Cancer Prediction using Naive Bayes Classifier

MINI PROJECT REPORT

Submitted by

Vigneshwaran G (210701307)

Jagathratchahan V (210701701)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE,

CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Cancer Prediction using Naive Bayes Classifier** " is the bonafide work of "***Vigneshwaran G (210701307),Jagathratchahan V (210701701)***" who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

MR.Karthick V

Assistant Professor,

Department of Computer Science and Engineering,

Rajalakshmi Engineering College,

Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

Abstract

In response to the high incidence and poor prognosis of lung cancer, this study tends to develop a generalizable lung-cancer prediction model by using machine learning to define high-risk groups and realize the early identification and prevention of lung cancer. We included 467,888 participants from UK Biobank, using lung cancer incidence as an outcome variable, including 49 previously known high-risk factors and less studied or unstudied predictors. We developed multivariate prediction models using multiple machine learning models, namely logistic regression, naïve Bayes, random forest, and extreme gradient boosting models. The performance of the models was evaluated by calculating the areas under their receiver operating characteristic curves, Brier loss, log loss, precision, recall, and F1 scores. The Shapley additive explanations interpreter was used to visualize the models. Three were ultimately 4299 cases of lung cancer that were diagnosed in our sample. The model containing all the predictors had good predictive power, and the extreme gradient boosting model had the best performance with an area under curve of 0.998. New important predictive factors for lung cancer were also identified, namely hip circumference, waist circumference, number of cigarettes previously smoked daily, neuroticism score, age, and forced expiratory volume in 1 second. The predictive model established by incorporating novel predictive factors can be of value in the early identification of lung cancer. It may be helpful in stratifying individuals and selecting those at higher risk for inclusion in screening programs.

Abbreviations: AUC = area under curve, BMI = body mass index, CRP = C-reactive protein, FEV1 = forced expiratory volume in 1 second, LC = lung cancer, LDCT = low-dose computed tomography, LLP = Liverpool Lung Cancer Project, LR = logistic regression, ML = machine learning, RF = random forest, SHAP = SHapley Additive exPlanations, UKB = UK Biobank, XGBoost = eXtreme gradient boosting.

Keywords: latent factor, lung cancer, machine learning, noninvasive, prediction model

1. Introduction

According to the latest data from the International Agency for Research on Cancer, lung cancer (LC) is the second most common cancer and the leading cause of cancer deaths worldwide.[1] As lung cancer has no obvious symptoms in early stage, most patients are delayed in seeking treatment after being diagnosed in the middle or late stage. The five-year survival rate of stage I LC is as high as 90% after treatment, while the five-year survival rate of stage IV LC can be <10%.[2] This suggests that early detection is an additional condition for radical treatment and can offer a good prognosis for LC patients.

Early screening, diagnosis and treatment of LC are critical secondary preventive measures to lower the risk of mortality from lung cancer. The most efficient screening approach for the prevention of LC is low-dose computed tomography (LDCT), which has been proven to lower the mortality rate of LC.[3] However, using LDCT for LC screening may lead to issues including false positives, radiation risks, overdiagnosis,

This work was supported by the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant [No. LHDMD23H160001]; and Ningbo Top Medical and Health Research Program [No. 2022030208].

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Supplemental Digital Content is available for this article.

[a] The Second School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China, [b] Department of Cardiothoracic Surgery, Ningbo No. 2 Hospital, Ningbo, China, [c] School of Public Health, Medical College of Soochow University, Suzhou, China, [d] Center for Cardiovascular and Cerebrovascular Epidemiology and Translational Medicine, Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo, China.

** Correspondence: Ting Cai, Center for Cardiovascular and Cerebrovascular Epidemiology and Translational Medicine, Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo 315000, China (e-mail: caiting12316@126.com).*

How to cite this article: Zhang S, Yang L, Xu W, Wang Y, Han L, Zhao G, Cai T. Predicting the risk of lung cancer using machine learning: A large study based on UK Biobank. Medicine 2024;103:16(e37879).

Received: 29 January 2024 / Received in final form: 25 February 2024 / Accepted: 21 March 2024

http://dx.doi.org/10.1097/MD.0000000000037879

and overtreatment.[4] A bigger financial burden might also come from widespread public usage of LDCT. Therefore, accurate identification of LC high-risk groups can improve the detection rate of LC and optimize resource allocation.

Predictive models, which can help with the initial screening of high-risk groups in the general population, use characteristics with particular predictive significance to assess the likelihood that a person would acquire LC in a given time frame. Previous classic LC risk prediction models such as the Bach model,[5] the Liverpool Lung Cancer Project (LLP) risk model,[6] and the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial lung-cancer risk- prediction model[7] were developed based on identified high-risk factors such as smoking, occupational exposure, and family history of lung cancer, etc, and the study population was mostly limited to smokers. Subsequently, researchers have attempted to improve the identification of lung cancer by adding some new predictors, such as genetic, clinical indicators, and imaging.[8–10] Additionally, there are still some LC prediction models with defects such as single method and low grade of evidence-based medicine. With the continuous development of medical data, machine learning (ML) has made good progress in the clinical practice and research of LC,[11,12] which brings a new direction for the prediction of LC.

In this study, we aimed to incorporate these variables based on UK Biobank (UKB) data into a new, wider range of variables, and then develop ML models and explain the association between risk factors and LC. This will help medical practitioners to quickly and accurately identify people at high risk of LC for targeted intervention and health management, and will also help patients better understand their risk of developing the disease more clearly and improve their awareness of disease risk factors.

2. Materials and methods

2.1. Data

The data used in this study were obtained from UKB, a large biomedical database and research resource. Since its establishment in 2006, it has collected blood, urine, and saliva samples from over 500,000 participants across the UK, as well as complete demographic, socio-economic, lifestyle, and health information from participants, which can be used to reliably assess the etiological relationship between exposure and cancer outcomes. All of the participants have provided informed written consent, and the UKB study was granted ethics approval from the North West Multicenter Research Ethics Committee.[13] Figure S1, Supplemental Digital Content, http://links.lww.com/MD/M249 shows the patient selection criteria and study flow chart.

