International Journal of Chronic Obstructive Pulmonary Disease

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Construction and Validation of a Machine Learning Model Based on Clinical and Microbiomic Features for Predicting High Mucus Secretion in COPD

Qingqing Liu[1], Hui Zhao[2]

[1]General Practice Department, Changzhi People' s Hospital, Changzhi City, Shanxi Province, 046000, People's Republic of China; [2]Department of Respiratory and Critical Care Medicine, The second Hospital of Shanxi Medical University, Taiyuan City, Shanxi Province, 030000, People's Republic of China

Correspondence: Hui Zhao; Qingqing Liu, Email hui_zhao@sxmu.edu.cn; 18234125163@139.com

Objective: To evaluate clinical and airway microbiome features of excessive mucus secretion (CMH) in COPD progression and apply machine learning for CMH status identification.

Methods: A total of 319 COPD patients from Changzhi People's Hospital (May 2020–March 2024) were consecutively enrolled and divided by sputum volume and characteristics into a high mucus secretion group (n=173) and a non-high mucus secretion group (n=146). Patients were randomly assigned to training (80%) and testing (20%) sets. Airway microbiome structure was analyzed via 16S rRNA sequencing. From clinical and microbiome data, 70 features were extracted. Six machine learning algorithms (SVM, KNN, RF, BN, GBDT, NN) were used to build classification models. Feature selection employed filtering methods, and hyperparameters were optimized by 10-fold cross-validation. Model performance was assessed using sensitivity, specificity, accuracy, and AUC.

Results: The CMH group and the non-CMH group differed significantly in a number of factors, including age, the length of the disease, and pulmonary function indices, according to a comparison of baseline patient data. Analysis of airway microbiome characteristics revealed that the CMH group had significantly lower observed ASVs and Shannon indices ($p<0.001$), along with significant enrichment of potentially pathogenic bacterial genera such as Haemophilus and Pseudomonas. Following feature selection, disease duration, Haemophilus abundance, history of AECOPD, Pseudomonas abundance, and predicted FEV1% were identified as significant predictive factors. With a sensitivity of 0.867, specificity of 0.789, PPV of 0.805, NPV of 0.855, and AUC of 0.911, the Bayesian Network (BN) model outperformed the other six machine learning models on the testing sets; its generalization ability was significantly superior to other algorithms such as SVM and RF.

Conclusion: CMH in COPD is linked to airway dysbiosis and pathogen enrichment. The BN model effectively identifies this phenotype with strong generalization ability.

Keywords: COPD, CMH, airway microbiota, machine learning, BN

Introduction

One of the main characteristics of chronic obstructive pulmonary disease (COPD), a prevalent chronic respiratory disorder that mostly consists of emphysema and chronic bronchitis (CB), is chronic mucus hypersecretion (CMH).[1] The prevalence of this disease is approximately 8.6% among people aged 20 and older, and as high as 12.64% among those aged 40 and older. The 5-year mortality rate is 50%, and the 7.7-year mortality rate is 75%. It severely impacts patients' quality of life and has become a major cause of death and disability worldwide.[2] When compared to COPD patients without CMH, those with CMH usually have more severe bacterial airway infections, more noticeable lung function reductions, and worse health.[3] Because the airways are directly exposed to the external environment, they are susceptible to invasion by external pathogenic microorganisms, which can trigger a local immune inflammatory

https://doi.org/10.2147/COPD.S620360
Received: 27 April 2026
Accepted: 13 July 2026
Published: 22 July 2026

response, thereby further exacerbating the progression of CMH.[4] At the same time, as environmental pollution worsens, the incidence of COPD continues to rise, placing a heavy medical and economic burden on society.[5] Therefore, early identification and intervention for CMH, along with understanding the characteristics of the airway microbiome in the progression of COPD, are of great significance for slowing disease progression and improving patient outcomes.

Artificial intelligence can analyze large amounts of data through machine learning to identify complex, nonlinear relationships, thereby helping to improve the accuracy of early disease diagnosis. In recent years, with the rapid development of artificial intelligence in the medical field, techniques such as support vector machines (SVM), random forests (RF), and Bayesian models (BN) have been applied to the early screening and diagnosis of COPD, as well as clinical staging and risk assessment.[6] Nowadays, specimens obtained from oropharyngeal swabs, induced sputum, bronchoscopic brushings, bronchoalveolar lavage fluid, and nasal mucosal brushings are frequently analyzed using 16S rRNA gene sequencing to ascertain the microbial species composition and community structure.[7,8] Research has demonstrated that as the disease progresses and clinical features change, the makeup of the airway microbiota in COPD patients varies considerably.[9] Although existing studies have preliminarily revealed associations between the airway microbiome and clinical characteristics in COPD, most predictive models still rely primarily on traditional clinical information or single biomarkers. Furthermore, existing research has largely focused on the overall airway microbiome characteristics of COPD patients or has only compared groups based on the presence or absence of CMH, lacking systematic exploration and functional analysis of the microbial community structure specific to the CMH-associated COPD phenotype. Therefore, this study combines clinical information with pulmonary microbiome data and employs machine learning methods to construct a model for identifying CMH-associated COPD and to explore the characteristics of its microbial community structure. The aim is to provide new microbiological evidence and auxiliary diagnostic strategies for the early identification, disease assessment, and intervention of COPD with CMH.

Materials and Methods

Study Population

This research is cross-sectional in nature. We collected information from 319 COPD patients who received care at Changzhi People's Hospital between May 2020 and March 2024. Clinical information was collected for each patient, including general demographic data, admission details, clinical symptoms, pulmonary function test results, laboratory tests, diagnoses, and treatments. All data were collected at a single time point prior to the patients' discharge. Based on sputum volume and characteristics, patients were split into two groups: the high mucus secretion group (n=173), which was defined as having daily sputum volume≥30 mL or sputum that was purulent or mucopurulent, and cough and sputum production symptoms that persisted for more than three months on most days for at least two consecutive years;[10] Non-CMH (n=146): daily sputum volume<30 mL and sputum characterized as mucous or no sputum. Patients were randomly divided into two groups, with 80% serving as the training set and 20% as the test set, for model development and internal validation.

