cancer, particularly in adenocarcinoma (9, 10). The diagnosis still depends on tumor tissue biopsy (11), through broncho fiberscope, puncture biopsy, or surgery, which have significant risks, such as surgery-related and time costs. Hence, non-invasive and fast method to confirm EGFR mutation is needed in clinical work. From logistic regression (LR) to machine learning, different methods were adopted to evaluation genotype mutation with AUC ranging from 0.65 to 0.75. Recent studies made use of CT or PET-CT images, and the sensitivity was increased to 0.81 (12).

In the study, an explainable model finds the significant influential factors for EGFR mutation with SHAP value. We utilized retrospective data including demographic characteristics and clinical examination in a large sample of patients with LA and finally selected the model with the best performance. It is expected that this type of model would be used for reference by clinicians.

MATERIALS AND METHODS

Study Participants

We retrospectively collected the clinical records of 7,413 patients who underwent LA diagnosis in West China Hospital of the Sichuan University, from April 2015 to June 2019.

Statistical Methods

Continuous data are presented as the mean ± standard deviation (SD) or median (Q1, Q3), and categorical data are described as numbers (%). Student's t-test or one-way ANOVA (analysis of variance) was used for normally distributed continuous variables. The Newman–Keuls or Student–Newman–Keuls method was used for multiple samples. The Kruskal–Wallis test was used for the non-normally distributed continuous variables. The Chi-square test and the Fisher discriminant analysis were used to evaluate the difference in categorical variables such as sex and smoking status. We assessed the predictive performance according to the AUC, to evaluate the prediction and accuracy of various machine learning models abovementioned in the test set, and bootstrap methods with 1,000 bootstrap replicates were used to derive 95% confidence interval (CI).

Machine Learning Models

We used the following supervised machine learning methods to develop the predictive models, which were novel and traditional machine learning methods used for the problem of classification: LR, random forest (RF), LightGBM, support vector machine (SVM), multi-layer perceptron (MLP), extreme gradient boosting (XGBoost), and decision tree. LR is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieve satisfied performance with linearly separable classes. It is an extensively employed algorithm for classification in medical study. Multilayer perception, known as artificial neural network (ANN), is a model that is inspired by the human brain and the way it functions. A standard ANN has an input layer, an output layer, and at least one hidden layer between input and output. ANN always has several layers of nodes, definite link patterns and layer connections, connection weights, and node (neuron) and activation functions that map weighted inputs to outputs. Throughout the training process, the weights are changed. The backpropagation algorithm is a technique to train ANNs, and it has the following two key stages: propagation and weight update. SVMs are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. It performs well in high-dimensional spaces and is still effective in cases where the number of dimensions is greater than the number of samples. In other hand, it uses a subset of training points in the decision function (called support vectors), so it is also memory-efficient. Decision tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In decision analysis, a decision tree can be visually and explicitly used to represent decisions and decision-making. Furthermore, decision tree is the basis of the following three models: RF, XGBoost, and LightGBM. RF consists of many random decision trees. It uses random sample of original data and random subset features to build the model. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way. Three main forms of gradient boosting are supported: gradient boosting algorithm, stochastic gradient boosting, and regularized gradient boosting.

In the end, LightGBM also uses tree-based learning algorithms. It can be used in classification, regression, and many more machine learning tasks. It is designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency, support of parallel learning, and capable of handling large-scale data. All data of 7,413 patients were included, in which, among the seven models, the best model to predict EGFR mutation is found (13–15).

RESULTS

Of the 30,052 patients with lung cancer, we gathered the clinical statistics of 7,413 patients with LA who meet the inclusion criteria from April 2015 to June 2019 in West China Hospital. All 7,413 patients with LA are confirmed by gene examination. Among them, 3,527 patients (47.6%) were diagnosed with EGFR mutation, and the remaining 3,886 patients were negative ones, in which the incidence was consistent with previous studies in Asia (16). Moreover, the mean age was 56.93 years. EGFR mutation cases were typically women, with less exposure to smoke and drink compared with controls (all $P < 0.001$). Ninety-one demographics and blood markers were included, and all features were from the first visit. In addition, nearly all variables, the differences between the train and test sets, were non-significant. Because of a large number of parameters included, we only show parameters included in our final model in **Table 1**. The full information could be found in **Table S1**.