All of the participants were followed from the time of recruitment until the diagnosis of primary invasive LC, their death, or their loss to follow-up, whichever occurred first, which was defined as the assignment of any of the C34 codes in the 10th Revision of the International Classification of Diseases. These codes are allocated to 6 subtypes of LC, characterized in terms of their location: C34.0 (main bronchus), C34.1 (upper lobe, bronchus, or lung), C34.2 (middle lobe, bronchus, or lung), C34.3 (lower lobe, bronchus, or lung), C34.8 (overlapping lesion of bronchus and lung), and C34.9 (bronchus or lung, unspecified). LC cases were identified by querying hospitalization data and family doctors in the primary care system and by examining death registration data from UKB.

2.2. Data processing

We excluded participants with more than 50% missing data, and those diagnosed with LC or other cancers at baseline, resulting in a final sample of 467,888 participants. The ratio of LC-to-non-LC cases in our sample was approximately 1:109, indicating an imbalance. Therefore, we adopted the synthetic minority over-sampling technique to process the raw data. This technique uses the k-nearest neighbor approach to generate new instances.[14] The sampled data were divided into a training set (70%) and a test set (30%), and the training set was used to create an ML model, while the test set was used to evaluate the model.

2.3. Feature selection

Predictor variables are one of the critical criteria to determine the validity of risk prediction models. We selected possible predictors of LC based on data availability and results from previous studies, including demographic characteristics (n = 3, e.g., gender, age, education), lifestyle and health information (n = 26, e.g., smoking, sleeplessness, sleep duration, overall health, lung diseases excluding lung cancer), laboratory data (n = 11, e.g., c-reactive protein, forced expiratory volume in 1 second), work environment (n = 4, e.g., workplace full of chemical or other fumes), and anthropometry (n = 5, e.g., body mass index, waist circumference, hip circumference). Most of these predictors have been repeatedly shown to be strongly associated with LC, and several predictors are highly correlated, such as smoking, occupational exposure, and lung disease,[5–7,9,10,12,15–19] and a few have not been extensively studied. Then, the predictors are ranked in importance by eXtreme Gradient Boosting (XGBoost).

2.4. Model development

In order to obtain better prediction performance, we used 4 commonly used ML methods—logistic regression (LR), naïve Bayes, random forest (RF), and XGBoost methods—to build LC risk-prediction models, all of which included all of the predictors. LR models are the models of choice in many medical data classification tasks, as they have the major advantage of retaining many features of linear regression in their binary outcome analyses, so they search for risk factors in the simplest way possible.[20] The naïve Bayes algorithm assumes attribute independence within the dataset, rendering it a straightforward and robust classification technique widely employed in practice.[21] The outstanding features of RF models are high prediction accuracy, less over-fitting, and strong anti-noise ability as an RF is a collection of many decision trees, so its prediction is made by aggregating the predictions of all trees by "majority voting."[22] XGBoost is an algorithm based on gradient boosting decision tree that utilizes regularized learning objective and second-order Taylor formula to simplify the model and avoid over-fitting problems, which is particularly important on multi-parameter datasets.[23] ML models tend to have better predictive accuracy than linear models, but they lack the interpretability of linear models. SHapley Additive exPlanations (SHAP) has better interpretability than other methods.[24,25] Therefore, SHAP was used to interpret the output of our model and to measure the importance of features. It allows to rank of the importance of features in the final model and assesses the contribution of each feature to the occurrence of LC.

2.5. Statistical analysis

Continuous variables are represented as means with standard deviation (SDs; normal distribution), and categorical variables are expressed as percentages. The statistics of participants with LC and those without LC were compared using t-tests, Wilcoxon rank-sum tests, and chi-square tests. The quality of the model was evaluated by the degree of differentiation and calibration, and the calibration curve was discretized by continuous data. Brier loss, log loss, precision, recall, F1, and areas under their receiver operating characteristic curves (AUCs) were calculated and used to compare the performances of models, with the AUC used as an overall measure of discrimination. Finally, we employed SHAP to visualize how

much each feature influenced the target outcome. R 4.1.2, SPSS.26, and Python 3.9.13 software were used for statistical analyses.

3. Results

3.1. Baseline characteristics

Our sample for analysis included 467,888 participants, 4299 of whom ultimately developed LC. Table 1 lists the characteristics of all participants. At baseline, the average age of the participants was 56 years, and 259,414 (55.44%) were women. In terms of body mass index (BMI), hip circumference, and work environment, the participants with LC performed similarly to whose without LC. Also there is no significant gender difference. However, compared to other participants, those who developed LC were older, in poorer health status, and more likely to be smokers.

3.2. Model accuracy

The model is evaluated using test sets. As mentioned, AUCs, Brier loss, log loss, recall, precision, and F1 score were used to

Table 1

The baseline characteristics of UK Biobank participants included in the study by lung cancer.

