scientific reports

Check for updates

OPEN

Exploring non-invasive biomarkers for pulmonary nodule detection based on salivary microbiomics and machine learning algorithms

Chunxia Huang[1,3], Qiong Ma[1,3], Xiao Zeng[1,3], Jiawei He[1], Fengming You[1,2]✉, Xi Fu[1,2]✉ & Yifeng Ren[1,2]✉

Microorganisms are one of the most promising biomarkers for cancer, and the relationship between microorganisms and lung cancer occurrence and development provides significant potential for pulmonary nodule (PN) diagnosis from a microbiological perspective. This study aimed to analyze the salivary microbiota features of patients with PN and assess the potential of the salivary microbiota as a non-invasive PN biomarker. We collected saliva smples from 153 patients with PN and 40 controls. Using 16 S rRNA gene sequencing, differences in α- and β-diversity and community composition between the group with PN and controls were analyzed. Subsequently, specific microbial variables were selected using six models were trained on the selected salivary microbial features. The models were evaluated using metrics, such as the area under the receiver operating characteristic curve (AUC), to identify the best-performing model. Furthermore, the Bayesian optimization algorithm was used to optimize this best-performing model. Finally, the SHapley Additive exPlanations (SHAP) interpretability framework was used to interpret the output of the optimal model and identify potential PN biomarkers. Significant differences in α- and β-diversity were observed between the group with PN and controls. Although the predominant genera were consistent between the groups, significant disparities were observed in their relative abundances. By leveraging the random forest algorithm, ten characteristic microbial variables were identified and incorporated into six models, which effectively facilitated PN diagnosis. The XGBoost model demonstrated the best performance. Further optimization of the XGBoost model resulted in a Bayesian Optimization-based XGBoost (BOXGB) model. Based on the BOXGB model, an online saliva microbiota-based PN prediction platform was developed. Lastly, SHAP analysis suggested *Defluviitaleaceae_UCG*-011, *Aggregatibacter*, *Oribacterium*, *Bacillus*, and *Prevotalla* are promising non-invasive PN biomarkers. This study proved salivary microbiota as a non-invasive PN biomarker, expanding the clinical diagnostic approaches for PN.

Keywords Salivary microbiota, Pulmonary nodule, Machine learning, SHapley additive additive explanations (SHAP), Non-invasive biomarker

Lung cancer remains the leading cause of incidence and mortality among malignant tumors in China[1], causing concerns regarding early screening. Although computed tomography (CT) scanning is the primary screening tool and significantly enhances early lung cancer detection rates, it has concurrently led to a consistent increase in detecting pulmonary nodule (PN). Considering the potential risk of PN for developing into lung cancer, multiple guidelines recommend CT surveillance as the primary monitoring method. Nevertheless, CT has limitations, including radiation exposure, elevated false-positive rates[2], and restricted applicability to some demographic groups, exacerbating the physical and psychological burden of prolonged surveillance, which has become a societal concern. Consequently, relying solely on imaging modalities for PN diagnosis is insufficient, underscoring the pressing need to expand PN diagnostic methodologies.

[1]Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu 610072, Sichuan Province, China. [2]TCM Regulating Metabolic Diseases Key Laboratory of Sichuan Province, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu 610072, Sichuan Province, China. [3]Chunxia Huang and Qiong Ma contributed equally to this work. ✉email: yfmdoc@163.com; fuxi884853@163.com; ryftcm.dr@yahoo.com

Microorganisms, regarded as the second human genome, play crucial roles in health and disease. This topic, which has been identified as one of the 125 most cutting-edge scientific questions globally[3], has been extensively explored. Findings from the Human Microbiome Project indicate that the oral cavity is a crucial site of microbial aggregation in the human body, with saliva and dental plaque microbiota diversity ranking second to that of the gut[4]. Moreover, recent studies have confirmed a close association exists between the salivary microbiota and various malignant tumors[5,6]. Significant differences exist in salivary microbiota profiles before and after treatment in patients with oral cancer[7]. Studies have analyzed the relationship between changes in salivary microbiota and symptoms in patients with head and neck squamous cell carcinoma[8]. Additionally, an increase in *Veillonella* and *Streptococcus* exist in the salivary microbiota of patients with non-small-cell lung cancer, whereas *Porphyromonas gingivalis* in the gum microbiota may promote malignant lung cancer progression[9]. Recent studies have further proposed saliva as the preferred biological specimen for disease diagnosis[10]. Therefore, the following questions arise: can the salivary microbiota serve as a potential PN biomarker, and how can they be applied in PN diagnosis?

With the advent of artificial intelligence, machine learning (ML) has been widely applied in medical data analysis[11], providing the scientific and technological means to answer these questions. ML allows the identification of potentially important predictor variables by permitting interaction among variables and finding optimization algorithms between effective outcomes and predictor variables, thereby establishing models with performance surpassing traditional methods for predicting and monitoring disease quickly and accurately[12]. Moreover, some studies have integrated ML with microbiome data to establish risk prediction models for diseases, such as malaria, lung and liver cancer[13–15], indicating that combining ML and microbiome data is an emerging direction for predictive modeling. However, no study exists on constructing a risk prediction model for PN based on microbiome data.

Accordingly, the objective of this study is to demonstrate the differential microbial profiles of patients with PN and controls. In addition, we imported a variety of machine learning algorithms to find the optimal model for predicting PN based on microbial variables. Finally, combining the SHapley Additive exPlanations (SHAP) algorithm to explore promising non-invasive biomarkers provides a new way to expand the clinical diagnosis of PN (Fig. 1).

Results

Clinical characteristics of the participants

A total of 271 subjects were recruited in this study. After 69 subjects were excluded by reference to the inclusion criteria, saliva microbiome sequencing was performed on the remaining 202 subjects. Moreover, 9 samples with unqualified 16 S rRNA detection, such as discrete point, low read microbiota, OD260/280 < 1.8, were excluded twice. Therefore, 193 samples were finally included in the analysis, comprising 153 and 40 individuals with PN and controls (Fig. 2). The groups were matched for clinical information such as age, sex, and smoking status. The baseline characteristics of all participants are presented in Table 1. The median age of the enrolled patients was 32 years, and the male-to-female ratio was approximately 3:7. There was no significant differences in the baseline characteristics.

Salivary microbiota map of patients with PN and controls

Through Venn diagram analysis (Fig. 3A), we found 267 shared genera between the groups, with 119 unique genera identified in the group with PN and 61 in the controls. Subsequently, α-diversity analysis was conducted, wherein we observed a trend toward a plateau in the rarefaction curve based on the Sobs index, indicating a progressively reasonable sequencing data volume (Fig. 3B). Using the Wilcoxon rank-sum test, the Chao, ACE, and Sobs indices were higher in the controls, suggesting a lower community abundance in the group with PN than in the controls. However, the Simpsoneven index was slightly lower in the controls, indicating a more even community distribution than the group with PN. These results collectively indicate significant differences in α-diversity between the groups (Fig. 3C–F, $P < 0.05$). Additionally, β-diversity inter-group differences were analyzed using the Bray-Curtis distance algorithm, showing significant differences in distance values between the groups ($P = 0.03$) (Fig. 3I). Furthermore, β-diversity analysis results were presented in reduced dimensions using PCoA and NMDS based on the weighted_unifrac distance algorithm (Fig. 3G, H), revealing distinct individual clusters between the groups, suggesting significant differences in community structure between the groups (Adonis $R^2 = 0.01$, $P = 0.03$).

Community bar plots were constructed to illustrate the microbial differences between the groups (Fig. 4A). The results indicated that the predominant genera in the groups were consistent, with *Streptococcus*, *Rothia*, and *Neisseria* ranking as the top three genera; however, significant differences were observed in their relative abundance. The heatmap visually depicts the distribution of the dominant genera across different samples (Fig. 4B). The Circos plot provides a visual representation of the one-to-one correspondence between the intergroup species in the groups (Fig. 4C). Further analysis using LEfSe revealed specific taxonomic groups between the two groups at the genus level. Genera, such as *Streptococcus*, *Coprococcus*, and *Lautropia*, were associated with the controls, whereas *Prevotella*, *Actinomyces*, and *TM7x* were associated with the group with PN (Fig. 4D).