We established final models with 12 features elected by the least absolute shrinkage and selection operator (LASSO) algorithm and clinical experience using machine learning algorithms (MLAs) above, including smoking consumption, sex, cholesterol, age, albumin globulin ratio, glutamyl transpeptidase, hemoglobin, carcinoembryonic antigen (CEA), platelet, neutrophils, lymphocytes, and aspartate aminotransferase. In the test set, RF achieved great performance in terms of predicting EGFR mutation with AUC (0.771, 95% CI: 0.770, 0.772), which was similar to XGBoost with AUC (0.740, 95% CI: 0.739, 0.741). The quantitative performance and the ROC curves had been established on **Table 2** and **Figure 1**.

To elucidate the features that influenced this prediction model further, we adopted the SHAP summary value of RF. This figure shows if the features were strongly connected to EGFR mutation. In other words, the higher the SHAP value, the greater possibility of EGFR mutation. As shown in the picture,

TABLE 1 | Patient characteristics and blood markers.

Variables	EGFR–Wild Type	EGFR-Mutation	P-Value
Patient population, n (n%)	3,886	3,527	
Demographic data			
Gender, n (%)			<0.001
Female	1,446 (37.210)	1,970 (55.855)	
Male	2,440 (62.790)	1,557 (44.145)	
Smoking Consumption, n (%)			<0.001
No	1,814 (46.873)	2,490 (71.001)	
Yes	2,056 (53.127)	1,017 (28.999)	
Age (year), mean (SD)	56.965 (10.749)	56.893 (10.197)	0.768
Blood routine			
Hemoglobin (g/L), mean (SD)	120.959 (17.833)	122.760 (17.470)	<0.001
Platelet(10^9/L), mean (SD)	215.069 (83.880)	211.581 (81.704)	0.079
Neutrophils%, mean (SD)	64.071 (12.673)	63.993 (11.888)	0.794
Lymphocyte%, mean (SD)	24.315 (10.758)	24.825 (10.108)	0.043
Blood biochemistry			
Cholesterol (mmol/L), mean (SD)	4.773 (1.006)	4.726 (1.002)	0.053
Albumin Globulin Ratio, mean (SD)	1. 527 (0.329)	1.587 (0.320)	<0.001
Glutamyl Transpeptidase (IU/L), mean (SD)	36.925 (22.541)	33.808 (22.600)	<0.001
Aspartate Aminotransferase (IU/L), mean (SD)	24.854 (8.265)	24.952 (8.444)	0.623
Tumor markers			
Carcinoembryonic Antigen (ng/ml), median [Q1, Q3]	6.120 [2.440, 25.348]	6.640 [2.300, 30.797]	0.881

TABLE 2 | The quantitative performance and the ROC curves of included models.

Model	AUC	Youden_Index	Sensitivity	Specificity
RF	0.825 (0.823, 0.827)	0.510 (0.506, 0.514)	0.738 (0.736, 0.74)	0.752 (0.75, 0.754)
XGBoost	0.826 (0.824, 0.828)	0.513 (0.509, 0.517)	0.749 (0.747, 0.751)	0.751 (0.749, 0.753)
LightGBM	0.819 (0.817, 0.821)	0.517 (0.512, 0.522)	0.749 (0.746, 0.752)	0.751 (0.748, 0.754)
Decision Tree	0.648 (0.647, 0.649)	0.277 (0.275, 0.279)	0.306 (0.305, 0.307)	0.804 (0.803, 0.805)
LR	0.695 (0.693, 0.697)	0.299 (0.295, 0.303)	0.633 (0.631, 0.635)	0.636 (0.634, 0.638)
SVM	0.795 (0.793, 0.797)	0.472 (0.468, 0.476)	0.719 (0.716, 0.722)	0.727 (0.725, 0.729)
MLP	0.774 (0.772, 0.776)	0.442 (0.437, 0.447)	0.711 (0.708, 0.714)	0.714 (0.711, 0.717)