Characteristics	Total	Non-lung cancer	Lung cancer
Age, y, mean (SD)	56.15 ± 8.07	56.10 ± 8.07	61.73 ± 5.83
Sex, n (%)			
Female	259,414 (55.44)	257,333 (55.51)	2081(48.41)
Male	208,474 (44.56)	206,256 (44.49)	2218(51.59)
BMI, kg/m^2, mean (SD)	27.36 ± 4.73	27.36 ± 4.73	27.42 ± 4.75
Waist circumference, cm, mean (SD)	89.94 ± 13.28	89.91 ± 13.28	93.04 ± 13.48
Hip circumference, cm, mean (SD)	103.31 ± 9.12	103.32 ± 9.12	102.81 ± 9.27
Neuroticism score, mean (SD)	4.12 ± 2.92	4.12 ± 2.92	4.24 ± 2.93
Forced expiratory volume in 1 second, L, mean (SD)	2.87 ± 0.65	2.87 ± 0.65	2.52 ± 0.64
Forced vital capacity, L, mean (SD)	3.79 ± 0.83	3.79 ± 0.83	3.51 ± 0.78
Met, min/week, mean (SD)	2660.22 ± 2428.24	2660.60 ± 2427.86	2618.53 ± 2468.20
Sleep duration, n (%)			
>8 h/day	34,056 (7.28)	33,563 (7.24)	493 (11.47)
7–8 h/day	319,613 (68.31)	317,010 (68.38)	2603(60.55)
<7 h/day	114,219 (24.41)	113,016 (24.38)	1203(27.98)
Education, n%			
O levels/GCSEs/CSEs or equivalent	124,415 (26.59)	123,472 (26.63)	943 (21.94)
A levels/AS levels or equivalent	52,281 (11.17)	51,972 (11.21)	309 (7.19)
College or University degree	152,985 (32.70)	152,319 (32.86)	666 (15.49)
Other	138,207 (29.54)	135,826 (29.30)	2381(55.38)
Over health, n%			
Excellent	78,644 (16.81)	78,307 (16.89)	337 (7.84)
Good	272,968 (58.34)	270,933 (58.44)	2035(47.34)
Fair	98,069 (20.96)	96,618 (20.84)	1451(33.75)
Poor	18,207 (3.89)	17,731 (3.82)	476 (11.07)
Smoking, n (%)			
Never	192,615 (41.17)	192,165 (41.45)	450 (10.47)
Previous	240,681 (51.44)	238,423 (51.43)	2258(52.52)
Current	34,592 (7.39)	33,001 (7.12)	1591 (37.01)
Drinking, n (%)			
Never	22.075 (4.72)	21,909 (4.73)	166 (3.86)
Previous	15,855 (3.39)	15,544 (3.35)	311 (7.23)
Current	429,958 (91.89)	426,136 (91.92)	3822(88.90)
Workplace very dusty, n (%)			
Rarely/never	448,633 (95.88)	444,413 (95.86)	4220(98.16)
Sometimes	15,225 (3.25)	15,160 (3.27)	65 (1.51)
Often	4030(0.86)	4016(0.87)	14 (0.33)
Chemical other fumes, n (%)			
Rarely/never	438,678 (93.72)	434,497 (93.72)	4181(97.26)
Sometimes	22,700 (4.85)	22,609 (4.88)	91 (2.12)
Often	6510(1.39)	6483(1.40)	27 (0.63)
Diesel exhaust, n (%)			
Rarely/never	458,319 (97.95)	454,075 (97.95)	4244(98.72)
Sometimes	7692(1.64)	7655(1.65)	37 (0.86)
Often	1877(0.40)	1859(0.40)	18 (0.42)
Nap during day, n (%)			
Rarely/never	267,962 (57.07)	265,980 (57.37)	1982(46.10)
Sometimes	176,638 (37.75)	174,722 (37.69)	1916(44.57)
Often	23,288 (4.98)	22,887 (4.94)	101 (9.33)
Sleeplessness, n (%)			
Rarely/never	117,614 (24.50)	113,736 (24.53)	878 (20.42)
Sometimes	223,231 (47.71)	221,257 (47.73)	1974(45.92)
Often	130,043 (27.79)	128,596(27.74)	1447(33.66)
Snoring, n (%)			
Rarely/never	194,523 (41.57)	192,619 (41.55)	1904(44.29)
Sometimes	273,365 (58.43)	270,970 (58.45)	2395(55.71)

* Values are mean (standard deviation) or number (percentages) unless otherwise indicated. SD = standard deviation.

evaluate the performance of the model (as shown in Table 2, Fig. 1). The AUCs value calculated from the ROC curve is one of the most important indicators of the classification quality of a binary classification model, and the ROC curve is the performance of the classifier under different thresholds. The F1 score is the reconciled average of precision and recall, which can help us to consider the precision and recall of the model in a comprehensive way. The Brier loss and the log loss are used to measure the difference between the predicted probability of the algorithm and the true result. The closer the score is to 0, the more accurate the predicted result is. Among them, the XGBoost model outperforms other machine learning models with the largest area under the ROC curve (AUC = 0.998), Brier loss (0.009), log loss (0.042), precision (0.997), recall (0.981), and F1 score (0.989). Therefore, we chose the XGBoost model to analyze the importance of features. We also used calibration curves to evaluate model performance, which is not common in similar studies in the past. Although a diagonal calibration curve would be preferable, the calibration capability of our model is still satisfactory (as shown in Fig. 2).

3.3. Key variables

The XGBoost Feature Importance Ranking Histogram shows that non-laboratory variables have a more significant influence, with hip circumference, neuroticism score, waist circumference, number of current cigarettes smoked, and Metabolic Equivalent Task identified as the top 5 crucial variables, and most of them have not received a lot of attention (Fig. 3).

3.4. Model visualization

We used SHAP to interpret the output of the model. First, we average the hash values of each feature to draw a standard bar graph to obtain the global importance (Fig. S2, Supplemental Digital Content, http://links.lww.com/MD/M250). Then, the relationship between the magnitude of the feature values and the predicted outcome is determined by scatterplotting the SHAP values of each feature for every sample (Fig. 4). It can be seen that the range of hip circumference in both charts was the widest, indicating that it had significant predictive power. Waist circumference, number of cigarettes previously smoked daily, forced expiratory volume in 1 second (FEV1), and age also had relatively high predictive power. The horizontal axis is the SHAP value, where the magnitude of the value is proportional to the impact of the corresponding feature on the prediction results of the sample. A SHAP value >0 indicates a positive impact. In this study, a positive impact represented a high risk of lung cancer. The color indicates the high and low characteristic values, with red corresponding to high values and blue corresponding to low values. Thus, the specific effect of each predictor is indicated by its color and direction. For example, with respect to the number of cigarettes previously smoked daily, the higher the number, the greater the risk of LC. Similarly, older age, a high number of cigarettes previously smoked daily, a high blood concentration of C-reactive protein (CRP), and a high blood concentration of high-density lipoprotein cholesterol had a positive impact on the occurrence of LC. SHAP dependence plots reflect the influence of the trend of each feature on the prediction result. When other variables were held constant, the risk of LC was higher with age > 55 years, FEV1 < 2 L, FVC > 5 L, and waist circumference > 80 cm. Once smoking can induce LC, especially when smoking 20 to 40 cigarettes per day peak; on the contrary, regardless of neurotic score, are negatively correlated with LC (Fig. 5).

4. Discussion

A rational predictive model can be an adjunct to LDCT lung cancer screening, so our study aimed to develop a generalizable LC predictive model based on easily accessible predictors using ML algorithms. In addition, new predictors that have received little attention so far, such as neuroticism and waist circumference, were identified.