The inclusion criteria are as follows: (1) Meet the gold standard for the diagnosis of COPD;[11] (2) Patients with stable COPD are those whose symptoms have not gotten worse over a four-week period; the post-treatment condition is when the patient's respiratory symptoms have improved after being admitted to the hospital, enabling the patient to stop receiving treatment and be released;[12] (3) Age over 18 years; Exclusion criteria are as follows: (1) Patients with other respiratory diseases that may affect airway mucus secretion, such as asthma or bronchiectasis, are excluded; (2) Patients with severe complications involving the heart, lungs, liver, kidneys, or other organs are excluded; (3) Patients who are receiving treatment that may affect the research results (such as immunosuppressants or specific antibiotics) or have used antibiotics within the 4 weeks prior to sputum collection are excluded; (4) Patients with missing data are excluded. This study has been reviewed and approved by the Medical Ethics Committee of Changzhi People's Hospital (Approval No.: KYYJ-2024-079), and all operations followed the ethical principles of the Helsinki Declaration. All enrolled patients signed informed consent forms and voluntarily participated in this study.

Sample Size Estimation

The sample size was decided based on how many eligible patients were available during the study period (May 2020 to March 2024) and by looking at similar machine learning studies on COPD. Wang et al[13] created an AECOPD prediction model with similar features using a comparable number of participants. Since this study was more exploratory, we aimed to include at least 300 patients to make sure both CMH and non-CMH types were well represented. In the end, we enrolled 319 patients, with 173 CMH cases and 146 non-CMH cases.

Data Collection

Quantification of Clinical Symptoms

The Chronic Obstructive Pulmonary Disease Assessment Test (CAT) in Chinese,[14] the internationally recognized Modified Medical Research Council Dyspnea Scale (mMRC),[15] and the Breathing, Cough, and Sputum Scale (BCSS)[16] were used to systematically assess patients'clinical symptom burden, dyspnea severity, and severity of respiratory symptoms, respectively.

Pulmonary Function Testing

Forced vital capacity (FVC), forced expiratory volume in one second (FEV_1), the FEV_1/FVC ratio, and the percentage of FEV_1 relative to predicted values were measured for every patient using the Master Screen TM PFT series pulmonary function testing system from Vyaire Medical (Germany), closely following the most recent standard operating guidelines jointly published by the American Thoracic Society (ATS) and the European Respiratory Society (ERS). At least three trials of each test were conducted, and the curve data with the highest total of FEV_2 and FVC was chosen for recording and analysis.

HRCT Scanning and Image Analysis

After patients underwent standard breathing training, a chest HRCT examination was performed using a Siemens SOMATOM Definition AS+ 128-slice multislice spiral CT scanner. The scanning range extended from the lung apex to the diaphragm level, covering the entire lung field. The technical parameters were as follows: tube voltage 120 kV; tube current automatically modulated using Care Dose 4D technology to optimize radiation dose; slice thickness 1 mm; and reconstruction interval 0.7 mm. After acquiring the raw images, they were imported into Philips IntelliSpace Portal 12 for analysis to obtain data on pulmonary emphysema, bronchial wall thickening, and pulmonary bullae.

16S rRNA Gene Sequencing Analysis

A total of 120 patients were randomly selected from the study population (72 from the CMH group and 48 from the non-CMH group) for 16S rRNA sequencing analysis. To minimize contamination from oral bacteria, sputum plugs were separated from saliva prior to analysis and stored at −80°C until DNA extraction. In accordance with the manufacturer's instructions, sputum plugs were homogenized using a solution containing 0.1% dithiothreitol (DTT), and then bacterial genomic DNA was extracted using a lysozyme-based lysis process (Qiagen DNA Mini Kit, Qiagen, CA, USA).

The V4 hypervariable region of the 16S rRNA gene was the focus of an amplicon library. Primers 515F (5′-GTGCCAGCMGCCCGGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) (28 cycles) were used for PCR amplification. The forward primer was connected to a 12-bp Golay barcode sequence and Illumina sequencing adapters. The Illumina HiSeq 2500 platform was used for bidirectional sequencing with a read length of 250 bp. In addition to DNA extraction and PCR negative controls to check for reagent contamination, each sequencing run contained commercial simulated community DNA (ZymoBIOMICS microbial DNA standards) as a positive control.

Sequencing data were processed using the QIIME v1.9.1 workflow. First, data were quality-filtered after removing non-Illumina adapter sequences (excluding reads <20 bp in length). Fastq-join was used to merge paired-end sequences with a minimum overlap of 10 bp. Phred quality scores (≥20) were used to further filter the sequences, and UCHIME was used to eliminate chimeras. Based on 97% sequence similarity, the Greengenes database (v13_8) and the Ribosomal Database Project classifier were used to identify operational taxonomic units (OTUs) and annotate species. Based on the sparsity curve and the maximum sequencing depth of unsuccessful samples, all COPD samples were standardized to a read depth of 31,750. To analyze OTU richness and evenness among samples, alpha diversity indices were computed;

main coordinate analysis plots were used to display β-diversity, which was assessed using the Bray-Curtis distance matrix. Additionally, the ratio of the Gamma-proteobacteria to Firmicutes (GP: F) and the ratio of Firmicutes to Bacteroidetes (F: B) were calculated as univariate microbiome parameters for association analysis with clinical indicators. Microbiome ratios were obtained by applying a log2 transformation to the proportion of corresponding microbial sequences in the samples.

Machine Learning Data Processing

From Clinical Information and the previously mentioned airway microbiome data, we extracted a total of 70 input features and 1 output feature (confirmed clinical diagnosis of CMH). We then built six machine learning models: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests (RF), Bayesian Models (BN), Generalized Bayesian Decision Trees (GBDT), and Neural Networks (NN). We evaluated and compared the models using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and area under the ROC curve (AUC). Before building the classification models, we transformed the categorical features using one-hot encoding or label encoding. For handling missing data, we applied the following imputation strategies:

1) The most common category of a categorical feature is used to replace any missing data;
2) The mean value of the missing data is substituted if it is a continuous characteristic. All features were then standardized to take inter-subject variability and feature differences into consideration.

Feature Selection

Filtering, enveloping, and embedding are the three general categories into which feature selection techniques fall.[17] However, only filtering makes it possible to carry out the feature selection procedure without using any machine learning methods, which is essential for preventing potential biases that can occur when comparing various techniques. The Kolmogorov–Smirnov test was utilized in this study to evaluate the data's normality, and feature selection was carried out using the filtering method (using the chi-square test or Fisher's exact test for categorical features, along with information gain, and the *t*-test or Wilcoxon rank-sum test for continuous features). Only features with a positive information gain score and those shown to be statistically significant by hypothesis testing were included.