Correlation between salivary microbiota and clinical characteristics of PN

Furthermore, we assessed the correlation between the patients' clinical characteristics and salivary microbiota composition using Spearman's correlation analysis. Before conducting the correlation analysis, we screened the environmental factors affecting the microbial community using VIF (VIF < 5). The VIF values of the clinical environmental factors remained unchanged before and after identification, indicating their suitability for subsequent analyses (Table S1). Overall, age, sex, and family tumor history were positively or negatively

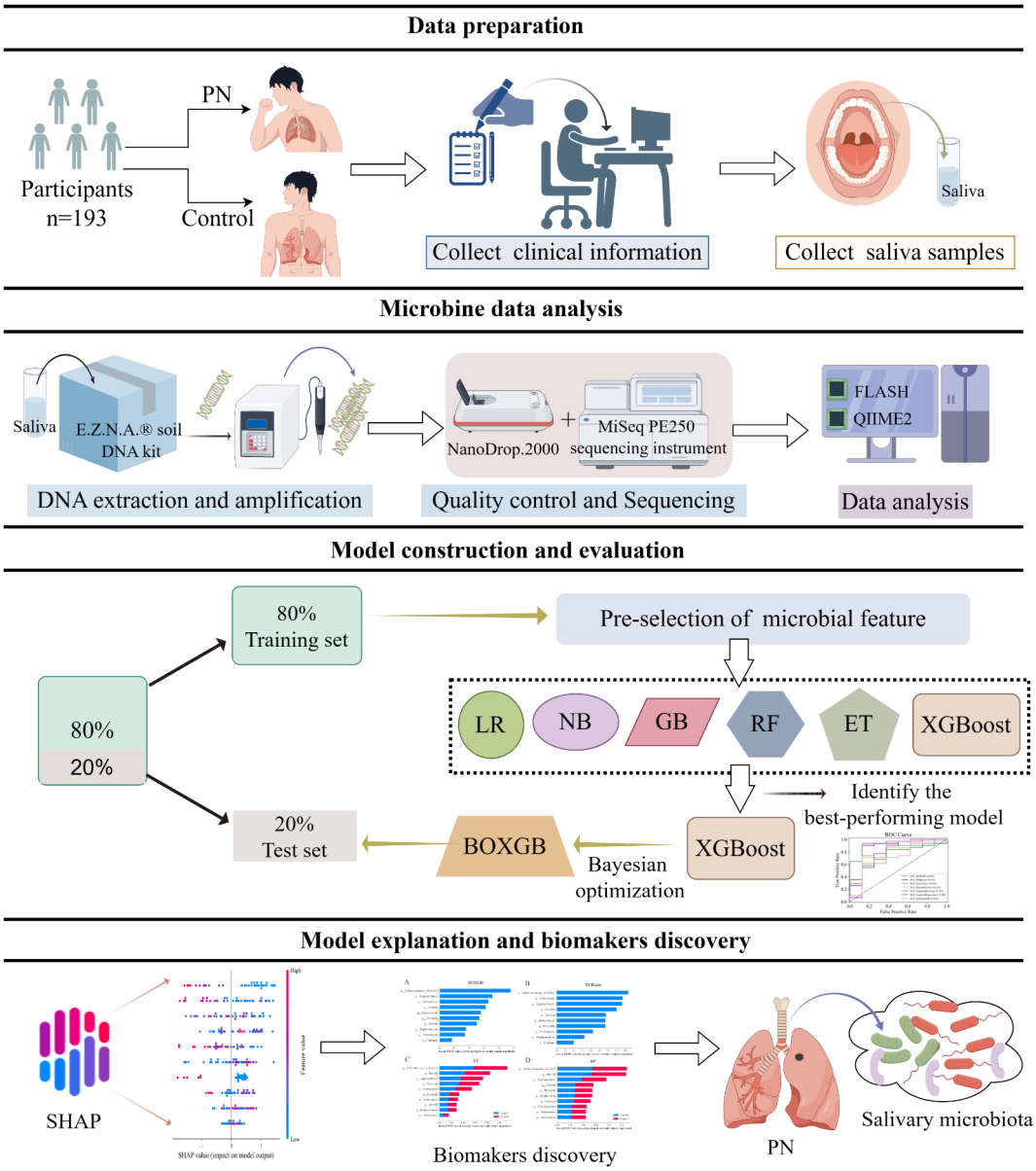

Fig. 1. Overview of the study this study.

correlated with salivary microbiota (Fig. 4E). Age was positively correlated with *Porphyromonas* (r = 0.30), *Neisseria* (r = 0.20), *Peptostreptococcus* (r = 0.20), *Capnocytophaga* (r = 0.16), and *Selenomonas* (r = 0.16) and negatively correlated with *Streptococcus* (r= -0.20) and *Haemophilus* (r= -0.18). Sex was positively correlated with *Porphyromonas*. Family tumor history was positively correlated with *TM7x* (r = 0.162) but showed no significant correlation with smoking or cancer history (Tables S2, S3). Subsequently, the results of the db-RDA based on the Bray-Curtis distance showed that all sample points formed two clusters, demonstrating the close association between the patients' clinical characteristics and salivary microbiota (Fig. 4F).

Prediction and analysis of saliva microbial function in PN

To further explore the potential biological relationship between microbial community data and pulmonary nodules, we performed a functional predictive analysis using PICRUSt2. Among them, the orthogonality (KO) function richness results based on the Kyoto Encyclopedia of Genes and Genomes (KEGG)[16–18] showed that, compared with the HC group, excinuclease ABC subunit A, DNA-binding protein HU-beta, U32 family peptidase, protein-Tyrosine phosphatase were more abundant in PN group. polar amino acid transport system permease protein, peptide/nickel transport system substrate-binding protein abundance is lower (Fig. 5A). Meanwhile, KEGG pathway difference analysis was conducted (Fig. 5B), and the results showed that, PN group Biosynthesis of other secondary metabolites, Translation, Energy metabolism, Metabolism of cofactors and so on There is a significant up-regulation in the abundance of vitamins pathway related genes, suggesting that salivary microbes may be involved in the evolution of pulmonary nodules through metabolic pathways.

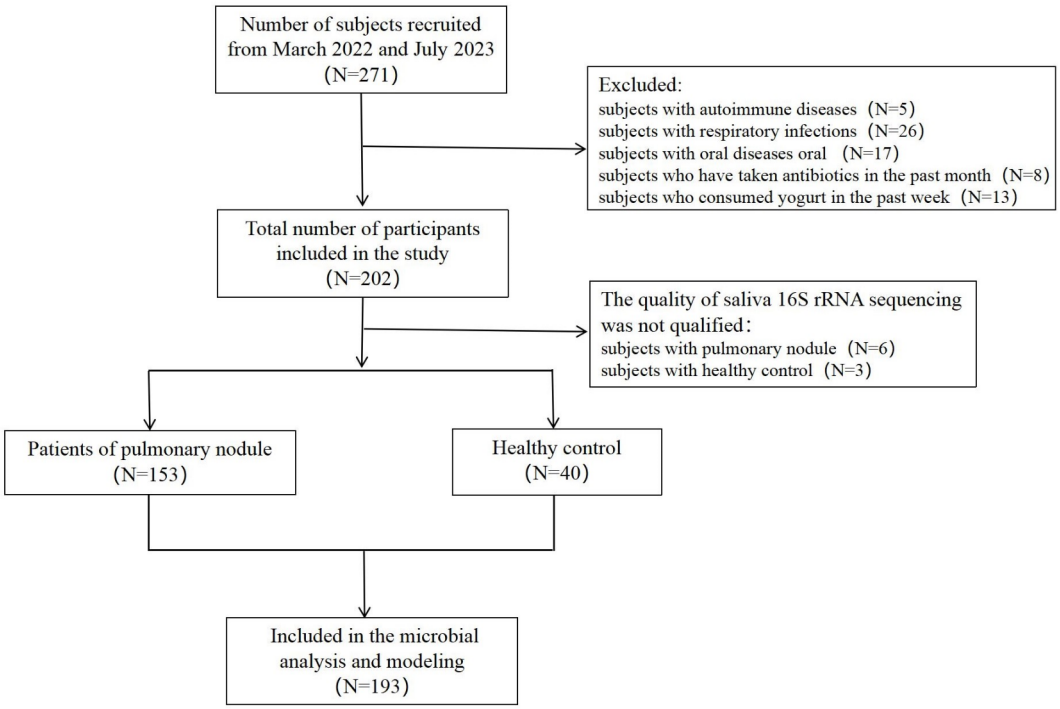

Fig. 2. Recruitment flow chart for this study.

	Total (n = 193)	Controls (n = 40)	PN (n = 153)	P value
Age(year)				0.06
Median	32.0	25.0	33.0	
[IQR]	[25.0, 39.0]	[25.0, 39.0]	[25.0, 39.0]	
Sex (%)				0.75
Female	136 (70.5)	29 (72.5)	107 (69.9)	
Male	57 (29.5)	11 (27.5)	11 (27.5)	
Personal cancer history (%)				0.07
No	173 (89.6)	39 (97.5)	134 (87.6)	
Yes	20 (10.4)	1 (2.5)	19 (12.4)	
Family cancer history (%)				0.10
No	139 (72.0)	33 (82.5)	106 (69.3)	
Yes	54 (28.0)	7 (17.5)	47 (30.7)	
Smoking status (%)				0.05
Never	171 (88.6)	39 (97.5)	132 (86.3)	
Ever	22 (11.4)	1 (2.5)	21 (13.7)	

Table 1. Baseline characteristics of study population. PN: pulmonary nodule, IQR: interquartile range. $P < 0.05$ was considered to indicate statistical significance.