The original Bach model and LLP model were relatively simple; the extended Bach and LLP models and the Spitz model enhance the prediction of lung cancer by incorporating CT imaging and laboratory indicators.[9,10] In recent years, Gould et al's model based on routine laboratory data identified the potential value of normal blood counts as a risk factor for LC and found that the model was more accurate than the USPSTF Eligibility Criteria for Lung Cancer Screening and mPLCOm2012 for the early identification of LC.[12] The success of these models suggests that adding more meaningful predictor variables may increase the efficacy of the models. However, none of these studies have focused on abnormal signals in the early stages of the organism. The association between factors such as neuroticism, waist circumference, and LC has been described,[26–29] but existing models have not yet used them as predictors. In past studies, BMI was often used to predict lung cancer. Olson[30] and Nitsche[31] et al pointed out that abdominal obesity had a more significant effect on the development of LC than overall obesity, and that BMI assessed comprehensive body mass index and waist circumference was used to evaluate abdominal obesity, and to recognize further the link between obesity and LC, waist circumference was included as a predictive variable in the present study. Stereotypes suggest that LC patients have insomnia, anxiety, depression, and other adverse psychological effects

Table 2

The performance of various machine learning models in training sets (A) and testing sets (B).

A					
Classifier	**Brier loss**	**Log loss**	**Precision**	**Recall**	**F1**
Logistic regression	0.164452	0.504315	0.766332	0.751650	0.758920
Naive Bayes	0.255612	1.384820	0.631035	0.958931	0.761172
XGBoost	0.007229	0.034846	0.997804	0.987607	0.992680
Random forest	0.001472	0.018227	0.999992	1.000000	0.999996
B					
Classifier	**Brier loss**	**Log loss**	**Precision**	**Recall**	**F1**
Logistic regression	0.160016	0.493003	0.772840	0.775874	0.774354
Naive Bayes	0.251972	1.306871	0.633452	0.964391	0.764650
XGBoost	0.009395	0.041846	0.997173	0.981290	0.989167
Random forest	0.111088	0.366342	0.985262	0.686312	0.809054

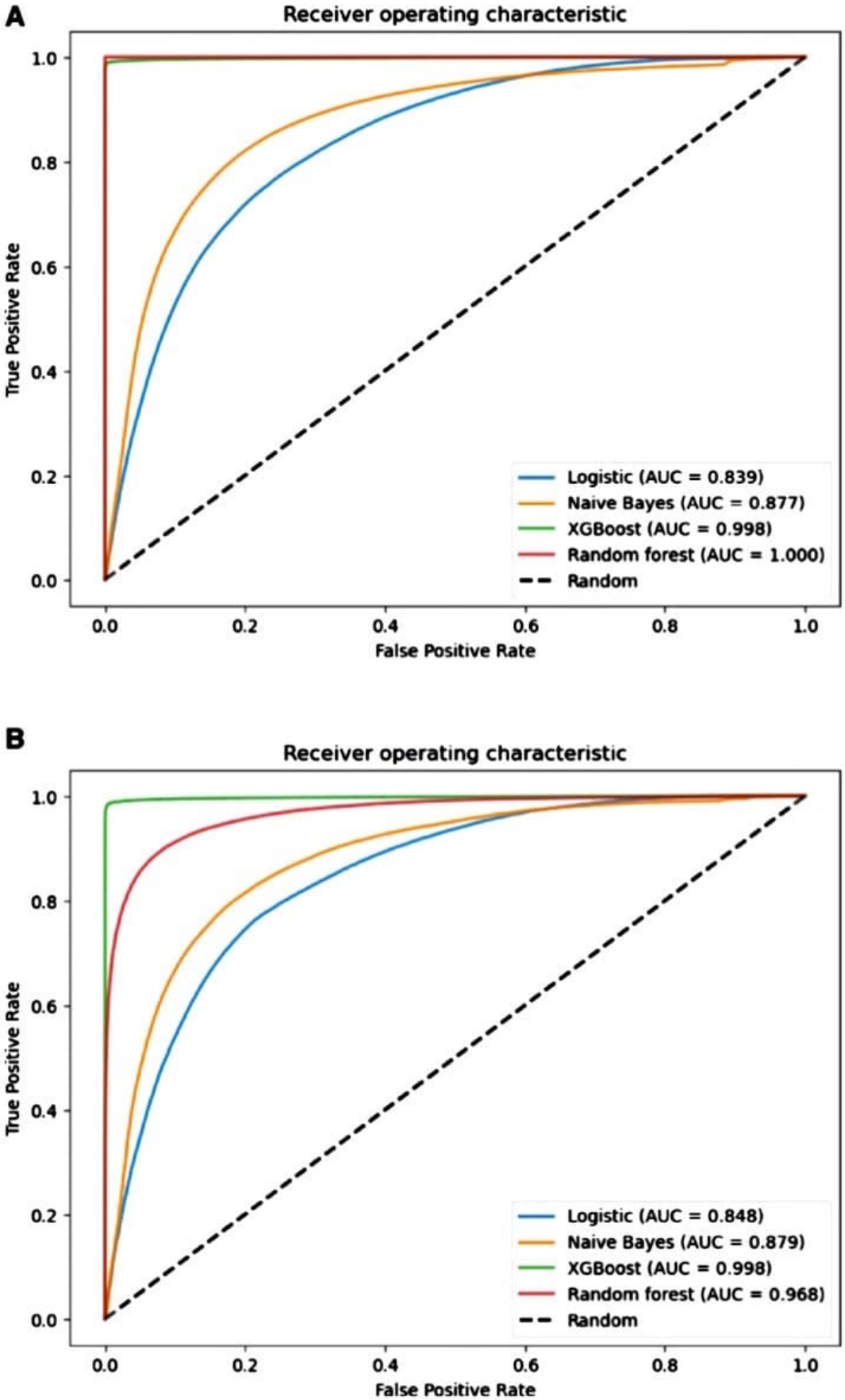
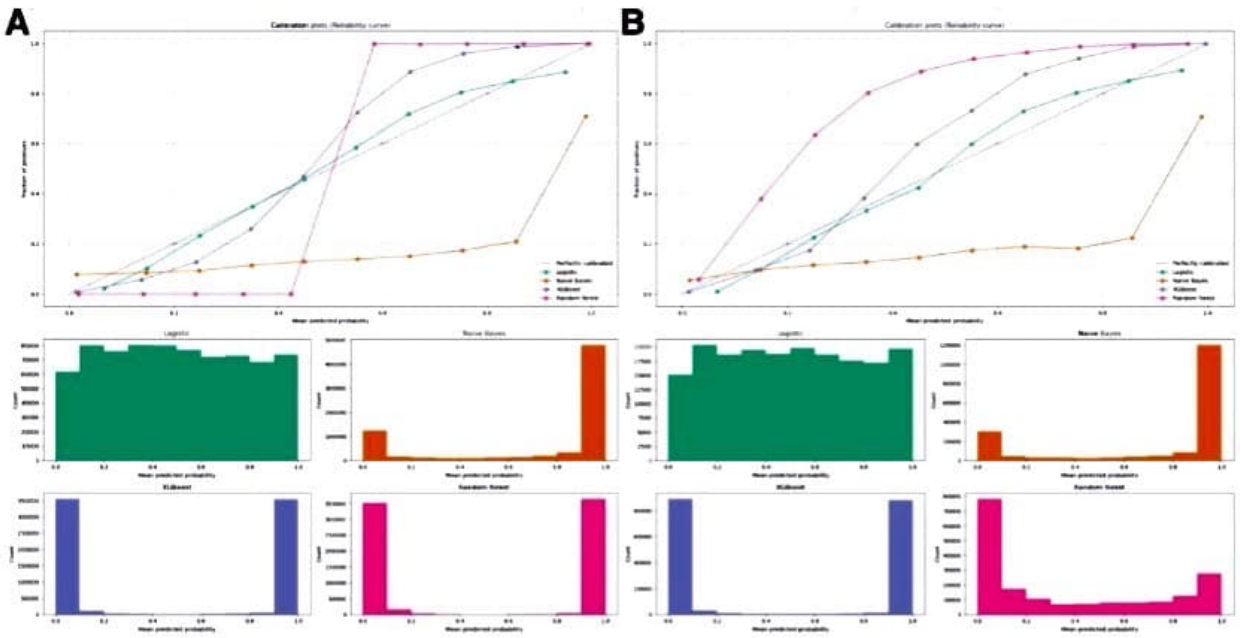