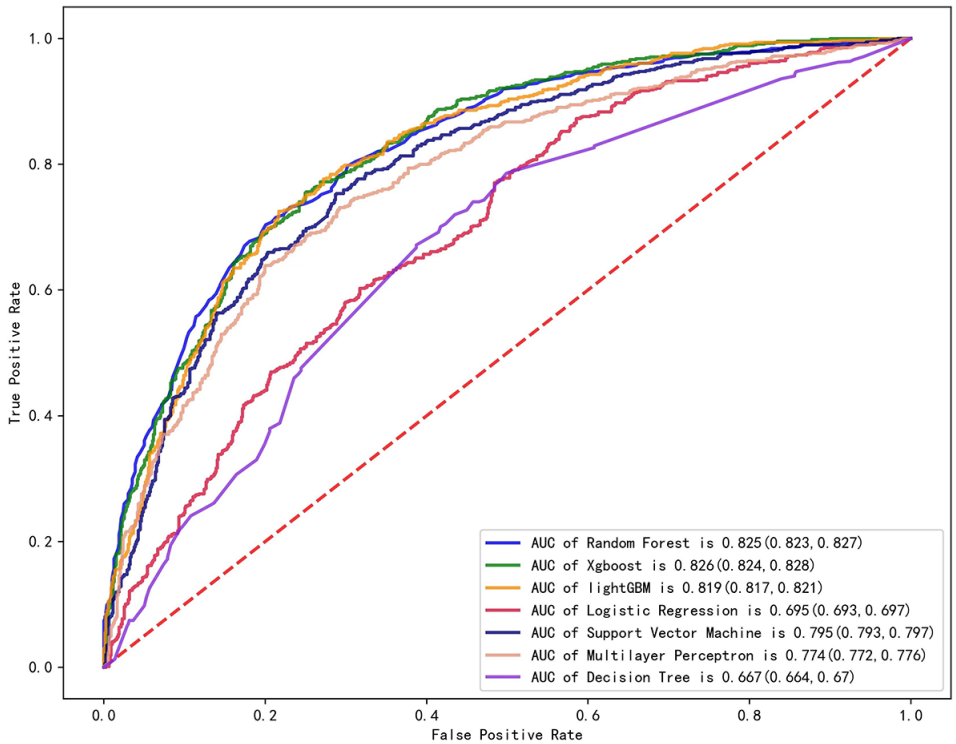

FIGURE 1 | Comparison of AUCs among seven machine learning models with ROC; RF got the greatest AUC for single model prediction.

smoking consumption, gender, cholesterol, age, and albumin globulin ratio ranked as the top five among all variables. The SHAP dependence plot can also be used to explain how a single factor affects the result of the model. The y-axis shows the SHAP value of single feature, and the value of different features is showed in the x-axis. We could vividly describe the tendency of each feature with the changing plots. SHAP value for specific features exceeding zero represents an increased risk of incidence of EGFR mutation.

DISCUSSION

Lung cancer is traditionally divided into two broad histologic categories: non–small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC represents more than 80% to 85% of lung cancers, of which approximately 40% are adenocarcinoma (11). In a development of precision medicine, we have found the important influence of gene mutation in oncotherapy and prognosis. EGFR mutation is one of the most common genotypes of lung cancer and occurs in at least 50% of NSCLC in Asia (17). TKIs show marked clinical benefits in EGFR mutation patients, compared with the conventional chemical therapy (18). In clinical work, gene examination requires biopsy and sequence testing, which cannot be detected in some cases, because of insufficient samples, the financial situation of patients, and so on. This impedes patients' therapy and prognosis in some degree. Moreover, obtaining samples usually depends on invasive methods, which have risk of bleeding, excessive damage, and so on. Hence, non-invasive method to predict EGFR mutation is pressing. In the present prediction models, the LR is commonly used, with the AUC ranging from 0.6 to 0.8 (12, 19).

In this retrospective study, we included more advanced MLAs using 91 features and filtrated models by LASSO algorithm and clinical experience. It indicated that RF gained better performance (**Figure 1**). To describe visually, we used a bar graph (**Figure 2**) to interpret the discrepancy between the traditional and novel machine learning methods, and we found that RF, XGBoost, and light GBM were superior compared with the traditional ones, like decision tree, multilayer perceptron, and SVM. The RF model got the best performance, whereas the LR model got the worst.