Model Development

The machine learning methods utilize the following R packages: caret, ipred, ranger, arm, nnet, and gbm. All models employ 10-fold cross-validation. Hyperparameters were tuned via grid search as follows:

We used "max_depth" and "n_estimators" with ranges of 1 to 20 and a step size of 1, "min_samples_split" with a range of 2 to 20 and a step size of 1, and "min_samples_leaf" with a range of 1 to 30 and a step size of 1 for the RF model.

Three kernel functions—linear, polynomial, and radial basis function—were chosen for SVM testing. SVM's C and γ are two more crucial hyperparameters. In our investigation, γ was chosen from 0 to 30 with a step size of 0.1 and C from 0 to 200 with a step size of 0.01.

For KNN, we selected eight distance metrics for testing: Minkowski, cosine, Euclidean, and Spearman.[18] For the key KNN parameter k, we selected odd numbers ranging from 1 to 21.

For the BN model, we selected the K2 algorithm for network structure learning, with the maximum number of parent nodes set to 3–5. Parameter estimation employed maximum likelihood estimation (MLE) for learning the conditional probability table Bayesian optimization was used for adaptive tuning of the BN's hyperparameters, including the search space for network structure and the smoothing factor for parameter estimation, with the optimal configuration selected via 10-fold cross-validation.

For the GBDT model, we chose "n_estimators" between 50 and 500 with a step size of 50, "learning_rate" between 0.01 and 0.3 with a step size of 0.01, "max_depth" between 3 and 10 with a step size of 1, "subsample" between 0.5 and 1.0 with a step size of 0.1, "min_samples_split" between 2 and 20 with a step size of 1, and "min_samples_leaf" between

1 and 30 with a step size of 1. When the validation set performance did not increase for ten rounds in a row, an early stopping technique was used to end training early.

We chose one to three hidden layers for the NN model, each with 32–256 neurons in increments of 32. The output layer produced binary classification probabilities using the sigmoid activation function, whereas the hidden layers employed the ReLU activation function. The Adam optimizer was used, and the starting learning rate was set in increments of 0.0001 between 0.0001 and 0.01. The L2 regularization coefficient has a step size of 0.0001 and a range of 0.0001 to 0.1. The maximum number of iterations is set to 300, and the batch size is set to 32, 64, or 128. When performance on the validation set does not improve for 20 consecutive epochs, an early stopping technique is used to end training early.

Apart from the hyperparameters mentioned above, all other hyperparameters for each model are set to their default values.

Statistical Analysis

R 4.2.3 and SPSS 26.0 were used for all statistical analyses. Categorical variables were expressed as frequency (percentage) [n (%)], and comparisons between groups were carried out using the chi-square (χ^2) test or Friedman test; continuous variables were described and compared according to their distribution characteristics: those with a normal distribution were expressed as mean±standard deviation ($\bar{x} \pm s$), and comparisons between groups were carried out using the independent samples *t*-test or analysis of variance (ANOVA). Statistical significance was defined as a *P*-value of less than 0.05.

Results

Patient Demographics

Following manual review, a total of 319 clinical records that met the inclusion criteria were ultimately identified; the remaining records were excluded due to the exclusion criteria. The CMH group and the non-CMH group comprised 173 and 146 patients, respectively. The mean age of patients in the CMH group was (69.3±8.7) years, while that of the non-CMH group was (68.1±9.2) years. Thirty-six input features were extracted from each Clinical Information, including 7 basic patient information features, 4 vital signs, 4 pulmonary function indicators, 3 comorbidities, and 18 symptoms. The results in Table 1 show that there were significant differences among 16 of these features.

Characteristics of the Airway Microbiome

The observed ASVs, Shannon index, and Faith's phylogenetic diversity index were all considerably lower in the CMH group than in the non-CMH group, according to alpha diversity analysis (all $P < 0.001$; Figure 1A–C). The microbial communities of the two groups grouped together and were easily distinguished, according to PCoA analysis based on the Bray-Curtis distance matrix (Figure 1D). This shows that the microbial communities of the CMH and non-CMH groups differ significantly, with the CMH group having lower microbial diversity. Firmicutes, Proteobacteria, and Bacteroidetes were the predominant phyla at the phylum level in both the CMH and non-CMH groups. The non-CMH group had a larger relative abundance of Firmicutes than the CMH group, whereas Proteobacteria showed the opposite trend (Figure 1F). Thirty-five significant genera were found in both groupings at the genus level. Pseudomonas, Moraxella, and Haemophilus had greater relative abundances in the CMH group than in the non-CMH group. Prevotella, Veillonella, and Streptococcus were more prevalent in the non-CMH group (Figure 1G). The main taxa with varying abundance are displayed in Figure 1E.

Linear discriminant analysis of effects (LEfSe) revealed 33 distinguishing genera between the CMH and non-CMH groups. In the CMH group, genera such as Haemophilus, Pseudomonas, Veillonella, Moraxella, Streptococcus, Klebsiella, and Neisseria were significantly enriched, indicating that a state of high mucus secretion is closely associated with increased colonization by various potentially pathogenic bacteria. In the non-CMH group, commensal genera such as Prevotella, Lactobacillus, Fusobacterium, Porphyromonas, Leptotrichia, Actinobacillus, and Bacteroides were dominant (Figure 2).