Figure 1. The area under the receiver operator characteristic (ROC) curves for 4 machine learning algorithms were generated for both the training (A) and testing (B) sets.

due to fear of the disease while ignoring whether emotions may also have an impact on lung cancer. Although laboratory data has a higher degree of accuracy, abnormalities in the body can often be monitored by ourselves, so we focus on these self-measurable indicators.

Based on the UKB database, 4 machine learning models, LR, naïve Bayes, RF, and XGBoost, were built to predict the risk of LC development. The XGBoost model showed the best performance by evaluating the metrics, including AUC (0.998), Brier loss (0.009), Log loss (0.042), precision (0.997), recall (0.981),



*The solid line indicates the actual prediction effect, and the diagonal dashed line indicates the ideal effect, and the closer the two, the better the prediction effect.

Figure 2. Calibration curves for different machine learning models for both the training (A) and testing (B) sets. *The solid line indicates the actual prediction effect, and the diagonal dashed line indicates the ideal effect, and the closer the 2, the better the prediction effect.

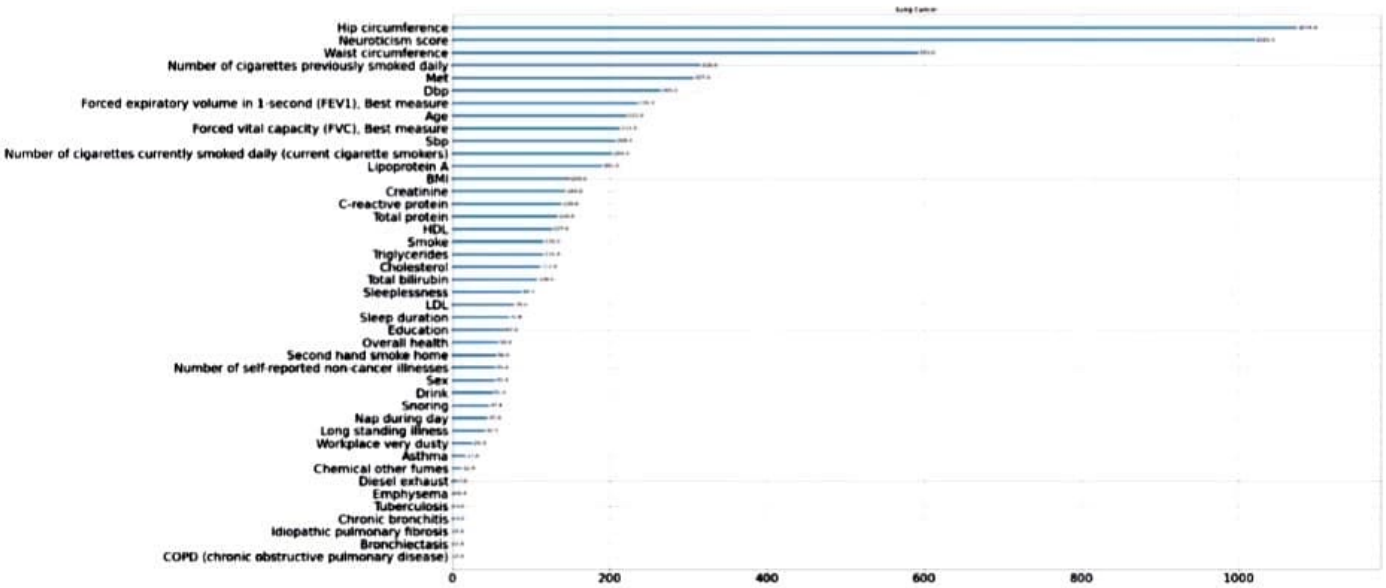


Figure 3. Feature importance ranking using extreme gradient boosting. BMI = body mass index; Dbp = diastolic blood pressure; HDL = high-density lipoprotein; LDL = low-density lipoprotein; Met = metabolic equivalent task; Sbp = systolic blood pressure.

and F1 score (0.989) showed the best performance overall. Although the efficacy cannot be directly compared with that of other models, it cannot be denied that our model possesses excellent discrimination and calibration. The higher accuracy of the model may be related to the additional contribution of the new predictors. The model algorithm is also a pivotal point in improving its accuracy. Traditional statistical analysis methods such as linear or logistic regression usually consider the covariance between the variables, as they rely on a linear relationship between them. In these methods, the coefficient estimates of the model may become unstable and rugged to interpret if there is a high degree of covariance characterizing the model. However, tree-based ML algorithms such as XGBoost are usually less sensitive to covariance. This is because tree models do not rely on linear relationships but instead rely on nonlinear relationships of features to model them. Tree models can split features of their choosing and handle highly correlated features. As a result, covariance is usually not something to worry too much about when using algorithms such as XGBoost.