Model construction and evaluation

First, the 193 patients were randomly divided into training and test sets at 80:20, with 154 and 39 patients included in the training and test sets, respectively. Subsequently, based on the random forest algorithm, the top 10 characteristic genera were selected as the final features for model construction, including *Oribacterium*, *Defluviitaleace UCG-011*, *Gemella*, *Aggregatibacter*, *Streptococcus*, *Lautropia*, *Bacillus*, *Actinomyces*, *Porphyromonas*, and *Prevotella* (Fig. S1). To assess the stability of the selected features, we conducted 100 independent experiments utilizing an 80:20 data split and documented the top 10 significant microbial features identified in each trial. The findings revealed that the identified microbial features consistently ranked among the top across repeated experiments, indicating substantial robustness and consistency. To further evaluate the model's robustness, these features were integrated into six machine learning models (LR, NB, GB, RF, ET, and XGBoost) for multiple training sessions. The results demonstrated that all models achieved satisfactory classification performance, with average AUC values ranging from 0.7352 ± 0.0777 to 0.8608 ± 0.0124 (Table 2).

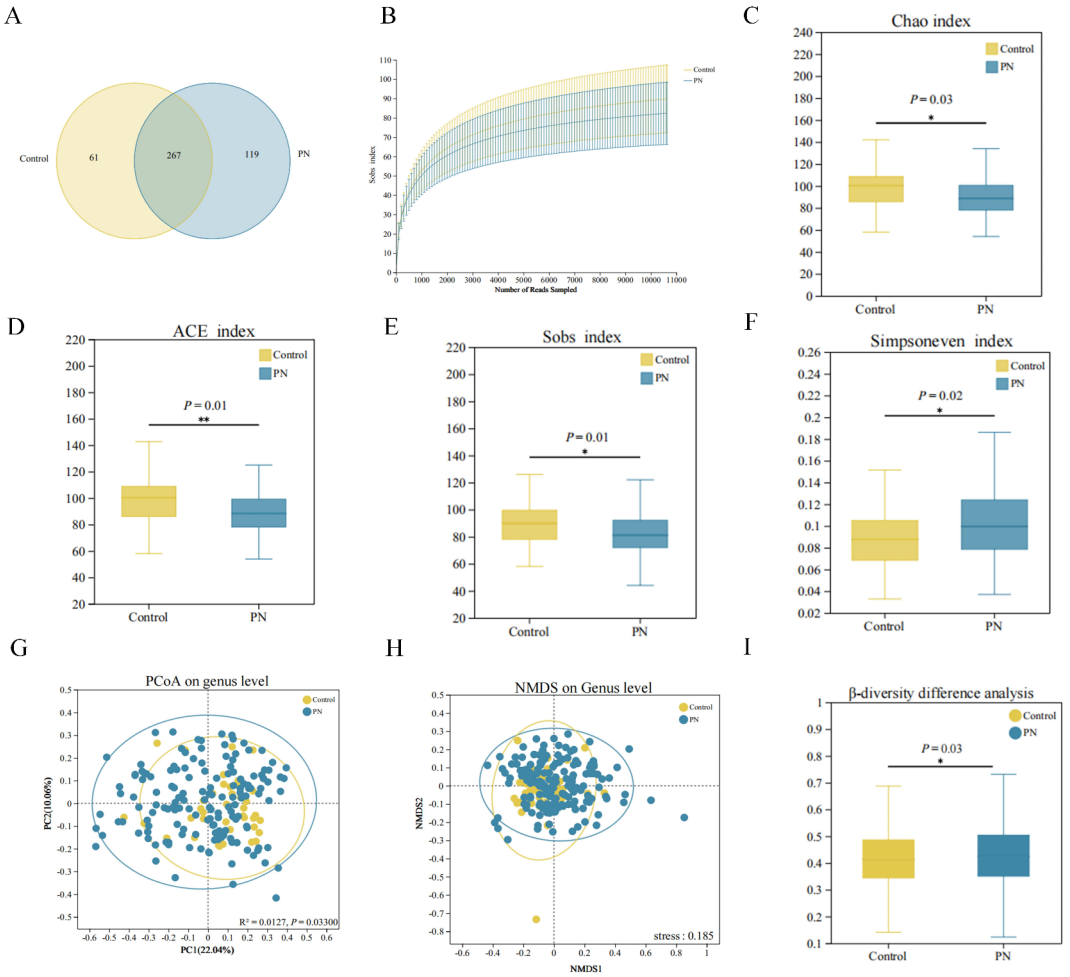

Fig. 3. Community diversity analysis between the groups. (**A**) Venn map. (**B**) Rarefaction curves based on sob. (**C**) α-diversity analysis based on Chao index. (**D**) α-diversity analysis based on ACE index. (**E**) α-diversity analysis based on Sobs index. (**F**) α-diversity analysis based on simpsoneven index. (**G**) PCoA analysis. (**H**) NMDS analysis. (**I**) β-diversity analysis. (PN: pulmonary nodule, $P<0.05$ was considered to indicate statistical significance).

A comprehensive assessment of model performance, based on metrics such as AUC, accuracy, and *F1*_score, identified the XGBoost model as the top performer (AUC = 0.8608 ± 0.0124). Subsequently, to further enhance model performance, the XGBoost model was fine-tuned using the Bayesian optimization algorithm, and the optimized model (BOXGB) was validated through 5-fold cross-validation. AUC of the optimized model increased to 0.8831 (Fig. 6A), indicating an improvement in predictive performance. To visually illustrate the correspondence between the model's predicted probabilities and actual observations, we constructed a calibration curve, which demonstrated a strong fit for the BOXGB model (Fig. 6B). Furthermore, a confusion matrix corroborated the model's reliable predictive capability (Fig. 6C).

To further assess the specificity of the selected features, additional experiments were conducted to compare the performance of a feature set comprising the top 10 selected features against a set utilizing all available features across six machine learning models. The experimental outcomes revealed that, in most instances, models trained with these 10 meticulously chosen features exhibited superior accuracy in predicting lung nodules compared to those trained on the complete feature set (Fig. 6A, E). This finding suggests that concentrating on a smaller, optimized subset of features can reduce model complexity, enhance computational efficiency, and potentially improve the model's generalization capability without compromising predictive performance. Moreover, considering potential variations in smoking status between the two sample groups, we evaluated the ability of the selected features to differentiate between smokers and non-smokers. The results generally indicated low AUC values, ranging from 0.3929 to 0.6071, for the selected models (Fig. 6F), suggesting a limited capacity of these features to distinguish between smokers and non-smokers. This finding substantiates our hypothesis that the predominant microbial features among the top 10 are primarily indicative of alterations in the microbial community associated with lung nodules, rather than variations directly attributable to smoking status.