Table 1 Baseline Characteristics of Patients

Baseline Indicators	CMH (n=173)	Non-CMH (n=146)	t/χ^2	P
Demographic Information				
Age (years, $\bar{x} \pm s$)	69.3±8.7	68.1±9.2	t = 1.19	0.232
Gender [Male, n (%)]	131 (75.7)	105 (71.9)	χ^2 = 0.59	0.440
BMI (kg/m^2, $\bar{x} \pm s$)	23.8±4.1	24.2±4.3	t = −0.84	0.397
Course of the disease (years, $\bar{x} \pm s$)	10.2±5.3	8.7±4.8	t = 2.63	0.009
Current smokers [n (%)]	107 (61.8)	86 (58.9)	χ^2 = 0.29	0.592
Current drinkers [n (%)]	65 (37.6)	53 (36.3)	χ^2 = 0.05	0.815
History of AECOPD[n (%)]	118 (68.2)	79 (54.1)	χ^2 = 6.66	0.010
Vital signs				
Heart rate (beats per minute, $\bar{x} \pm s$)	89.4±12.1	86.2±11.3	t = 2.43	0.016
Diastolic blood pressure (mmHg, $\bar{x} \pm s$)	78.3±10.7	79.0±10.5	t = −0.59	0.558
Systolic blood pressure (mmHg, $\bar{x} \pm s$)	132.6±18.4	133.1±17.9	t = −0.24	0.807
pH	7.38±0.05	7.40±0.04	t = −3.89	<0.001
Lung function parameters				
FEV_1 (L, $\bar{x} \pm s$)	1.24±0.51	1.38±0.55	t = −2.36	0.019
FVC (L, $\bar{x} \pm s$)	2.41±0.78	2.56±0.81	t = −1.68	0.094
FEV_1/FVC ($\bar{x} \pm s$)	51.4±11.2	53.9±11.8	t = −1.94	0.054
Predicted FEV_1 (%, $\bar{x} \pm s$)	49.2±15.3	52.8±16.1	t = −2.04	0.042
Complications [n (%)]				
High blood pressure	85 (49.1)	70 (47.9)	χ^2 = 0.04	0.833
Diabetes	39 (22.5)	32 (21.9)	χ^2 = 0.02	0.894
Fatty liver	26 (15.0)	21 (14.4)	χ^2 = 0.03	0.871
Clinical symptoms [n (%)]				
BCSS score (points, $\bar{x} \pm s$)	6.8±2.1	5.2±1.9	t = 7.08	<0.001
CAT score (points, $\bar{x} \pm s$)	21.5±6.7	17.3±6.9	t = 15.50	<0.001
Clinical Symptom Burden [n (%)]			χ^2 = 19.48	<0.001
Mild (CAT < 10)	12 (6.9)	29 (19.9)		
Moderate (CAT 10–20)	69 (39.9)	70 (47.9)		
Severe (CAT > 20)	92 (53.2)	47 (32.2)		
mMRC Grade [n (%)]			χ^2 = 6.32	0.042
Grade 0-1	42 (24.3)	53 (36.3)		
Grade 2	75 (43.4)	59 (40.4)		
Grade 3-4	56 (32.4)	34 (23.3)		
CT Imaging Features				
Emphysema [n (%)]	122 (70.5)	92 (63.0)	χ^2 = 2.02	0.155
Bronchial wall thickening [n (%)]	98 (56.6)	73 (50.0)	χ^2 = 1.41	0.236
Pulmonary bullae [n (%)]	64 (37.0)	48 (32.9)	χ^2 = 0.59	0.443
Mosaic perfusion [n (%)]	40 (23.1)	28 (19.2)	χ^2 = 0.73	0.392
Total lung volume (L, $\bar{x} \pm s$)	5.42±1.18	5.21±1.23	t = 1.55	0.121
Mean lung density (HU, $\bar{x} \pm s$)	−856.4±38.7	−843.2±36.5	t = 3.11	0.002
Percentage of airway wall area (%, $\bar{x} \pm s$)	68.4±7.2	66.1±7.5	t = 2.79	0.005
LAA% (< −950 HU, $\bar{x} \pm s$)	24.7±12.3	21.3±11.9	t = 2.50	0.013
Tree-bud sign [n (%)]	22 (12.7)	15 (10.3)	χ^2 = 0.46	0.497
Pulmonary parenchymal infiltration [n (%)]	19 (11.0)	13 (8.9)	χ^2 = 0.38	0.538
Pleural effusion [n (%)]	14 (8.1)	9 (6.2)	χ^2 = 0.44	0.507
Arterial Blood Gas Analysis				
PaO_2 (mmHg, $\bar{x} \pm s$)	68.3±11.4	71.2±10.8	t = −2.32	0.021
$PaCO_2$ (mmHg, $\bar{x} \pm s$)	44.7±7.6	42.1±6.9	t = 3.17	0.002