Neuroticism score is the most critical factor in our model. Neuroticism reflects an individual's emotional processes, and individuals with high neuroticism scores are often in negative moods such as anxiety and depression. High levels of neuroticism are also a significant cause of decreased sleep quality.[32] It has been found that unhealthy sleep duration (<7 h or >8 h/d) may increase the incidence of LC by affecting the body's circadian pattern, immune-inflammatory balance, and other aspects.[33] Interestingly, high and low neuroticism scores are negatively associated with the occurrence of LC. We hypothesize that this is because individuals with these psychiatric symptoms

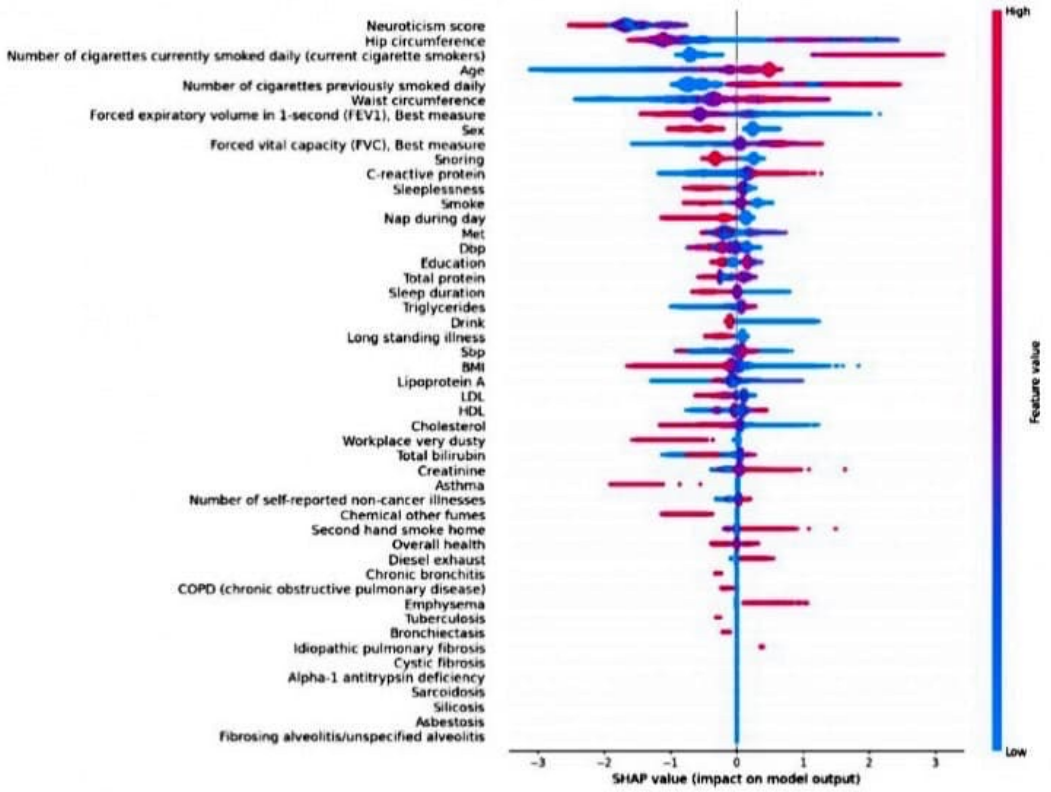


Figure 4. SHapley Additive exPlanations visualization plot of selected predictors. *The width of the horizontal bar range can be interpreted as the effect on the model's predictions, with the larger the range, the greater the effect. The color of the horizontal bar indicates the magnitude of the predictor, showing a gradient from blue (low) to red (high). the direction on the x-axis indicates the likelihood of having LC (right) or being healthy (left). The reader can infer the likelihood of having LC. BMI = body mass index; Dbp = diastolic blood pressure; HDL = high-density lipoprotein; LDL = low-density lipoprotein; Met = metabolic equivalent task; Sbp = systolic blood pressure; XGBoost = extreme gradient boosting.

may be more aware of their health status compared to other individuals, and therefore will seek medical attention when a precursor of LC occurs, leading to effective prevention. Another noteworthy point is that BMI, hip circumference, and waist circumference showed different results despite all being indicators that can reflect obesity. BMI was negatively associated with LC risk in our study, but this association may be due to the inverse effect of disease progression on appetite and interference from smoking behavior.[34] In contrast, waist circumference was positively associated with LC incidence,[28,29] and we also found that although hip circumference showed more influence than waist circumference, the results showed a more evident association between waist circumference and LC. A study found that centrally obese people with a BMI below 25 kg/m^2 but with a large waist circumference could have a 40% higher risk of LC than those with a higher BMI but average waist circumference,[29] Hidayat et al[35] found that among never-smokers a larger waist circumference could increase the risk of LC by 11%. Furthermore, this study also revealed a positive correlation between waist circumference and the likelihood of LC. It suggests that it may be the type of obesity influencing LC, reminding us that a well-proportioned body is more important than a lower BMI. While past studies have shown that model accuracy can be improved by incorporating expensive/invasive predictor variables, our results suggest that simple predictors can also improve model predictive ability and have a greater range of applications, especially in low economic areas and groups, reducing the burden of expenditure. In addition, mood, sleep, and obesity can be addressed early and can be intervened in, which can help physicians propose individualized preventive measures.

In addition to the new predictors described above, many familiar factors contributed significantly to our model. Cigarette smoking is the most common risk factor for LC,[36] and we have also determined that current smoking behavior is more harmful than previous smoking behavior, which partly reflects the fact that the risk of LC occurrence increases with increasing tobacco dose.[37] However, we observed that LC risk did not increase when smoking more than 40 cigarettes per day, possibly because some long-term smokers have a robust self-repair system to avoid carcinogenic gene mutations.[38] Even healthy older adults inevitably experience age-related declines in physiologic reserve and increased tissue and cellular susceptibility to carcinogens.[19] We observed that subjects with elevated blood levels of CRP were more susceptible to LC than subjects with normal blood levels of CRP, and Lyu Z et al[39] demonstrated that the inclusion of metabolic markers such as CRP and low-density lipoprotein in the prediction model improved the discriminatory properties of the model. Reduced FEV1 usually indicates airway obstruction or impaired lung function. Mendelian analyses have shown that reduced FEV1 increases the Mendelian analysis suggests that reduced FEV1 increases the risk of lung squamous cell carcinoma. LC risk prediction models incorporating lung function have found a strong negative correlation between maximal FEV1 and LC irrespective of smoking status, which could improve screening sensitivity.[16] This is similar to the conclusion we reached in the present study. Pulmonary comorbidities, such as chronic obstructive pulmonary disease and emphysema, have also been associated with LC,[17,18] but this was not seen in the present study. This may be due to our sample's characteristics; therefore, the importance of these comorbidities cannot be overlooked.