Furthermore, our SHAP summary plot (**Figure 3**) provided an explicable machine learning model to predict EGFR mutation in LA with large scales of samples and clinical features. Smoking consumption, sex, cholesterol, age, and albumin globulin ratio were in the top five variables related to EGFR mutation. From the previous epidemiological studies, it has been found that non-smoking characteristics among female patients are non-negligible (18, 20–23); consistent with our study, female patients who are non-smokers were more likely to have EGFR mutation (**Figures 4A, B**). Interestingly, we found that people under 40 years old were not prone to this mutation, but with age, the trend of change has no significant characteristics

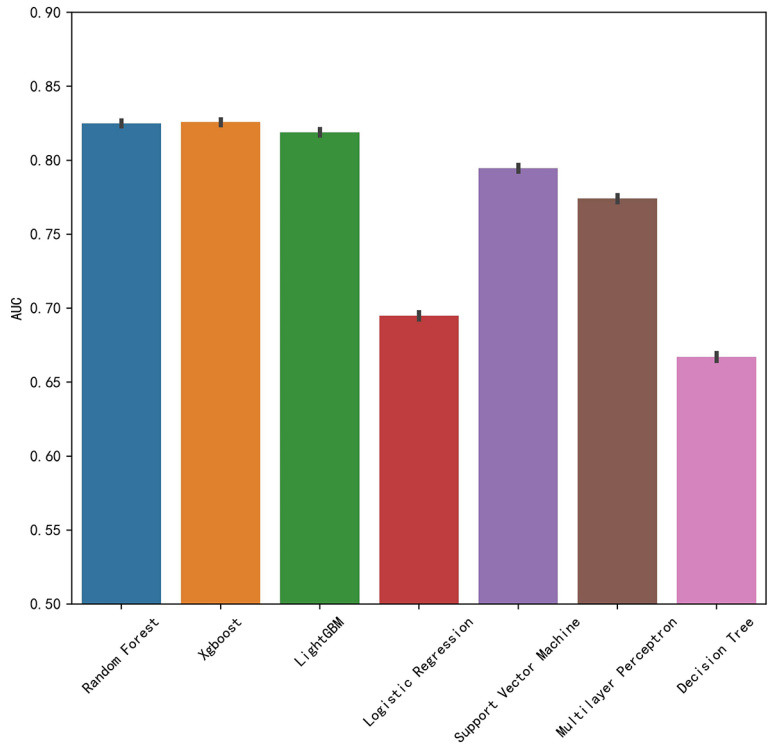

FIGURE 2 | Comparison of AUCs among seven machine learning models with bar graph.

(**Figure 4C**). The possible reason is that there are more co-influential factors such as aging, so the independent age change cannot exhibit a specific effect. As for metabolic indicators, mutation risk is elevated for participants with lower cholesterol and higher albumin globulin ratio (**Figures 5A, B**). The increased albumin globulin ratio might be explained by suppressed immune function or high and aberrant expression of normal proteins (24), which were called tumor-associated antigens. The relationship and the specific mechanism need further studies. Increased CEA (25, 26), which is often

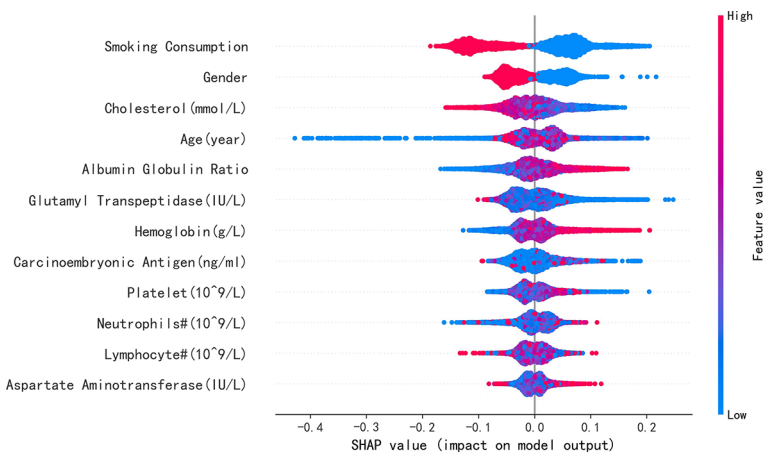

FIGURE 3 | SHAP summary plot of the 12 features of the RF model. The higher the SHAP value of single feature, the higher the possibility of EGFR mutation. Red represents closer with this mutation, and blue represents apposite possibility.