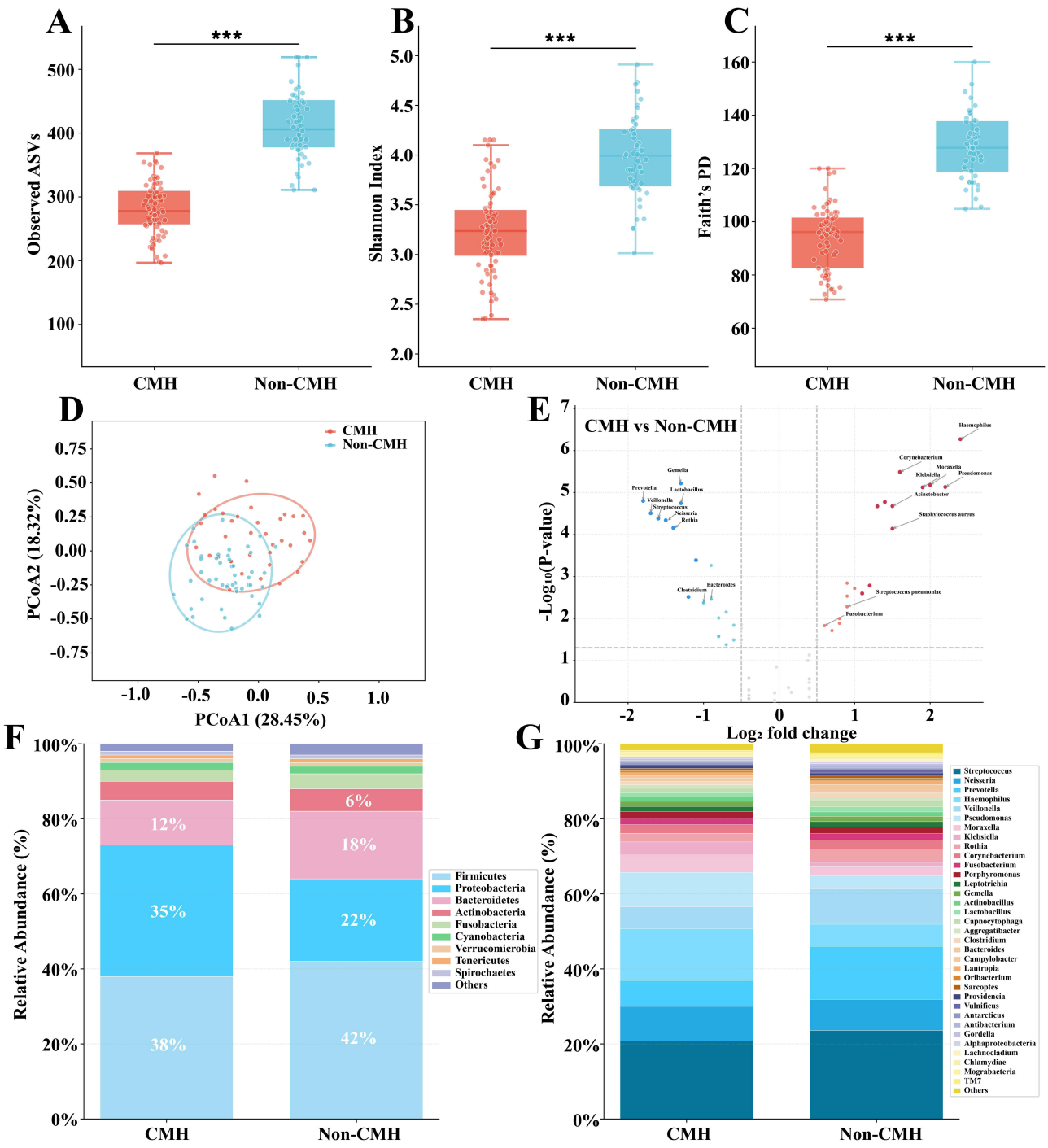

Figure 1 Shows the respiratory microbiome's diversity and location in patients with and without CMH.
Notes: ***P<0.001. The Shannon index, Faith PD index, and OTU count between the two groups (**A**–**C**). Bray Curtis-based PCoA plots (**D**). Different taxonomic unit visualizations in the non-CMH group and the CMH group (**E**). The x axis displays the fold change, whereas the y axis displays -log10 P (FDR adjusted). A sample is represented by each dot in the diagram. The red dots represent $p < 0.05$, | logFC |> 1; Green dots represent $p > 0.05$, | logFC |> 1; Gray dot represents $p > 0.05$, | logFC |< 1. The black boxes represent the taxa that belong to top 35 Histogram of the relative abundance of the top 10 phyla (**F**) and the top 35 genera (**G**). genera.

Feature Selection

The results of feature selection, which combined information gain and hypothesis testing, showed that disease duration was the most important feature, followed by Haemophilus, history of AECOPD, Pseudomonas, and FEV1% predicted.

https://doi.org/10.2147/COPD.S620360

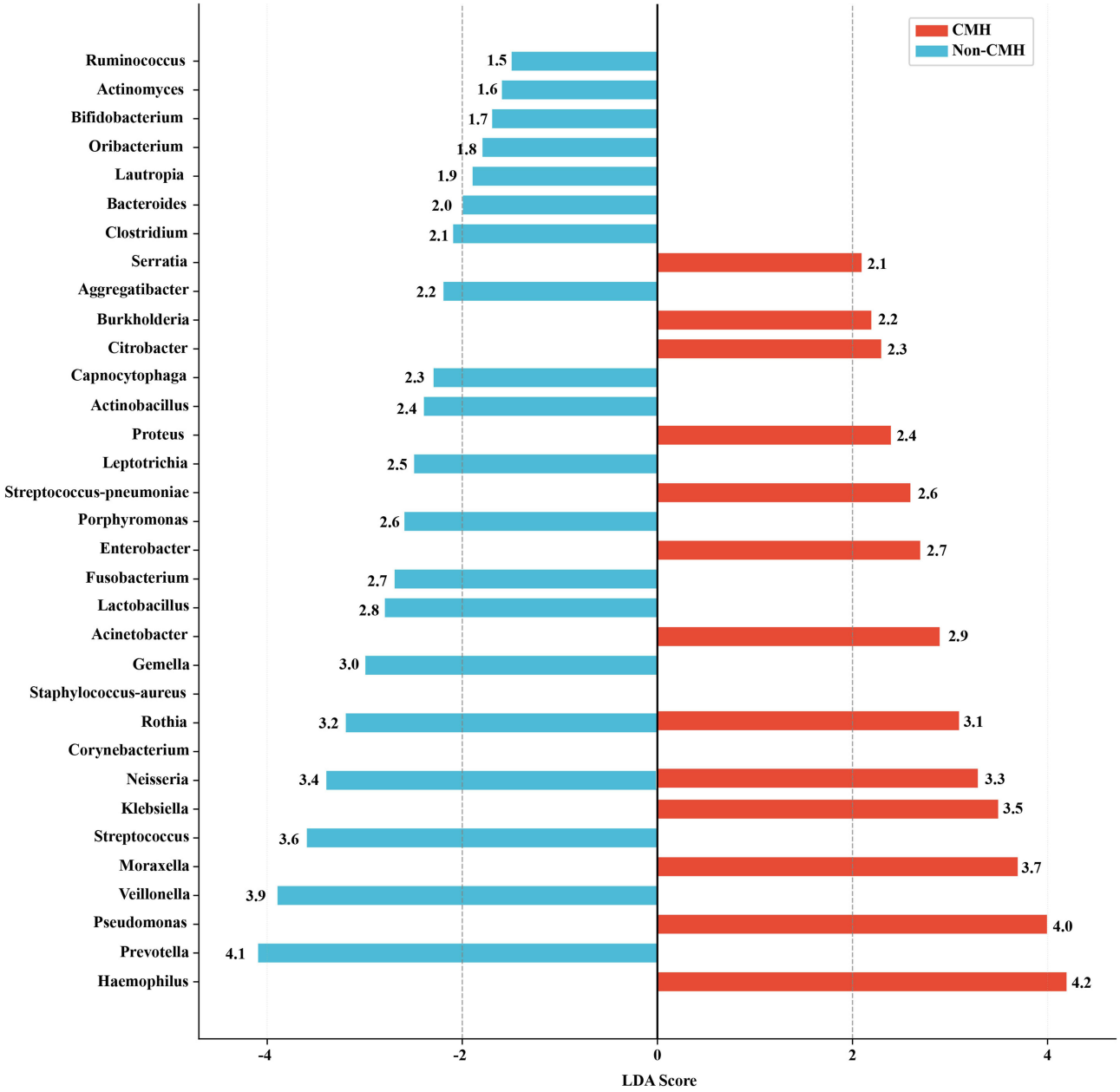

Figure 2 LEfSe analysis and LDA histogram of the airway microbiome in the CMH group and the non-CMH group.

The remaining features were excluded because their information gain was zero or they did not demonstrate statistical significance (Figure 3).

Model Performance

Table 2 and Table 3 summarize the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC obtained by each recognition algorithm under feature selection cross-validation. The results show that the SVM algorithm performed best on the training set, but its AUC dropped to 0.483±0.027 on the test set. The BN algorithm achieved an AUC of 0.917±0.013 on the training set, with sensitivity, specificity, and other metrics all exceeding 0.8. It also maintained the best generalization ability on the test set, with an AUC of 0.911±0.014. It outperformed the other five algorithms in terms of sensitivity, specificity, PPV, and NPV, demonstrating good stability.