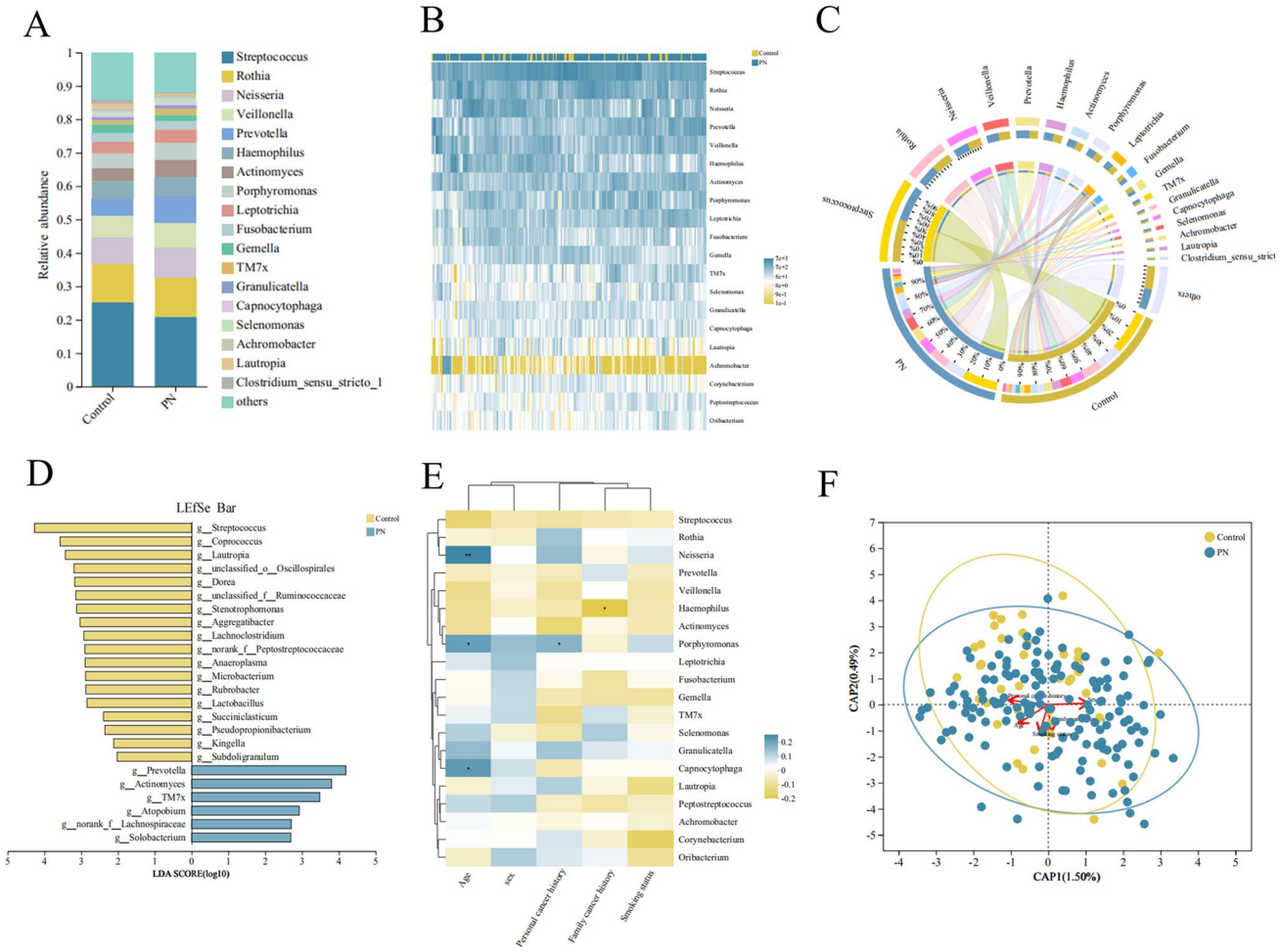

Fig. 4. Differences in microbial composition between the groups, and analysis of the correlation between clinical characteristics and microbiota. (**A**) Community Bar plot. (**B**) Heatmap shows the distribution of Top dominant species in the groups. (**C**) Circos sample and species relationship diagram. Half represents the species composition of the groups, while the other half represents the distribution proportions of species at the genus level in the groups. (**D**) LEfSe analysis. (**E**) A heatmap of clinical features and salivary genera based on Spearman correlation analysis. (**F**) db-RDA assessing the impact of clinical characteristics on microbial community composition based on Bray-Curtis distance. (PN: pulmonary nodule. *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$)

Model interpretation

Based on these results, we attempted to determine how the BOXGB model predicts PN using SHAP values. A summary plot of the SHAP values illustrated the impact and ranking importance of each microbial feature in the model output (Fig. 6D). The most influential microbial features in descending order were *Defluviitaleace_UCG-011*, *Aggregatibacter*, *Oribacterium*, *Gemella*, *Streptococcus*, *Prevotella*, *Bacillus*, *Porphyromonas*, *Actinomyces*, and *Lautropia*. To visually demonstrate the contribution of each microbial feature in the BOXGB model in each patient, we used the SHAP method to illustrate specific prediction scenarios for two representative samples, thereby elucidating the interpretability of the optimal model (Fig. 7A, B). For instance, features such as *Defluviitaleace_UCG-011* = 0, *Oribacterium* = 2, *Gemella* = 22, *Prevotella* = 502, *Bacillus* = 0, *Porphyromonas* = 112, and *Actinomyces* = 1.0 promoted predictions biased towards patients with PN, whereas features such as *Defluviitaleace_UCG-011* = 2, *Aggregatibacter* = 17, *Oribacterium* = 6, *Gemella* = 47, *Streptococcus* = 478, and *Porphyromonas* = 8 promoted predictions biased towards the controls.

Model display and application

Considering all performance evaluation metrics, the BOXGB model demonstrated the best performance. To facilitate the application of this study's findings clinically and among relevant researchers, we developed an interactive online prediction platform based on the BOXGB algorithm, which can be accessed for assessment at the following website: https://predictive-model-of-pulmonary-nodules.streamlit.app/ (Fig. S2). Users could input their salivary flora information into the artificial intelligence model interface. After clicking the "Submit" button, the model would calculate the risk of PN for the user. In some cases, if the web-based calculator has

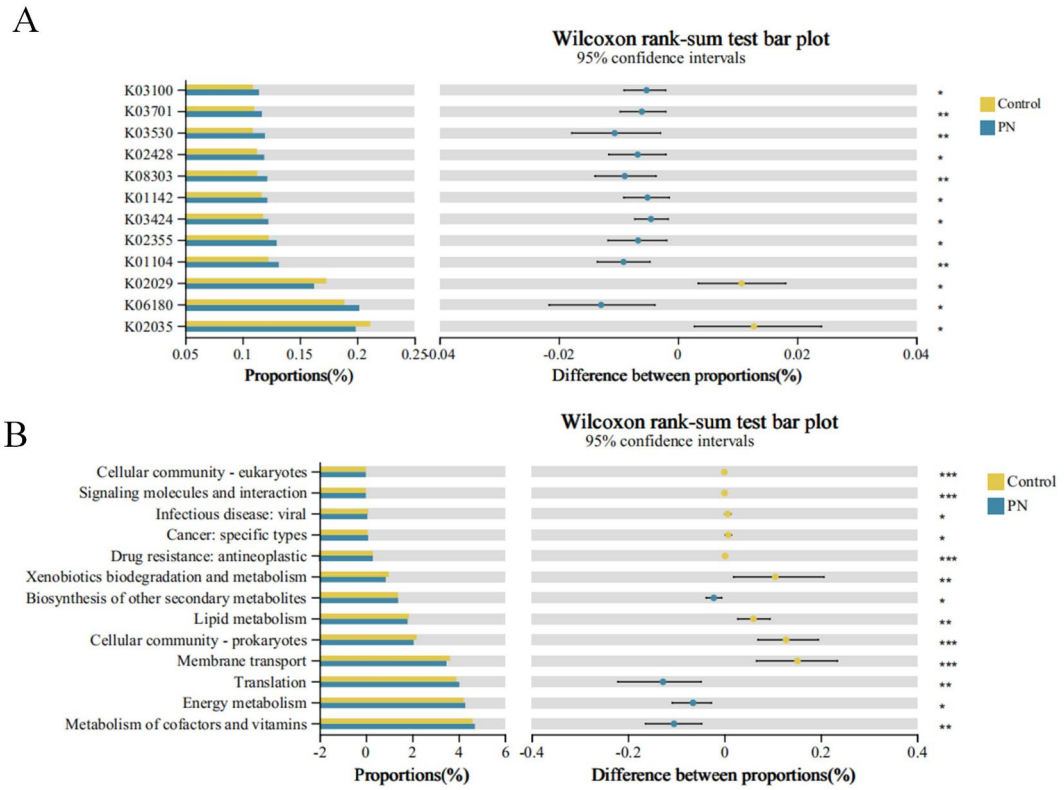

Fig. 5. Function prediction analysis was performed by PICRUSt2. (**A**) KO analysis showed that there were significant differences between Control and PN group. (**B**) Partial difference paths between Control and PN under KEGG primary classification (PN: pulmonary nodule. *: $P < 0.05$, * *: $P < 0.01$, * * *: $P < 0.001$.

Model	AUC	Accuracy	Precision	Recall	F1_score
LR	0.7432 ± 0.0507	0.7393 ± 0.0300	0.8815 ± 0.0309	0.7795 ± 0.0689	0.8250 ± 0.0270
NB	0.7560 ± 0.0580	0.7564 ± 0.0269	0.8395 ± 0.0248	0.8602 ± 0.0634	0.8479 ± 0.0239
GB	0.7896 ± 0.0445	0.7948 ± 0.0280	0.8648 ± 0.0347	0.8817 ± 0.0263	0.8725 ± 0.0154
ET	0.7352 ± 0.0777	0.7350 ± 0.0310	0.8858 ± 0.0457	0.7688 ± 0.0377	0.8218 ± 0.0193
RF	0.8017 ± 0.0521	0.8034 ± 0.0386	0.8220 ± 0.0315	0.9623 ± 0.0131	0.8864 ± 0.0205
XGboost	0.8608 ± 0.0124	0.7991 ± 0.0299	0.8715 ± 0.0519	0.8817 ± 0.0333	0.8750 ± 0.0148

Table 2. PN prediction performance of ML models for test sets (mean ± SD). PN, Pulmonary nodules; SD, Standard Deviation; LR, Logistic Regression; NB, Naïve Bayes; GB, Gradient Boosting; ET, Extra Tree; RF, Random Forest; XGBoost, eXtreme Gradient Boosting; AUC, Area under the receiver operating characteristic curve.

gone to sleep (shut down), it can be accessed by clicking "Yes, get this app back up!". The web-based calculator required approximately 30 s to load.

Exploring potential non-invasive biomarkers for PN using the SHAP method

We selected models with an AUC > 0.8 from the final seven ML models, including BOXGB, XGBoost, ET, and RF, and used SHAP for global interpretation to explore potential non-invasive biomarkers for PN (Fig. 8). The mean |SHAP value| provides an intuitive quantitative measure of the global importance of the predictive factors. Specifically, a higher mean |SHAP value| for a microbial feature implies a greater impact on predictions, indicating greater potential as a biomarker. Even for the same characteristic microbial features, the mean |SHAP value| may vary across different ML models, possibly because of differences in the models. To objectively select biomarkers, we used a threshold of the mean |SHAP value| > 0.6 across the BOXGB and XGBoost model and, and the mean |SHAP value| > 0.06 across the ET and RF model, we found *Defluviitaleace_UCG-011*, *Aggregatibacter*, *Oribacterium*, *Bacillus*, and *Prevotella* may be the most promising non-invasive biomarkers for PN.