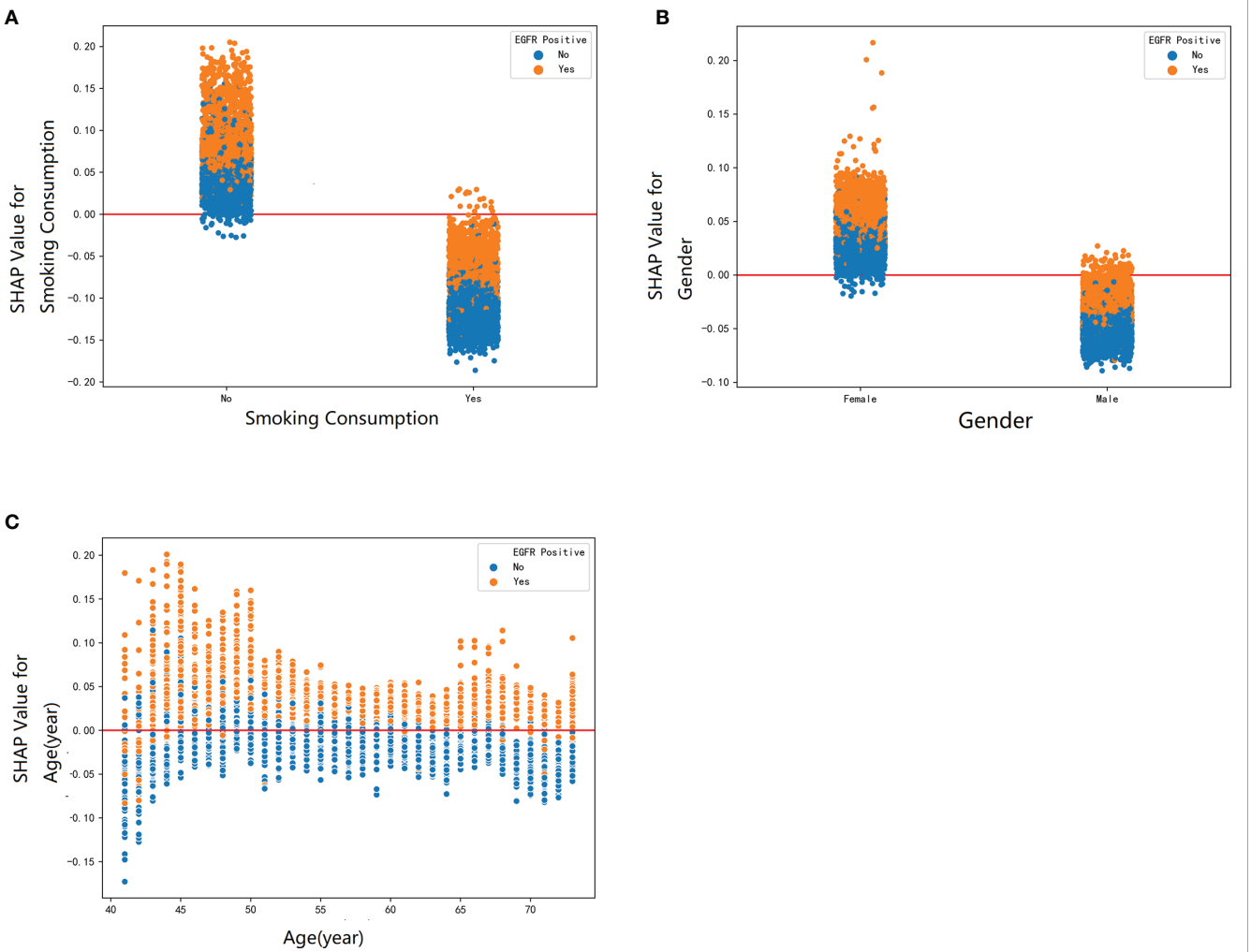

FIGURE 4 | The SHAP dependence plot was used to explain how a single factor affects the result in this RF model. SHAP value for specific feature exceeding zero represents an increased risk of incidence of EGFR mutation. The demographic factors: **(A)** smoking consumption, **(B)** gender, and **(C)** age.

observed in LA, was also included. We found that CEA (**Figure 5C**) was associated with the mutation but we also easily saw that there was no significant connection with the change, similar to glutamyl transpeptidase, neutrophils, and lymphocytes. As for platelet, we could also see that patients with platelets exceeding the normal range were close to EGFR mutation (**Figure 5D**). Lots of research had provided evidence of venous thromboembolism incidence in EGFR mutation patients and lead to an inconsistent conclusion (27–29). In our study, the result supported the positive change. Differences between genotypes and therapy choices need further study.