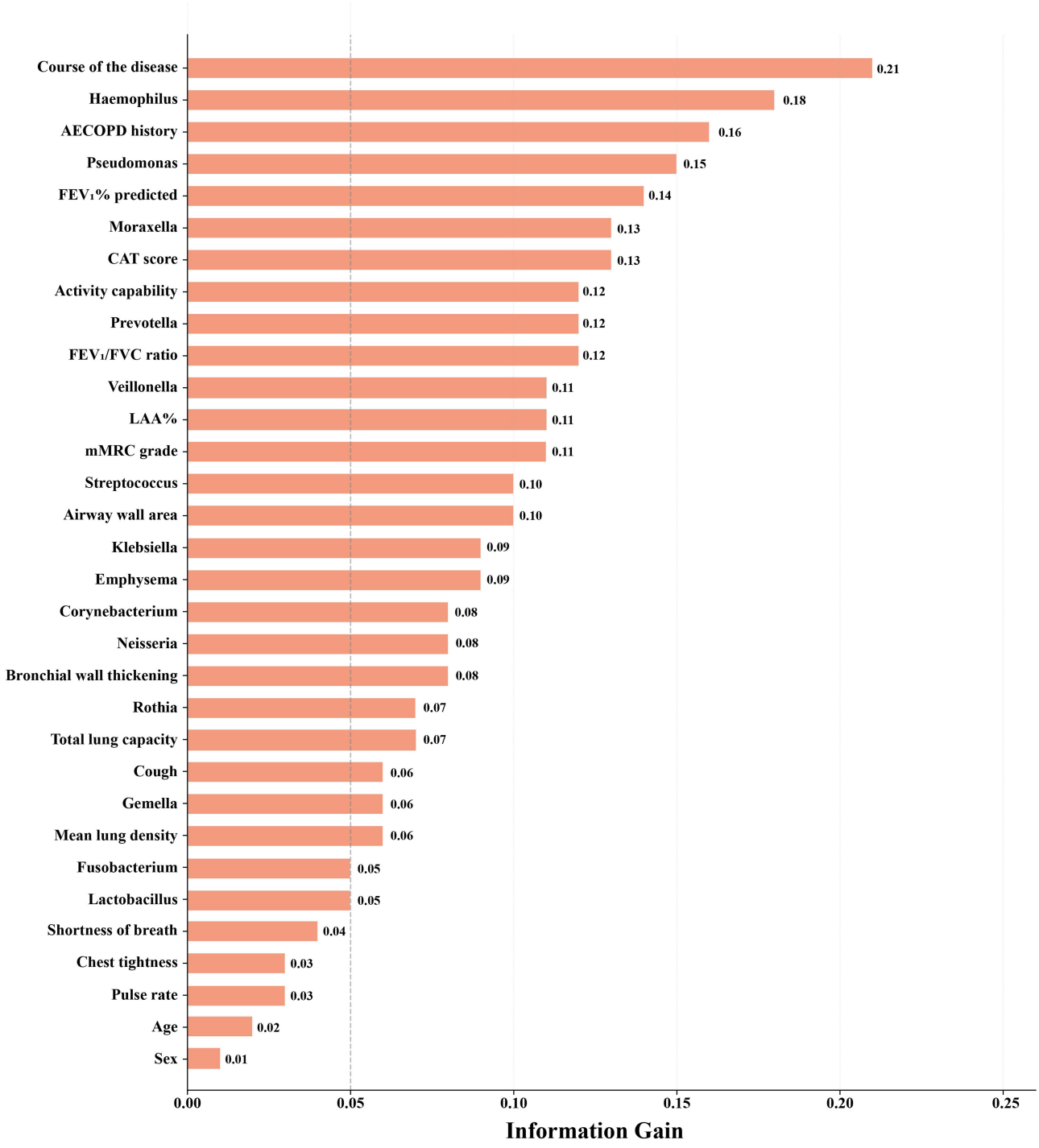

Figure 3 Information gain ranking.

Discussion

This study investigated the development of a machine learning-based predictive model for airway CMH in COPD and its associated microbiome research. The results showed that the CMH group had a longer disease duration, a higher proportion of AECOPD, and poorer lung function compared to the non-CMH group, indicating that CMH is an important marker of COPD severity. These findings are consistent with the study by Kim et al,[19] which reported a significantly higher incidence of mucus hypersecretion in COPD patients with a disease duration of ≥10 years. Differences were also observed in the composition of the airway microbiome between the CMH and non-CMH groups. Our data indicate

https://doi.org/10.2147/COPD.S620360

Table 2 Comparison of Model Performance on Training Set

	Sensitivity	Specificity	PPV	NPV	AUC
SVM	0.970±0.012	0.917±0.015	0.946±0.014	0.954±0.011	0.988±0.008
KNN	0.746±0.021	0.625±0.024	0.676 ±0.022	0.701±0.020	0.736±0.024
RF	0.852±0.020	0.773±0.018	0.831±0.017	0.819±0.019	0.825±0.016
BN	0.850±0.018	0.842±0.016	0.811±0.019	0.876±0.015	0.917±0.013
GBDT	0.628±0.023	0.764±0.020	0.533±0.025	0.827±0.018	0.762±0.019
NN	0.625±0.026	0.561±0.027	0.587±0.025	0.601±0.024	0.593±0.031

Table 3 Comparison of Model Performance on Test Set

	Sensitivity	Specificity	PPV	NPV	AUC
SVM	0.687±0.025	0.303±0.028	0.496±0.026	0.492±0.024	0.483±0.027
KNN	0.670±0.023	0.654±0.022	0.641±0.024	0.682±0.021	0.729±0.022
RF	0.827±0.020	0.764±0.019	0.749±0.022	0.736±0.023	0.758±0.021
BN	0.867±0.016	0.789±0.018	0.805±0.017	0.855±0.015	0.911±0.014
GBDT	0.650±0.024	0.690±0.022	0.474±0.026	0.822±0.019	0.711±0.020
NN	0.598±0.027	0.550±0.026	0.570±0.025	0.579±0.024	0.574±0.028

reduced microbial diversity in the CMH group. In the CMH group, the relative abundance of Proteobacteria was higher, whereas the non-CMH group was dominated by Firmicutes, with a more balanced and diverse microbial community. Potential pathogens such as Haemophilus, Pseudomonas, and Moraxella were enriched in the CMH group, and these genera were identified as predictors of CMH. Previous studies have shown that COPD patients have higher Moraxella concentrations and a significantly increased imbalance in sputum microbiota compared to healthy individuals.[20] Furthermore, bacterial diversity continues to decline in advanced-stage patients, who exhibit higher mortality rates.[21] Therefore, microbial community diversity is crucial for maintaining respiratory health and plays a vital role in resisting colonization, maintaining epithelial integrity, and immune regulation.[22,23]