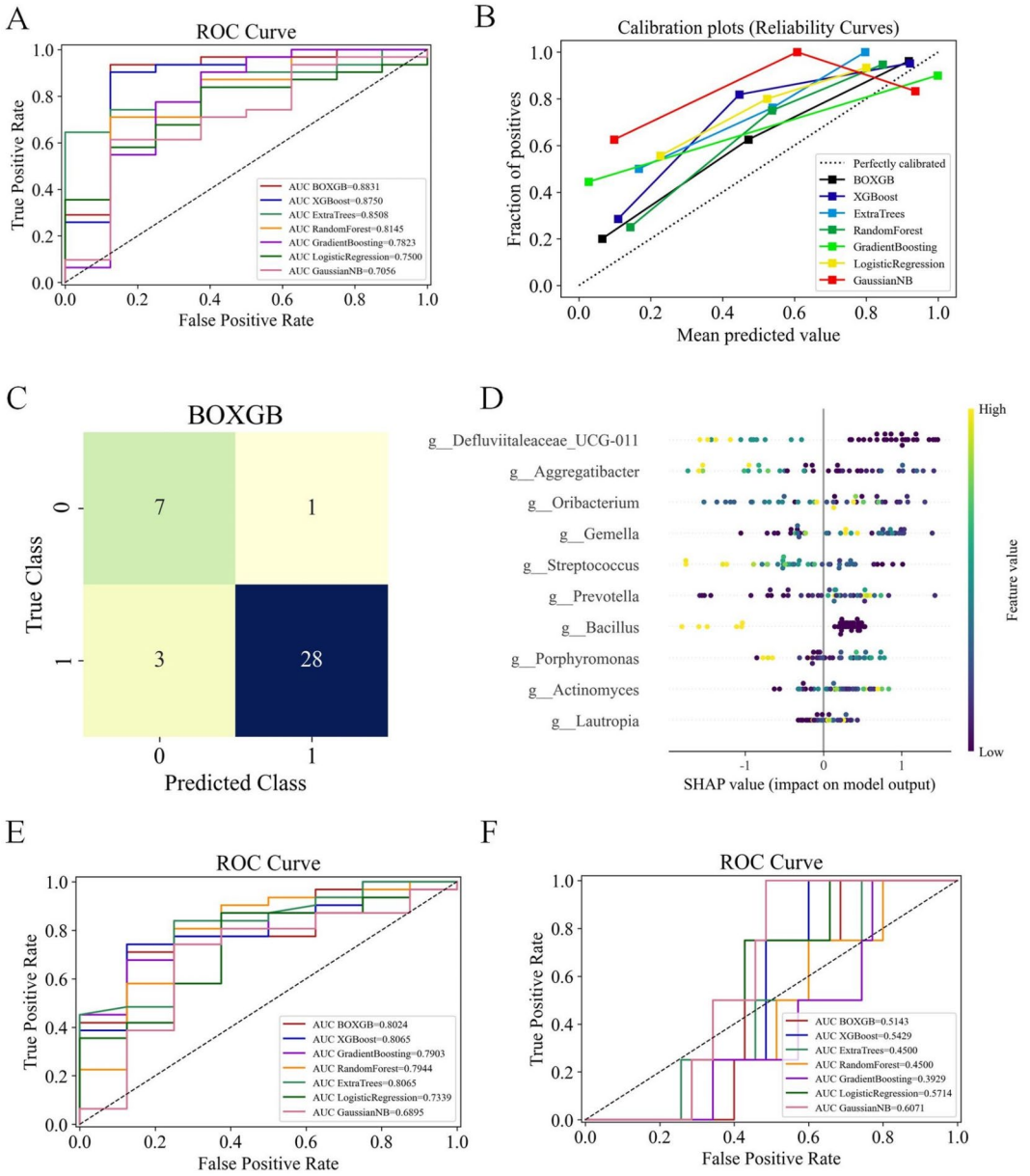

Fig. 6. The evaluation and interpretation of different Models. (**A**) The ROC curve of the test sets. (**B**) The calibration curve of the test sets. The closer the curve is to the 45-degree diagonal line, the more accurate the model's prediction probability. (**C**) The confusion matrix of the BOXGB model in the test set. (**D**) SHAP summary plot of the BOXGB model. It illustrates the distribution of the effects of 10 features on model outputs. SHAP creates a point for each feature attribution value for every patient, with points colored according to the corresponding patient's feature value and vertically stacked to depict density. Navy blue indicates higher feature encoding values, light yellow indicates lower feature encoding values, and the midpoint is represented by green. The further a point deviates from the baseline SHAP value of zero, the stronger its effect on the output. (**E**) Evaluating test set ROC curves based on all microbial characteristics. (**F**) Distinguishing smoking status based on Top10 microbial characteristics.

Discussion

PN refers to focal, round, increased-density opacities in the lungs with a diameter of ≤ 3 cm on imaging[19]. A sharp increase was observed in the global population of patients with PN who may be at risk of malignancy[20]. The salivary microbiota can influence the development of diseases such as lung cancer[6]. Studies have highlighted the high accessibility of saliva, making it a focus of biomarker studies[10]. This study aimed to compare the differences in salivary microbiota between patients with PN and controls and develop a diagnostic prediction model for PN by establishing various ML models. Additionally, we aimed to interpret the models and explore potential non-invasive biomarkers for PN using the SHAP method.

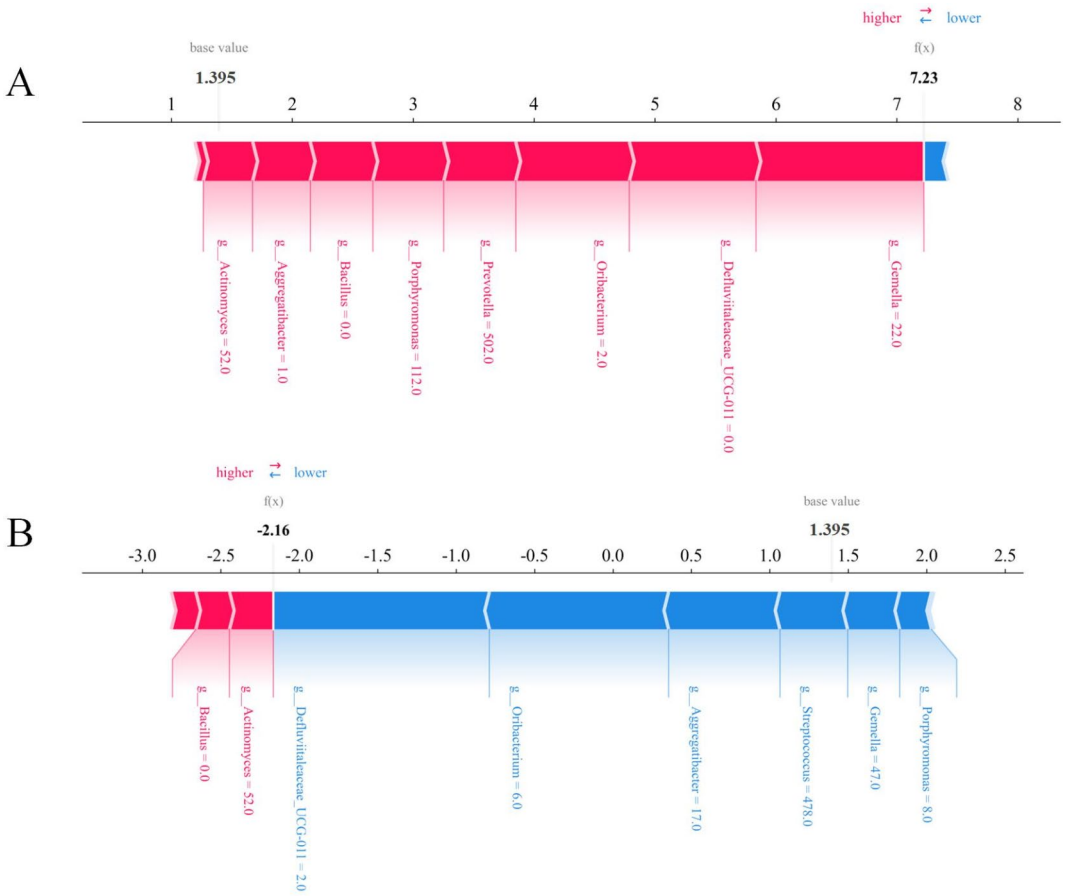

Fig. 7. SHAP models explain two typical predictions. (**A**) A high-risk SHAP interpretation model was shown for patients with pulmonary nodule. (**B**) A low-risk SHAP interpretation model for patients with pulmonary nodule. The force plots started at the base value (the average of all predictions). The base value is 1.395. Each predictor and its corresponding Shapley value were represented by an arrow, which either increased (shown in red) or decreased (shown in blue) the model's predicted value with respect to the base value. A predictor's importance was shown by the size of its arrow, where a larger arrow represents a more important predictor. Feature values were listed at the bottom of the plot. Finally, the predicted output value of the model was illustrated by the point where the red and blue arrows intersect.

In this study, we used 16 S rRNA gene sequencing to comprehensively describe the microbial profiles of the groups. Significant differences were observed in salivary microbiota between the groups using α- and β-diversity analyses. These results underscore the feasibility of using salivary microbiota to identify individuals with PN. Furthermore, this study revealed a significant association between the salivary microbiota and clinical characteristics of the participants. Prior research has confirmed age is a risk factor for PN[21]. In this study, variations in the salivary microbiota also showed the most significant correlation with age.