Despite the encouraging performance of our model, this study also has several limitations. First, although the sample size was relatively large in our retrospective study, we only examined patients in West China Hospital; moreover, it presented partial information about LA in the Western region. Hence, there was indeed bias in this study, and further multicenter and prospective study in future is needed. Second, we did not include the imaging statistics in our predicting model. Third, only EGFR mutation was included in our study; in future work, the relationship between EGFR and other genetic mutations, like c-ros oncogene 1 (ROS-1), ALK, and Kirsten rat sarcoma viral oncogene (KRAS), or different subtypes of EGFR, can be explored.

CONCLUSION

In brief, accurate and rapid EGFR discrimination is valuable for patients, both resectable and inaccessible to surgery. Traditional diagnosis methods are limited to a large proportion of patients. We established the explainable machine model to predict EGFR

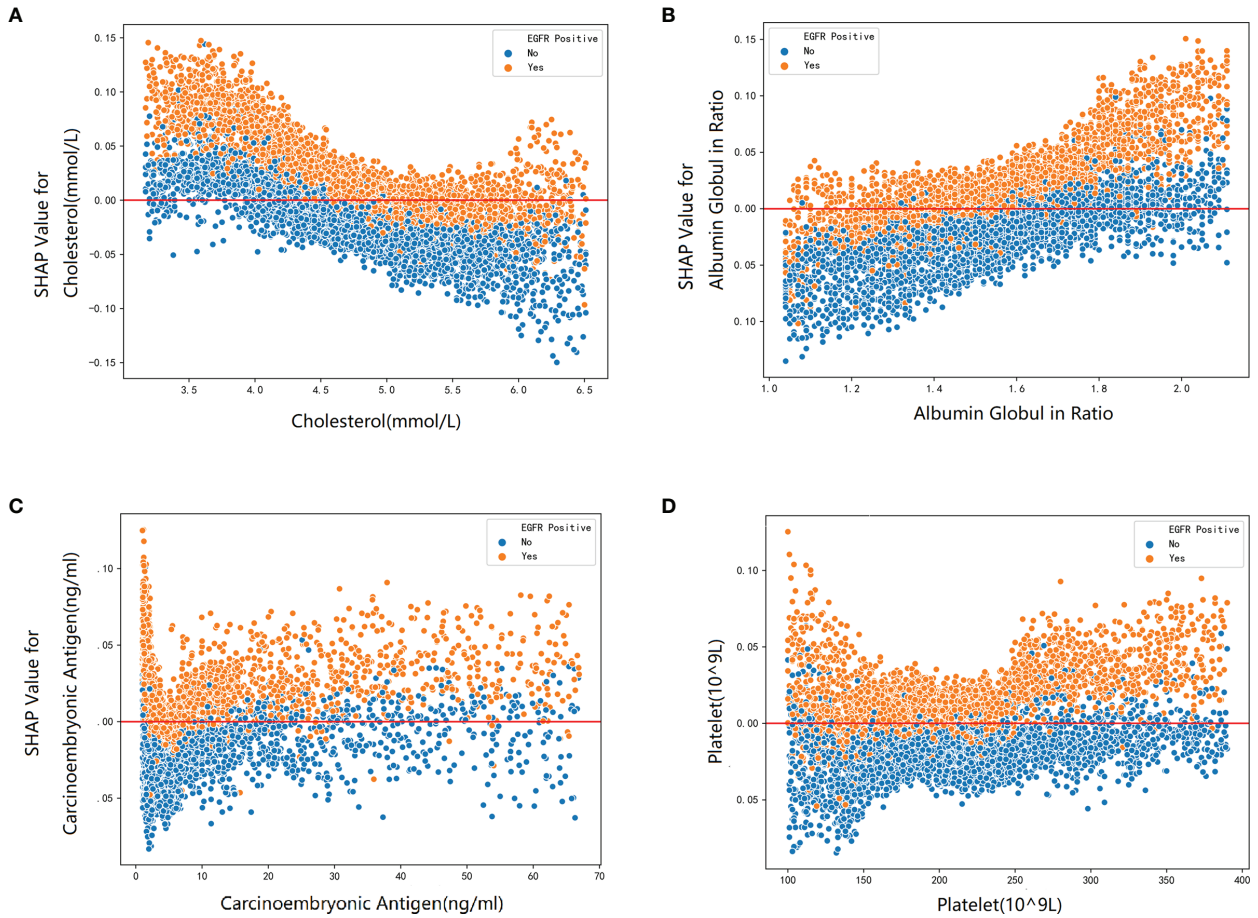

FIGURE 5 | The SHAP dependence plot about blood markers: **(A)** cholesterol, **(B)** albumin globulin ratio, **(C)** carcinoembryonic, and **(D)** platelet.