The strengths of this study are that our model is based on a large prospective cohort, covers a large number of participants and follow-up information, has a high evidence-based rating, and avoids selection bias and retrospective bias. Second, the proposed model utilizes predictors that can be quickly obtained through questionnaires, physical measurements, and simple clinical examinations, and the low-cost approach can significantly increase the population's participation rate and avoid

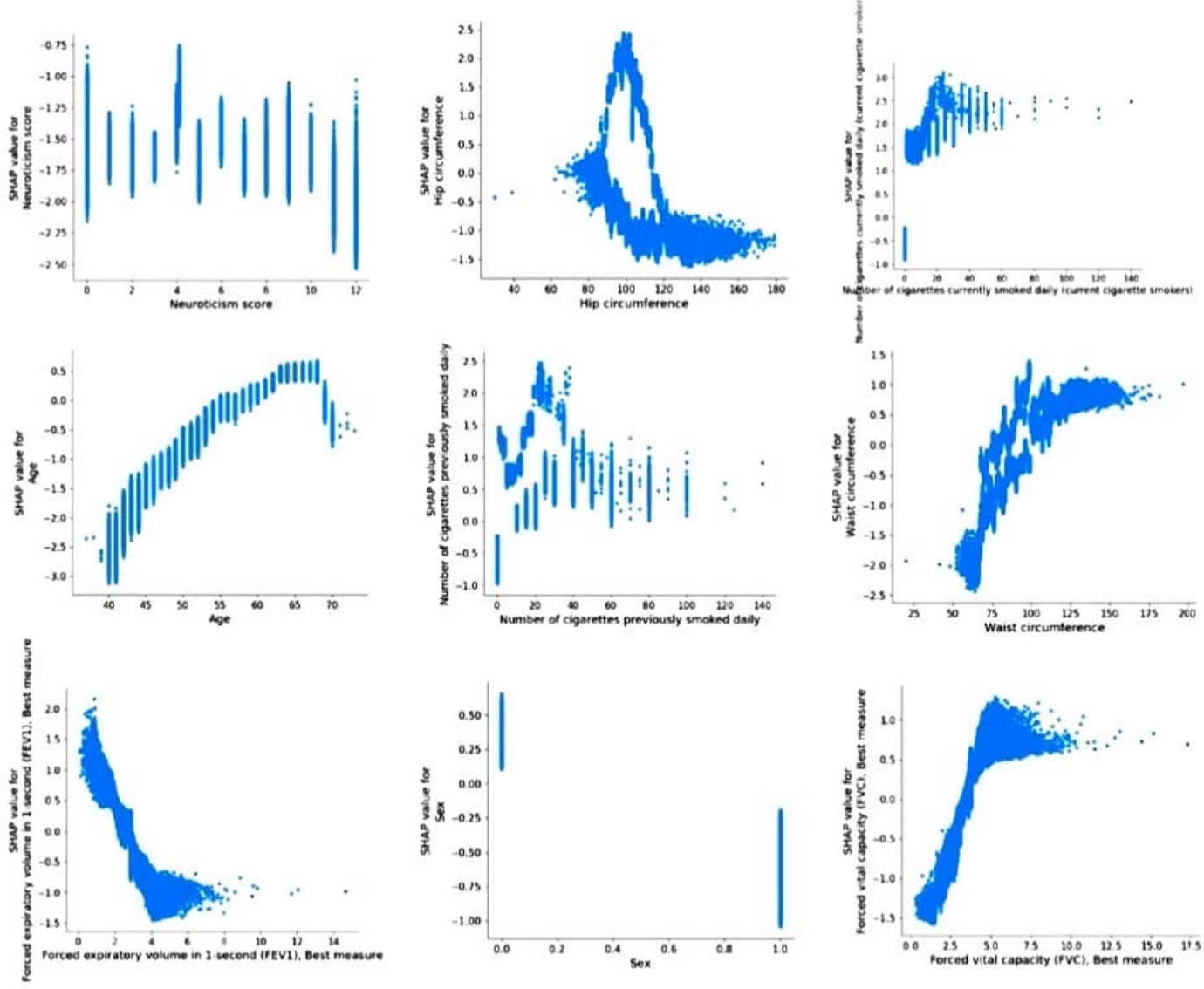


Figure 5. SHAP dependence plots for the main predictor variables.

unnecessary waste of resources. In addition, several new predictors, such as neuroticism, waist circumference, and hip circumference, played an influential role in this study, and these non-laboratory variables can help provide a reasonable risk prediction when laboratory results are unavailable, so it is worthwhile to pay more attention to their potential as predictors to improve model performance.

However, this study has some limitations. First, due to the health effect of volunteers,[40] people who volunteered to participate in UKB were younger, healthier, and better educated than the general population. Thus, our data are not representative of the current general population in many ways, which accounts for the low prevalence of LC. Second, non-laboratory variables such as neuroticism score and sleeplessness are subjective, so future studies should minimize reporting bias regarding these variables, where possible. Finally, the data used in our model came from the UK population; thus, their applicability to other ethnic groups needs further testing, as the model performance may change when applied to different samples.

5. Conclusion

In summary, we developed an effective LC prediction model based on machine learning algorithms, and the easily accessible predictors highlight the potential for our model to be used as a tool for rapid assessment.

Acknowledgments

Data supporting the results of this study are available on request from the UK Biobank. We would like to thank all the participants and researchers at the UK Biobank.