Owing to the high dimensionality and significant inter-individual variability of microbial communities, the correlation between microbial communities and PN has not yet been visualized through modeling analysis. Recent advancements in artificial intelligence and big data have led to the emergence of data-driven ML algorithms for medical research[22]. In this study, we partitioned the saliva microbial data into training and test sets at 8:2. Owing to the differences in computational efficiency, learning strategies, interpretability, and other aspects among the different models, we initially established six ML models for data training. XGBoost performed the best. Considering that hyperparameters have a more significant and crucial impact on model performance, this study further optimized the XGBoost model's hyperparameters, resulting in the BOXGB model. Notably, XGBoost, as a boosting algorithm, is a type of boosting tree model renowned for its Regularized Boosting technique[23]. This technique involves incorporating regularization terms into the cost function to control the model's complexity and prevent overfitting. In addition to Bayesian Optimization, other automatic parameter-tuning algorithms include grid and random searches[24,25]. Although random search has low time complexity, it may not precisely locate the optimal solution, whereas grid search, despite its high time complexity, yields better average results. Bayesian Optimization combines the advantages of both approaches, potentially providing superior and more stable performance. Consequently, this study opted for the Bayesian algorithm to optimize XGBoost with the following specific parameter settings: Scale_pose_weight = 0.7891633135336498, max_depth = 6, resulting in an increase in the AUC from 0.8750 to 0.8831. To facilitate the model's application in clinical

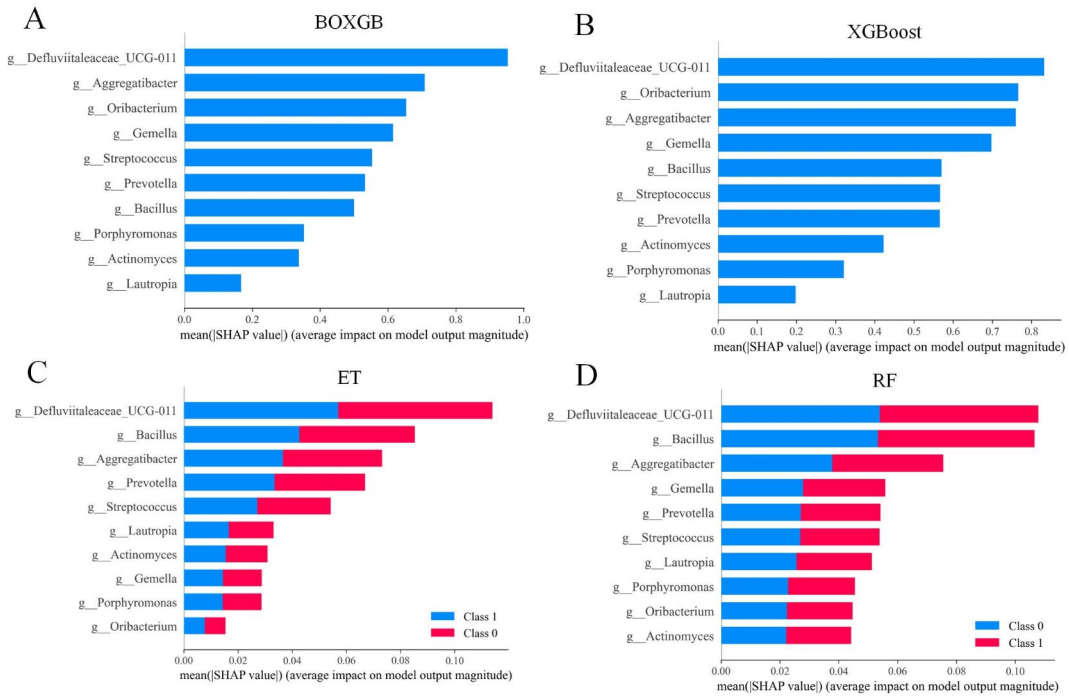

Fig. 8. SHapley Additive exPlanations (SHAP) feature importance shown according to the mean absolute SHAP value of each feature. (**A**) Summary Bar plot of the BOXGB model, (**B**) Summary Bar plot of the XGBoost model, (**C**) Summary Bar plot of the Extra Tree model, (**D**) Summary Bar plot of the Random Forest model. (Class 0: the controls, Class 1: pulmonary nodule).

and research settings, we developed an interactive online prediction platform where users can input relevant microbial information to obtain probability predictions for PN.

Additionally, this study used the SHAP method to explore potential non-invasive biomarkers for PN. The results revealed a complex and nonlinear relationship between salivary microbiota and the risk of PN, suggesting that salivary microbiota is a dynamic screening indicator for identifying high-risk individuals. Specifically, *Defluviitaleaceae_UCG-011* exhibited the highest mean |SHAP value|, indicating a decrease in the risk of PN with an increase in relative abundance, implying *Defluviitaleaceae_UCG-011* is a promising potential biomarker for PN and has a potential probiotic role in the human body. These findings are consistent with those of previous studies. For instance, Zhang et al. suggested that *Defluviitaleaceae_UCG-011* may have potential probiotic effects in preventing influenza, subacute thyroiditis, and hypothyroidism[26]. Furthermore, increasing the abundance of *Defluviitaleaceae_UCG-011* may improve conditions such as depression, cognitive impairment, and hyperlipidemia[27,28]. We found that *Defluviitaleaceae_UCG-011* is closely related to metabolic diseases, suggesting that it may be involved in the occurrence and development of PN through affecting metabolic pathways, which is consistent with our previous functional prediction results. In addition, Du et al. found that increasing the abundance of *Defluviitaleaceae_UCG-011* may promote the antifibrotic and cardioprotective effects of Astragaloside-IV[29]. Collectively, these studies confirmed that *Defluviitaleaceae_UCG-011* is a potential probiotic in the body, although the mechanisms related to its physiological and pathological changes in the human body remain unexplored. The genus *Aggregatibacter* includes species such as *Aggregatibacter actinomycetemcomitans* and *aphrophilus*. *Aggregatibacter actinomycetemcomitans* poses a threat to the host immune system and has been implicated in various periodontal diseases and malignant tumor progression. For example, its carriage is associated with an increased risk of pancreatic cancer[30]. *Aggregatibacter actinomycetemcomitans* can secrete leukotoxin (LtxA), and this toxin aids the bacteria in evading the host's immune response during infection. However, LtxA may serve as a novel candidate for the targeted biological therapy of leukocyte disorders, and preclinical drug development is currently underway[31]. Bacteria of the genus *Oribacterium* in patients with oral ulcers complicated by ulcerative colitis can serve as indicators for evaluating treatment efficacy[32]. Additionally, the abundance of *Oribacterium* in the oral cavity of patients with liver cancer was significantly increased compared with that in normal individuals, indicating a potential biomarker value for the auxiliary non-invasive diagnosis of liver cancer[33]. Most species within the genus *Bacillus* are harmless bacteria, with some strains used as probiotics, such as *Bacillus coagulans* and *subtilis*[34,35]. The genus *Prevotella* is the dominant microbial group in patients with PN and lung cancer. Roy et al. found a significant increase in the *Prevotella* genus in saliva samples of patients with lung adenocarcinoma[36]. Further, Tsay et al. also found that the enrichment of *Prevotella* may promote lung cancer progression by upregulating carcinogenic signaling pathways, such as PI3K and ERK[37]. Collectively, these studies support the potential of *Defluviitaleaceae_UCG-011*, *Aggregatibacter*, *Oribacterium*, *Bacillus*, and *Prevotella* as non-invasive biomarkers with the highest potential for PN.

In conclusion, this study elucidated the differential microbial profiles of saliva microbiota between patients with PN and controls. By constructing ML models and training them on important feature variables, we established a diagnostic model for PN based on salivary microbiota data and developed an online prediction platform. Furthermore, we proved that salivary microbiota could serve as non-invasive liquid biomarkers for PN diagnosis, providing more scientific evidence for intervention in PN from the perspective of microbiota and offering new insights for expanding clinical diagnostic approaches for PN. Future studies should validate the identified salivary microbiota biomarkers in larger and more diverse patient cohorts.

Limitations of the study

This study had some limitations. Firstly, the sample size was relatively small, which may have limited the generalizability of the results. To enhance the significance of this study's findings, future studies should focus on enlarging the sample size, including larger samples from different regions, to better control for potential confounders and individual differences, thereby improving the external validity of the results. Secondly, various factors, including smoking, dietary habits, and pharmacological interventions, have the potential to influence the composition of the oral microbiota. Subsequent research should aim to elucidate the specific effects of smoking and other lifestyle determinants on the salivary microbiota. Thirdly, the model established in this study could only distinguish between patients with PN and controls. Future studies should collect saliva samples from patients with benign and malignant PN to establish a benign-malignant discrimination model based on saliva microbiota information. Additionally, although potential biomarkers were identified through SHAP analysis, which needs to be validated by basic experiments and clinical trials.