mutation in patients with LA, which can be referred to in the clinical diagnosis as a non-invasive method, which may guide or assist treatment of patients who lack pathologic diagnosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee on Biomedical Research, West China Hospital of Sichuan University. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

WL devised concept of article. RY wrote the manuscript. XX and HW processed the data. XX and WL reviewed the article. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (Nos. 92159302, 81871890, and 91859203 to WL).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.924144/full#supplementary-material

REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin* (2021) 71(1):7–33. doi: 10.3322/caac.21654
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71 (3):209–49. doi: 10.3322/caac.21660
3. Qiu H, Cao S, Xu R. Cancer Incidence, Mortality, and Burden in China: A Time-Trend Analysis and Comparison With the United States and United Kingdom Based on the Global Epidemiological Data Released in 2020. *Cancer Commun (Lond)* (2021) 41(10):1037–48. doi: 10.1002/cac2.12197
4. Kocarnik JM, Compton K, Dean FE, Fu W, Gaw BL, Harvey JD, et al. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncol* (2022) 8(3):420–44. doi: 10.1001/jamaoncol.2021.6987
5. Chaft JE, Shyr Y, Sepesi B, Forde PM. Preoperative and Postoperative Systemic Therapy for Operable Non-Small-Cell Lung Cancer. *J Clin Oncol: Off J Am Soc Clin Oncol* (2022) 40(6):546–55. doi: 10.1200/JCO.21.01589
6. Saw SPL, Ong B-H, Chua KLM, Takano A, Tan DSW. Revisiting Neoadjuvant Therapy in Non-Small-Cell Lung Cancer. *Lancet Oncol* (2021) 22(11):e501–e16. doi: 10.1016/S1470-2045(21)00383-1
7. Miller M, Hanna N. Advances in Systemic Therapy for Non-Small Cell Lung Cancer. *BMJ (Clin Res ed)* (2021) 375:n2363. doi: 10.1136/bmj.n2363
8. Chaft JE, Rimner A, Weder W, Azzoli CG, Kris MG, Cascone T. Evolution of Systemic Therapy for Stages I-Iii Non-Metastatic Non-Small-Cell Lung Cancer. *Nat Rev Clin Oncol* (2021) 18(9):547–57. doi: 10.1038/s41571-021-00501-4
9. Robichaux JP, Elamin YY, Tan Z, Carter BW, Zhang S, Liu S, et al. Mechanisms and Clinical Activity of an Egfr and Her2 Exon 20-Selective Kinase Inhibitor in Non-Small Cell Lung Cancer. *Nat Med* (2018) 24(5):638–46. doi: 10.1038/s41591-018-0007-9
10. Meador CB, Sequist LV, Piotrowska Z. Targeting Exon 20 Insertions in Non-Small Cell Lung Cancer: Recent Advances and Clinical Updates. *Cancer Discovery* (2021) 11(9):2145–57. doi: 10.1158/2159-8290.CD-21-0226
11. Schabath MB, Cote ML. Cancer Progress and Priorities: Lung Cancer. *Cancer Epidemiol Biomarkers Prev: Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol* (2019) 28(10):1563–79. doi: 10.1158/1055-9965.EPI-19-0221
12. Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, et al. Predicting Egfr Mutation Status in Lung Adenocarcinoma on Computed Tomography Image Using Deep Learning. *Eur Respir J* (2019) 53(3):1800986. doi: 10.1183/13993003.00986-2018
13. H Xia, X Wei, Y Gao and H Lv eds. (2019). Traffic Prediction Based on Ensemble Machine Learning Strategies With Bagging and Lightgbm, in: *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. (2019), 1-6, doi: 10.1109/ICCW.2019.8757058
14. Livingston F. Implementation of Breiman's Random Forest Machine Learning Algorithm. *Ece591q Mach Learn J Paper* (2005) 1-13.