The diversity of the airway microbiome in the CMH group was lower than in the non-CMH group, which may be closely associated with the enrichment of potential pathogens in this group, such as Haemophilus, Pseudomonas, and Moraxella. Excessive mucus production and increased viscosity in the airways under CMH conditions create a selective microenvironment conducive to the colonization of specific bacteria. In a neutrophil-dominated airway inflammatory environment, the Proteobacteria phylum and Haemophilus genus predominate. Excessive neutrophil elastase impairs mucociliary clearance, induces the formation of Neutrophil Extracellular Traps (NETs), and promotes excessive mucus production.[24] NETs in sputum are closely associated with reduced microbial diversity and increased abundance of Haemophilus spp.[25] Haemophilus spp. can form biofilms in the airways to resist attacks from the host immune response, thereby establishing long-term colonization within the airways. This process further induces NET formation and exacerbates neutrophil-mediated inflammation, leading to the release of serum peroxidase, catalase, matrix metalloproteinase 8, matrix metalloproteinase 9, and neutrophil elastase in CMH patients.[26,27] Furthermore, the CMH and non-CMH states may correspond to different inflammatory phenotypes, leading to differences in the airway microbiome between the two groups. Findings by Jia et al[28] indicate that different phyla of the respiratory tract prokaryotic microbiota are associated with innate immune, metabolic, and inflammatory factors related to COPD. In eosinophilic inflammation, anaerobic commensals suppress inflammation by producing short-chain fatty acids; however, in moderate eosinophilic inflammation, these commensals may be replaced by pathogenic bacteria such as the Proteobacteria phylum and Haemophilus spp., triggering severe neutrophilic inflammation and exacerbating mucus plugging.[29] The pattern of Proteobacteria enrichment and Firmicutes depletion observed in the CMH group in this study suggests that the CMH state is dominated by neutrophilic inflammation, whereas the non-CMH group is characterized by a relatively balanced inflammatory state.

Based on the above baseline metrics and microbiome data, we evaluated six machine learning models for identifying excessive mucus secretion in COPD patients. The results showed that after feature selection using filtering methods, the performance of all five algorithms—excluding the NN—exceeded that of the baseline model. Among these five models, the BN demonstrated the best CMH identification capability, followed by the RF. This may be attributed to the BN model's superior interpretability, suitability for medical data, and ability to handle missing values.[30] In our study, the SVM model performed best on the internal test set but underperformed on the external test set, likely due to excessive model complexity. CMH identification models with high sensitivity and specificity can effectively provide decision support to clinicians, thereby helping to avoid misclassification of non-CMH cases and overestimation of CMH cases to some extent. Therefore, we believe that BN and RF hold promise in the development of CMH identification models. Most current studies focus on machine learning for the early prediction of AECOPD or the differentiation between COPD and asthma. For example, Wang et al[13] developed an AECOPD recognition model using five machine learning algorithms. After feature selection, they found that the SVM model achieved the best performance, with a sensitivity of 0.80, a specificity of 0.83, a positive predictive value of 0.81, a negative predictive value of 0.85, and an area under the receiver operating characteristic curve of 0.90. Fernandez et al[31] developed a predictive model using a probabilistic neural network that could predict the early onset of COPD exacerbation 4.8±1.8 days in advance, with a detection accuracy of 80.5%; however, the study enrolled only 16 patients, resulting in a relatively small sample size. Kocks et al[32] developed an asthma/COPD differential diagnosis tool based on AC/DC machine learning, which demonstrated higher average diagnostic accuracy than primary care physicians.

In our study, we constructed feature subsets by combining baseline indicators with microbiome data. Among these, the duration of COPD—which had an information gain of 0.21—proved to be a powerful indicator for identifying CMH. Therefore, in clinical practice, physicians should pay closer attention to the duration of COPD. Haemophilus yielded an information gain of 0.18, indicating that COPD patients with an abundance of Haemophilus in their airway microbiota are more prone to developing CMH than other patients. Additionally, a history of acute COPD exacerbations, Pseudomonas, and predicted FEV1% should also be identified as important indicators of CMH. For certain characteristics—such as current smokers, current drinkers, hypertension, and diabetes—the statistical significance or information gain was below zero. On the one hand, this may be because these characteristics genuinely lack the ability to identify CMH. However, uneven data quality may also be a significant factor contributing to this result. Although we employed imputation strategies to address this, such methods inevitably altered the original distribution and, to some extent, compromised the objectivity and validity of the data. Therefore, the relevant results do not accurately reflect the actual situation.

The feature screening results of this study showed that the abundance of Haemophilus and Pseudomonas are important predictive factors for CMH, suggesting that these two bacterial genera may play a role in CMH related airway microecological disorders. Based on this, we propose the following two-stage simplified screening strategy: In the first stage, routine clinical indicators such as disease duration, AECOPD history, FEV1%, etc are used for rapid risk assessment, which can be obtained immediately in daily outpatient diagnosis and treatment; In the second stage, sputum or throat swab samples are collected from patients who are initially screened positive or highly suspected clinically. qPCR technology is used to detect the abundance of Haemophilus and Pseudomonas, and a simplified scoring system is constructed based on the three clinical variables mentioned above. The scoring system can use logistic regression coefficient scoring method to correspond the total score with CMH risk probability, thereby achieving accurate identification of high-risk patients without relying on high-throughput sequencing. Previous studies have provided feasible evidence for this transformation pathway. Through longitudinal analysis of the AERIS cohort, Mayhew et al[33] found a significant positive correlation between changes in the abundance of *Haemophilus influenzae* in sputum and the risk of acute exacerbation of COPD. Monitoring based on this single bacterial genus can partially predict clinical deterioration events. In addition, Wang et al[13] screened three key bacterial genera, namely *Haemophilus influenzae*, *Moraxella catarrhalis*, and *Streptococcus pneumoniae*, based on sputum microbiome data, and constructed a risk scoring system for acute exacerbation of COPD in combination with clinical symptoms. The area under the working characteristic curve of the subjects reached 0.85. These studies indicate that replacing whole microbiome sequencing with a few key bacterial genera can significantly reduce technical barriers and detection costs while maintaining a certain level of

predictive accuracy. Furthermore, future research could attempt to develop rapid antigen detection test strips or isothermal amplification kits for Haemophilus and Pseudomonas, reducing detection time to less than 30 minutes and making them truly suitable as instant diagnostic tools for primary healthcare institutions. At the same time, the simplified scoring system can be embedded into an electronic medical record system or WeChat mini program, which automatically outputs the CMH risk probability after inputting clinical variables and/or qPCR test results, making it easier for clinical doctors to make quick decisions.