Methods

Ethics approval and consent to participate

We selected patients with PN between August 2022 and September 2023 at the Hospital of Chengdu University of TCM, Chengdu Integrated TCM & Western Medicine Hospital, and Sichuan Cancer Hospital (ChiCTR2200062140; 25/07/2022). Controls consisted of healthy individuals matched for clinical information, such as age, sex, and tumor history, with no PN evident on chest CT imaging. The participants were volunteers aged between 18 and 60 years. Participants were excluded based on the following criteria: history of untreated infectious diseases, history of autoimmune diseases, history of respiratory infections and oral diseases, oral intake of antibiotics within the past month, or consumption of yogurt within the past 7 days.

Sample information collection and 16 S rRNA gene sequencing

We collected clinical information on patients with PN and controls, such as age, sex, smoking history, tumor history, and family tumor history. Collection of saliva specimens: Sampling was performed in a sterile room by research personnel trained in sampling techniques. The participants were required to fast for 30 min before sample collection and rinse their mouths with 10 ml of saline solution before sampling. Non-stimulating saliva samples of 1–2 mL was collected in sterile Eppendorf tubes. Compared with stimulating saliva, non-stimulating saliva has a lower secretion rate, stays longer in the oral cavity, and is more likely to capture microorganisms from various parts of the oral cavity[38]. Subsequently, the samples were preserved on dry ice and transferred to the laboratory within 4 h.

Following the instructions of the E.Z.N.A.® soil DNA kit (Omega Bio.-tek, Norcross, GA, U.S.), saliva microbiota genomic DNA was extracted. The quality of the extracted genomic DNA was assessed using 1% agarose gel electrophoresis. DNA concentration and purity were quantified using a NanoDrop.2000 (Thermo Scientific, USA). The extracted DNA was then used as a template for amplifying the V3–V4 variable region of the genes using upstream primer 338 F (5'-ACTCCTACGGGAGGCAGCAG-3') and downstream primer 806R (5'-GGACTACHVGGGTWTCTAAT-3')[39] carrying Barcode sequences for 16 S rRNA gene amplification (ABI GeneAmp® 9700)[40]. Subsequently, the normalized equimolar concentrations of each amplicon were pooled and sequenced on the MiSeq PE250 sequencing instrument (Illumina, San Diego, CA, USA). Quality control of the raw sequences was performed using the FASTP online platform (version 0.19.6). Sequence merging was conducted using FLASH software (version 1.2.11) to obtain the optimized data. The DADA2 plugin was used to perform quality control and filter the sequence data, resulting in amplicon sequence variants (ASVs)[41] representing the sequence and abundance information. Species annotation of ASVs was performed using the silva138/16s_bacteria reference database, and the Naïve Bayes classifier in QIIME2[42] (version 2022.2) was used for the species annotation of ASVs with a confidence threshold of 70%. Sample information was analyzed based on ASVs, and various indicators were further analyzed, including sequencing depth. Community structure statistics were performed at the genus level based on taxonomic information to obtain the microbial community information for each sample.

Microbiome data analysis

We used Mothur (version v.1.30.2, https://mothur.org/wiki/calculators/) for α-diversity index analysis, generating inter-group difference plots (sobs, ace, chao, simpsoneven) and rarefaction curves. Principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS) were conducted using the weighted_unifrac algorithm, while inter-group differences in β-diversity were visualized based on the Bray-Curtis distance algorithm. Linear discriminant analysis (LDA) effect size (LEfSe) was performed using Galaxy Platform (http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload)[43] on samples based on different grouping conditions, with a threshold LDA score of > 2.0, based on the genus level. Additionally, community structure data tables at the genus level were used to generate community heatmap plots using the pheatmap package. Community Circos plots were generated using Circos-0.67-7 (http://circos.ca/). The R language vegan package was used for Variance Inflation Factor (VIF) analysis to select representative clinical features[44], enhancing

the credibility of subsequent analyses correlating with microbial communities. Distance-based redundancy analysis[45] (db-RDA) was used based on the Bray-Curtis distance to assess the impact of clinical features on the composition of salivary microbial communities. Spearman's correlation coefficients were used to evaluate the correlation between the salivary microbial community composition and clinical features, and heatmap plots were generated.

Model construction and optimization

To establish a PN diagnostic model based on salivary microbiota data, the dataset was randomly divided into training (80%) and test sets (20%) using random sampling. Considering the different characteristics of various models regarding computational speed, assumptions regarding the nature of the data, and interpretability, we initially established six ML classifiers: Logistic Regression (LR), Naïve Bayes (NB), Gradient Boosting (GB), Random Forest (RF), Extra Trees (ET), and eXtreme Gradient Boosting (XGBoost). These models were implemented using the Python scikit-learn library[46].

LR is one of the most widely used binary classification algorithms and is considered the gold standard for analyzing binary medical data[47]. Its primary function is to establish the relationship between dependent and independent variables, demonstrating extensive applicability across various domains and reliable modeling capabilities[48]. NB estimates the conditional probability of each category for each feature by assuming that P(x/yi) follows a Gaussian distribution. NB has the advantages of stable classification efficiency, a relatively simple algorithm, and good performance on small-scale data[49]. GB is an effective technique for handling diverse datasets and constructing powerful predictive models by integrating multiple weak learners. GB is capable of managing imbalanced datasets but is susceptible to overfitting[50]. In addition, the training time may be relatively long owing to the iterative training of multiple models. RF improves the predictive performance and generalization capability by aggregating decision trees[51]. RF exhibits significance against overfitting and applies to high-dimensional and large-scale datasets. ET is an ensemble learning algorithm that introduces more randomness in building decision trees, reducing overfitting risk and requiring fewer parameter adjustments. ET performs exceptionally well in high-dimensional and large-sample scenarios[52]. XGBoost is a ML algorithm based on the gradient boosting framework that aims to enhance predictive performance by integrating multiple decision tree models[23]. XGBoost uses first- and second-order gradient information to optimize the objective function and introduces regularization terms to control model complexity. XGBoost supports parallel processing and exhibits excellent performance when handling large-scale datasets and high-dimensional features.

Owing to the impact of hyperparameters on the performance of ML models, selecting hyperparameters is crucial. A Bayesian optimization algorithm was used to identify the optimal hyperparameter combination[53]. This algorithm is a global optimization method that rapidly identifies the global optimal solution by updating the prior probability model and considering the parameter information. We selected the best-performing model among the six initially established models and optimized its hyperparameters.

Model assessment and explanation

We evaluated the performance using metrics such as the area under the receiver operating characteristic curve (AUC) and F1 score. Thereafter, a web-based calculator for predicting PN was developed using the "Strealit" (https://share.streamlit.io/) application regarding the optimal mode[54]. Despite the potential benefits of ML algorithms in improving clinical decision-making accuracy for patients, their decision-making process is often perceived as a "black box," posing challenges for interpretation. This study addressed the "black box" challenge by using the SHAP algorithm to elucidate the optimal model[55]. SHAP is an additive explanation framework based on game theory perspectives developed by Lundberg et al., which elucidates the contribution of each feature value to the diagnosis[56]. Compared with other interpretability methods, the SHAP method possesses three desirable properties: local accuracy, missingness, and consistency[57]. The tree-based SHAP algorithm from Lundberg's Python 'SHAP' package was used. The SHAP algorithm iteratively removes each feature, computes changes in model accuracy, and predicts the malignancy rate to obtain Shapley values for specific features and their directional impact on outcomes. By visualizing the importance of the SHAP feature and summary plots, the contribution of the features to the model and their value in differentiating patients were depicted. To demonstrate the operation of the ML model, SHAP force plots were generated to explain the individual diagnoses of several representative cases from the test set, where the red and blue arrows represent increased and decreased diagnostic probabilities from the baseline, respectively.

Statistical analysis

Clinical baseline characteristics were analyzed using the SPSS software (version 26.0). Given the non-normal distribution of continuous variables in the clinical information of this study, Wilcoxon rank-sum tests were used for intergroup comparisons, with results presented as medians and interquartile ranges. Wilcoxon rank-sum tests were also used for testing inter-group differences in α- and β-diversity, and false discovery rate correction was applied for multiple testing adjustments. Adonis was chosen for intergroup difference testing in PCoA and NMDS analyses. LEfSe analysis used an all-against-all (stricter) multigroup comparison strategy. $P < 0.05$ was considered statistically significant.

Data availability

The raw metagenomic data generated in this study have been deposited in the NCBI Sequence Read Archive, with the accession code PRJNA1114406.