15. D Wang, Z Yang and Z Yi eds. (2017). Lightgbm: An Effective Mirna Classification Method, in: *in Breast Cancer Patients. Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics;* 2017
16. Zhang Y-L, Yuan J-Q, Wang K-F, Fu X-H, Han X-R, Threapleton D, et al. The Prevalence of Egfr Mutation in Patients With Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis. *Oncotarget* (2016) 7(48):78985–93. doi: 10.18632/oncotarget.12587
17. Harrison PT, Vyse S, Huang PH. Rare Epidermal Growth Factor Receptor (Egfr) Mutations in Non-Small Cell Lung Cancer. *Semin Cancer Biol* (2020) 61:167–79. doi: 10.1016/j.semcancer.2019.09.015
18. Duma N, Santana-Davila R, Molina JR. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. *Mayo Clin Proc* (2019) 94(8):1623–40. doi: 10.1016/j.mayocp.2019.01.013
19. Mohapatra PR, Punatar S, Prabhash K. Nomogram to Predict the Presence of Egfr Activating Mutation in Lung Adenocarcinoma. *Eur Respir J* (2012) 39 (6):1550–1. doi: 10.1183/09031936.00022112
20. Shi Y, Au JS-K, Thongprasert S, Srinivasan S, Tsai C-M, Khoa MT, et al. A Prospective, Molecular Epidemiology Study of Egfr Mutations in Asian Patients With Advanced Non-Small-Cell Lung Cancer of Adenocarcinoma Histology (Pioneer). *J Thorac Oncol* (2014) 9(2):154–62. doi: 10.1097/JTO.0000000000000033
21. Kawaguchi T, Koh Y, Ando M, Ito N, Takeo S, Adachi H, et al. Prospective Analysis of Oncogenic Driver Mutations and Environmental Factors: Japan Molecular Epidemiology for Lung Cancer Study. *J Clin Oncol: Off J Am Soc Clin Oncol* (2016) 34(19):2247–57. doi: 10.1200/JCO.2015.64.2322
22. Jemal A, Miller KD, Ma J, Siegel RL, Fedewa SA, Islami F, et al. Higher Lung Cancer Incidence in Young Women Than Young Men in the United States. *N Engl J Med* (2018) 378(21):1999–2009. doi: 10.1056/NEJMoa1715907
23. Chapman AM, Sun KY, Ruestow P, Cowan DM, Madl AK. Lung Cancer Mutation Profile of Egfr, Alk, and Kras: Meta-Analysis and Comparison of Never and Ever Smokers. *Lung Cancer (Amsterdam Netherlands)* (2016) 102:122–34. doi: 10.1016/j.lungcan.2016.10.010
24. Kumagai S, Koyama S, Nishikawa H. Antitumour Immunity Regulated by Aberrant Erbb Family Signalling. *Nat Rev Cancer* (2021) 21(3):181–97. doi: 10.1038/s41568-020-00322-0
25. Duffy MJ, O'Byrne K. Tissue and Blood Biomarkers in Lung Cancer: A Review. *Adv Clin Chem* (2018) 86:1–21. doi: 10.1016/bs.acc.2018.05.001
26. Kulpa J, Wójcik E, Reinfuss M, Kołodziejski L. Carcinoembryonic Antigen, Squamous Cell Carcinoma Antigen, Cyfra 21-1, and Neuron-Specific Enolase in Squamous Cell Lung Cancer Patients. *Clin Chem* (2002) 48(11):1931–7. doi: 10.1093/clinchem/48.11.1931
27. Wang J, Hu B, Li T, Miao J, Zhang W, Chen S, et al. The Egfr-Rearranged Adenocarcinoma Is Associated With a High Rate of Venous Thromboembolism. *Ann Trans Med* (2019) 7(23):724. doi: 10.21037/atm.2019.12.24
28. Davidsson E, Murgia N, Ortiz-Villalón C, Wiklundh E, Sköld M, Kölbeck KG, et al. Mutational Status Predicts the Risk of Thromboembolic Events in Lung Adenocarcinoma. *Multidiscip Respir Med* (2017) 12:16. doi: 10.1186/s40248-017-0097-0
29. Dou F, Li H, Zhu M, Liang L, Zhang Y, Yi J, et al. Association Between Oncogenic Status and Risk of Venous Thromboembolism in Patients With Non-Small Cell Lung Cancer. *Respir Res* (2018) 19(1):88. doi: 10.1186/s12931-018-0791-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.