Author contributions

Conceptualization: Siqi Zhang, Liyuan Han.
Data curation: Liyuan Han.
Formal analysis: Weiwen Xu, Yue Wang.
Funding acquisition: Guofang Zhao, Ting Cai.
Methodology: Siqi Zhang, Ting Cai.
Resources: Ting Cai.
Software: Yue Wang.
Validation: Guofang Zhao.
Visualization: Liangwei Yang, Liyuan Han.
Writing – original draft: Siqi Zhang.
Writing – review & editing: Siqi Zhang, Liangwei Yang, Weiwen Xu, Yue Wang, Liyuan Han, Guofang Zhao, Ting Cai.

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.

[2] Detterbeck FC, Boffa DJ, Kim AW, et al. The eighth edition lung cancer stage classification. Chest. 2017;151:193–203.

[3] Oudkerk M, Liu S, Heuvelmans MA, et al. Lung cancer LDCT screening and mortality reduction – evidence, pitfalls and future perspectives. Nat Rev Clin Oncol. 2021;18:135–51.
[4] Bonney A, Malouf R, Marchal C, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev. 2022;8:CD013829.
[5] Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003;95:470–8.
[6] Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer. 2008;98:270–6.
[7] Tammemagi CM, Pinsky PF, Caporaso NE, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. J Natl Cancer Inst. 2011;103:1058–68.
[8] Sattar M, Majid A, Kausar N, et al. Lung cancer prediction using multigene genetic programming by selecting automatic features from amino acid sequences. Comput Biol Chem. 2022;98:107638.
[9] Gray EP, Teare MD, Stevens J, et al. Risk prediction models for lung cancer: a systematic review. Clin Lung Cancer. 2016;17:95–106.
[10] Maisonneuve P, Bagnardi V, Bellomi M, et al. Lung cancer risk prediction to select smokers for screening CT--a model based on the Italian COSMOS trial. Cancer Prev Res (Phila). 2011;4:1778–89.
[11] Fatima FS, Jaiswal A, Sachdeva N. Lung cancer detection using machine learning techniques. Crit Rev Biomed Eng. 2022;50:45–58.
[12] Gould MK, Huang BZ, Tammemagi MC, et al. Machine learning for early lung cancer identification using routine clinical and laboratory data. Am J Respir Crit Care Med. 2021;204:445–53.
[13] Conroy MC, Lacey B, Bešević J, et al. UK Biobank: a globally important resource for cancer research. Br J Cancer. 2023;128:519–27.
[14] Swana EF, Doorsamy W, Bokoro P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors (Basel). 2022;22:3246.
[15] Štěpánek L, Ševčíková J, Horáková D, et al. Public health burden of secondhand smoking: case reports of lung cancer and a literature review. Int J Environ Res Public Health. 2022;19:13152.
[16] Muller DC, Johansson M, Brennan P. Lung cancer risk prediction model incorporating lung function: development and validation in the UK biobank prospective cohort study. J Clin Oncol. 2017;35:861–9.
[17] Qi C, Sun SW, Xiong XZ. From COPD to lung cancer: mechanisms linking, diagnosis, treatment, and prognosis. Int J Chron Obstruct Pulmon Dis. 2022;17:2603–21.
[18] Mouronte-Roibás C, Leiro-Fernández V, Fernández-Villar A, et al. COPD, emphysema and the onset of lung cancer. A systematic review. Cancer Lett. 2016;382:240–4.
[19] Barta JA, Zinner RG, Unger M. Lung cancer in the older patient. Clin Geriatr Med. 2017;33:563–77.
[20] LaValley MP. Logistic regression. Circulation. 2008;117:2395–9.
[21] Zhang Z. Naïve Bayes classification in R. Ann Transl Med. 2016;4:241.
[22] Breiman L. Random forests. Mach Learn. 2001;45:5–32.
[23] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
[24] Lundberg SM, Lee S-I, editors. A Unified Approach to Interpreting Model Predictions. Neural Information Processing Systems; 2017.
[25] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.
[26] Wei X, Jiang X, Zhang X, et al. Association between neuroticism and risk of lung cancer: results from observational and Mendelian randomization analyses. Front Oncol. 2022;12:836159.
[27] Nakaya N, Bidstrup PE, Saito-Nakaya K, et al. Personality traits and cancer risk and survival based on Finnish and Swedish registry data. Am J Epidemiol. 2010;172:377–85.
[28] Dewi NU, Boshuizen HC, Johansson M, et al. Anthropometry and the risk of lung cancer in EPIC. Am J Epidemiol. 2016;184:129–39.
[29] Kabat GC, Kim M, Hunt JR, et al. Body mass index and waist circumference in relation to lung cancer risk in the Women's Health Initiative. Am J Epidemiol. 2008;168:158–69.
[30] Olson JE, Yang P, Schmitz K, et al. Differential association of body mass index and fat distribution with three major histologic types of lung cancer: evidence from a cohort of older women. Am J Epidemiol. 2002;156:606–15.
[31] Nitsche L, Vedire Y, Kannisto E, et al. Visceral obesity in non-small cell lung cancer. Cancers (Basel). 2022;14:3450.
[32] Stephan Y, Sutin AR, Bayard S, et al. Personality and sleep quality: evidence from four prospective studies. Health Psychol. 2018;37:271–81.
[33] Xie J, Zhu M, Ji M, et al. Relationships between sleep traits and lung cancer risk: a prospective cohort study in UK Biobank. Sleep. 2021;44:zsab089.
[34] Zhou W, Liu G, Hung RJ, et al. Causal relationships between body mass index, smoking and lung cancer: univariable and multivariable Mendelian randomization. Int J Cancer. 2021;148:1077–86.
[35] Hidayat K, Du X, Chen G, et al. Abdominal obesity and lung cancer risk: systematic review and meta-analysis of prospective studies. Nutrients. 2016;8:810.
[36] Lemjabbar-Alaoui H, Hassan OU, Yang YW, et al. Lung cancer: biology and treatment options. Biochim Biophys Acta. 2015;1856:189–210.
[37] Warren GW, Cummings KM. Tobacco and lung cancer: risks, trends, and outcomes in patients with cancer. Am Soc Clin Oncol Educ Book. 2013;33:359–64.
[38] Huang Z, Sun S, Lee M, et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. Nat Genet. 2022;54:492–8.
[39] Lyu Z, Li N, Chen S, et al. Risk prediction model for lung cancer incorporating metabolic markers: development and internal validation in a Chinese population. Cancer Med. 2020;9:3983–94.
[40] Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol. 2017;186:1026–34.