Received: 22 November 2024; Accepted: 24 March 2025

Published online: 29 March 2025

References

1. Han, B. et al. Cancer incidence and mortality in China, 2022. *J. Natl. Cancer Cent.* **4**, 47–53. https://doi.org/10.1016/j.jncc.2024.01.006 (2024).
2. Guan, Y., Ren, M., Guo, D. & He, Y. Research progress on lung cancer screening. *Zhongguo Fei Ai Za Zhi* **23**, 954–960 (2020).
3. 125 Questions: Exploration and discovery. https://www.science.org/content/resource/125-questions-exploration-and-discovery. Accessed 17 July 2024.
4. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature* **569**, 641–648 (2019).
5. Farrell, J. J. et al. Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut* **61**, 582–588 (2012).
6. Ma, Q. et al. Mechanisms underlying the effects, and clinical applications, of oral microbiota in lung cancer: current challenges and prospects. *Crit. Rev. Microbiol.* **50**, 631–652. https://doi.org/10.1080/1040841X.2023.2247493 (2024).
7. Mäkinen, A. I. et al. Salivary Microbiome profiles of oral cancer patients analyzed before and after treatment. *Microbiome* **11**, 171 (2023).
8. Zuo, H. J. et al. Study on the salivary microbial alteration of men with head and neck cancer and its relationship with symptoms in Southwest China. *Front. Cell. Infect. Microbiol.* **10**, 514943 (2020).
9. Stasiewicz, M. & Karpinski, T. M. The oral microbiota and its role in carcinogenesis. *Semin. Cancer Biol.* **86**, 633–642 (2022).
10. Song, M., Bai, H., Zhang, P., Zhou, X. & Ying, B. Promising applications of human-derived saliva biomarker testing in clinical diagnostics. *Int. J. Oral Sci.* **15**, 2 (2023).
11. Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A. & &, J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* **186**, 1772–1791 (2023).
12. Shehab, M. et al. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **145**, 105458 (2022).
13. Liu, Q. et al. Gut mycobiome as a potential non-invasive tool in early detection of lung adenocarcinoma: A cross-sectional study. *BMC Med.* **21**, 409 (2023).
14. Liu, Y. et al. Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. *Cell. Metab.* **34**, 719–730e4 (2022).
15. Lee, Y. W., Choi, J. W. & Shin, E. H. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med.* **129**, 104151 (2021).
16. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: Biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
17. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
18. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
19. Au-Yong, I. T. H., Hamilton, W., Rawlinson, J. & Baldwin, D. R. Pulmonary nodules. *BMJ* **371**, m3673 (2020).
20. Adams, S. J. et al. Lung cancer screening. *Lancet* **401**, 390–408 (2023).
21. McWilliams, A. et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N. Engl. J. Med.* **369**, 910–919 (2013).
22. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
23. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* 785–794 (Association for Computing Machinery, 2016).
24. Kim, Y., Kim, K. H., Park, J., Yoon, H. I. & Sung, W. Prognosis prediction for glioblastoma multiforme patients using machine learning approaches: development of the clinically applicable model. *Radiother. Oncol.* **183**, 109617 (2023).
25. Tang, R., Luo, R., Tang, S., Song, H. & Chen, X. Machine learning in predicting antimicrobial resistance: A systematic review and meta-analysis. *Int. J. Antimicrob. Agents* **60**, 106684 (2022).
26. Zhang, X. et al. Causal effect of gut microbiota of defluviitaleaceae on the clinical pathway of Influenza-Subacute Thyroiditis-Hypothyroidism. *Front. Microbiol.* **15**, 1354989 (2024).
27. Zhang, Q. et al. The beneficial effects of *Lactobacillus brevis* FZU0713-fermented *Laminaria japonica* on lipid metabolism and intestinal microbiota in hyperlipidemic rats fed with a high-fat diet. *Food Funct.* **12**, 7145–7160 (2021).
28. Huang, H. S. et al. Anti-depressive-like and cognitive impairment alleviation effects of Gastrodia Elata Blume water extract is related to gut Microbiome remodeling in ApoE(-/-) mice exposed to unpredictable chronic mild stress. *J. Ethnopharmacol.* **302**, 115872 (2023).
29. Du, X. Q. et al. Astragaloside IV ameliorates Isoprenaline-Induced cardiac fibrosis in mice via modulating gut microbiota and fecal metabolites. *Front. Cell. Infect. Microbiol.* **12**, 836150 (2022).
30. Fan, X. et al. Human oral Microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* **67**, 120–127 (2018).
31. Kachlany, S. C. Aggregatibacter actinomycetemcomitans leukotoxin: From threat to therapy. *J. Dent. Res.* **89**, 561–570 (2010).
32. Xu, J. et al. The oral bacterial microbiota facilitates the stratification for ulcerative colitis patients with oral ulcers. *Ann. Clin. Microbiol. Antimicrob.* **22**, 99 (2023).
33. Li, B. L., Cheng, L., Zhou, X. D. & Peng, X. Research progress on the relationship between oral microbes and digestive system diseases. *Hua Xi Kou Qiang Yi Xue Za Zhi* **36**, 331–335 (2018).
34. Marzorati, M. et al. A. Bacillus subtilis HU58 and Bacillus coagulans SC208 probiotics reduced the effects of antibiotic-Induced gut Microbiome dysbiosis in an M-SHIME((R)) model. *Microorganisms* **8**, 1028 (2020).
35. Wauters, L. et al. Efficacy and safety of spore-forming probiotics in the treatment of functional dyspepsia: A pilot randomised, double-blind, placebo-controlled trial. *Lancet Gastroenterol. Hepatol.* **6**, 784–792 (2021).
36. Roy, P., Sarma, A., Kataki, A. C., Rai, A. K. & Chattopadhyay, I. Salivary microbial dysbiosis may predict lung adenocarcinoma: A pilot study. *Indian J. Pathol. Microbiol.* **65**, 123–128 (2022).
37. Tsay, J. J. et al. Airway microbiota is associated with upregulation of the PI3K pathway in lung cancer. *Am. J. Respir. Crit. Care Med.* **198**, 1188–1198 (2018).
38. Lu, H. et al. The sampling strategy of oral Microbiome. *Imeta* **1**, e23 (2022).
39. Liu, J. H. et al. Comparative studies of the composition of bacterial microbiota associated with the ruminal content, ruminal epithelium and in the faeces of lactating dairy cows. *Microb. Biotechnol.* **9**, 257–268 (2016).
40. Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
41. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
42. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
43. Boekel, J. et al. Multi-omic data analysis using galaxy. *Nat. Biotechnol.* **33**, 137–139 (2015).
44. Cheng, J., Sun, J., Yao, K., Xu, M. & Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim. Acta Mol. Biomol. Spectrosc.* **268**, 120652 (2022).

45. Jiang, R., Wang, M., Xie, T. & Chen, W. Site-specific ecological effect assessment at community level for polymetallic contaminated soil. *J. Hazard. Mater.* **445**, 130531 (2023).
46. Pedregosa, F. et al. Scikit-learn: Mmachine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Swensen, S. J., Silverstein, M. D., Ilstrup, D. M., Schleck, C. D. & Edell, E. S. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch. Intern. Med.* **157**, 849–855 (1997).
48. Dai, R. et al. BBPpred: Sequence-Based prediction of Blood-Brain barrier peptides with feature representation learning and logistic regression. *J. Chem. Inf. Model.* **61**, 525–534 (2021).
49. Cui, S. et al. Using Naive Bayes classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. *Injury* **49**, 1865–1870 (2018).
50. Cunningham, L. et al. Gradient boosting approaches can outperform logistic regression for risk prediction in cutaneous allergy. *Contact Dermat.* **86**, 165–174 (2022).
51. Wang, C. et al. Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease. *Comput. Methods Programs Biomed.* **188**, 105267 (2020).
52. Chen, A. Y., Lee, J., Damjanovic, A. & Brooks, B. R. Protein pK(a) prediction by Tree-Based machine learning. *J. Chem. Theory Comput.* **18**, 2673–2686 (2022).
53. Zheng, J. et al. Metabolic syndrome prediction model using bayesian optimization and XGBoost based on traditional Chinese medicine features. *Heliyon* **9**, e22727 (2023).
54. Li, Y. et al. Machine learning-based models to predict one-year mortality among Chinese older patients with coronary artery disease combined with impaired glucose tolerance or diabetes mellitus. *Cardiovasc. Diabetol.* **22**, 139 (2023).
55. Bernard, D. et al. Explainable machine learning framework to predict personalized physiological aging. *Aging Cell.* **22**, e13872 (2023).
56. Scott, M. & Lundberg & Su-In Lee. *A Unified Approach to Interpreting Model Predictions* 4768–4777 (2017).
57. Zhang, L. et al. An interpretable machine learning model based on contrast-enhanced CT parameters for predicting treatment response to conventional transarterial chemoembolization in patients with hepatocellular carcinoma. *Radiol. Med.* **129**, 353–367 (2024).

Acknowledgements

This work was supported by the National Nature Science Foundation of China (82405354), the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20230339, and the China Postdoctoral Science Foundation (2023MD744129), and Sichuan Province Science and Technology Support Program (2022ZDZX0022), and Science and Technology Research Special Project of Sichuan Provincial Administration of Traditional Chinese Medicine (2023ZD06).

Author contributions

C.H.: conceived and designed the experiments, wrote the first draft of the manuscript. Q.M.: conceived and designed the experiments, interpreted the results. X.Z.: analyzed the data and built models. J.H.: contributed to data collection, F.Y., X.F. and Y.R.: conceived the study, supervised the work, and reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study adhered to the principles of the Declaration of Helsinki and was approved by the Ethics Committee of the Hospital of Chengdu University of TCM (Ethical Approval No. 2022KL-051). The trial was registered with the China Clinical Trial Registration Center (Registration No. ChiCTR220062140).

Informed consent

All participants signed written informed consent.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-95692-6.

Correspondence and requests for materials should be addressed to F.Y., X.F. or Y.R.