There are a number of limitations to this study. First, our findings only apply to particular patients who match the inclusion and exclusion criteria; they do not represent the overall community of COPD patients with or without CMH. For instance, the suggested model's capacity to discriminate will be constrained if a patient also has asthma. Second, even though our data covers a substantial time span from 2020 to 2024, it comes from a single location, which would restrict how broadly the suggested approach can be applied. To confirm the model's generalizability, future research should use sizable, multicenter external validation datasets. Third, the performance of algorithms other than the six employed in this work was not compared by the authors. To assess the application of various machine learning methods for CMH detection, more investigation is required. The sample size of this study is relatively small. Despite using feature selection, the relatively small sample size may still limit the model's generalization ability and stability. Readers should interpret the performance of the model with caution. At the same time, low abundance microbial groups may also be underestimated in microbiome analysis. Future research should broaden the model's application by adding more thorough training data. To train the identification model and perform external validation, larger datasets from several sites are required.

Conclusions

Our study reveals that patients with and without CMH in COPD differ significantly in baseline indicators, including demographic information, pulmonary function parameters, and clinical symptoms. The observed changes in microbiome composition—characterized by reduced microbial diversity and increased abundance of potentially pathogenic bacteria—are associated with the presence of CMH, although the cross-sectional design precludes determination of causal direction. BN and RF models can effectively identify patients with CMH based on these features. Furthermore, we have provided a robust feature set for identifying CMH using machine learning methods. This study is expected to offer decision support when clinicians face difficulties in making a definitive diagnosis. Future work will focus on developing more comprehensive inclusion criteria, analyzing larger datasets, and employing longitudinal designs to clarify the temporal and causal relationships between CMH and airway microbiome alterations.

Data Sharing Statement

The datasets used and analysed during the current study available from the corresponding author Professor Hui Zhao on reasonable request.

Ethics Statement

This study has been reviewed and approved by the Medical Ethics Committee of Changzhi People's Hospital (Approval No.: KYYJ-2024-079), and all operations followed the ethical principles of the Helsinki Declaration. All enrolled patients signed informed consent forms and voluntarily participated in this study.

Author Contributions

Hui Zhao should be considered the first corresponding author and Qingqing Liu should be considered the second corresponding author.

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by the Shanxi Province Higher Education "10 Billion Project" under the grant titled "Establishment and Evaluation of an Integrated Cohort for COPD Comorbidities and Related Immunological Mechanisms" (Grant No: BY-ZB-2024007).

Disclosure

The authors report support for the paper from Shanxi Province Higher Education "10 Billion Project" and declared that they have no other conflicts of interest regarding this work.

References

1. Wang Q, Su W, Liu J, Zhao D. Advances in the investigation of the role of autophagy in the etiology of chronic obstructive pulmonary disease: a review. *Medicine*. 2023;102(47):e36390. doi:10.1097/MD.0000000000036390
2. Al Wachami N, Guennouni M, Iderdar Y, et al. Estimating the global prevalence of chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. *BMC Public Health*. 2024;24(1):297. doi:10.1186/s12889-024-17686-9
3. Ramos FL, Krahnke JS, Kim V. Clinical issues of mucus accumulation in COPD. *Int J Chronic Obstr*. 2014;9:139–150. doi:10.2147/COPD.S38938
4. Aghapour M, Ubags ND, Bruder D, et al. Role of air pollutants in airway epithelial barrier dysfunction in asthma and COPD. *Eur Resp Rev*. 2022;31(163):210112. doi:10.1183/16000617.0112-2021
5. Merino M, Villoro R, Hidalgo-Vega Á, Carmona C. Social economic costs of COPD in Extremadura (Spain): an observational study. *Int J Chronic Obstr*. 2018;13:2501–2514. doi:10.2147/COPD.S167357
6. Abulizi A, Zhou J, Abudukelimu N, Yehaiya G, Abudukelimu M, Abudukelimu H. A review of artificial intelligence-based research on chronic obstructive pulmonary disease. *Respir Med*. 2026;255:108778. doi:10.1016/j.rmed.2026.108778
7. Li S, Xu Y, Wu Y, Zhao Z. 16S rRNA sequencing-based analysis of sputum microbiome in patients with acute exacerbations of chronic obstructive pulmonary disease: retrospective cohort study. *J Thoracic Dis*. 2025;17(10):8876–8886. doi:10.21037/jtd-2025-995
8. Nielsen R, Xue Y, Jonassen I, et al. Repeated bronchoscopy in health and obstructive lung disease: is the airway microbiome stable? *BMC Pulm Med*. 2021;21(1):342. doi:10.1186/s12890-021-01687-0
9. Cai S, Gao J, Liu X, et al. Seasonal dynamics of the upper respiratory tract microbiome in chronic obstructive pulmonary disease. *Int J Chronic Obstr*. 2023;18:1267–1276. doi:10.2147/COPD.S403198
10. Shen Y, Huang S, Kang J, et al. Management of airway mucus hypersecretion in chronic airway inflammatory disease: chinese expert consensus (English edition). *Int J Chronic Obstr*. 2018;13:399–407. doi:10.2147/COPD.S144312
11. Du D, He S, Qin J, et al. Diagnostic performance of FEV(1)/FEV(6) in detecting airway obstruction and chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Chron Respir Dis*. 2026;23:14799731261431251. doi:10.1177/14799731261431251
12. Vogelmeier CF, Criner GJ, Martínez FJ, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report: GOLD executive summary. *Arch Bronconeumol*. 2017;53(3):128–149. doi:10.1016/j.arbres.2017.02.001
13. Wang C, Chen X, Du L, Zhan Q, Yang T, Fang Z. Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease. *Comput Meth Programs Biomed*. 2020;188:105267. doi:10.1016/j.cmpb.2019.105267
14. Chen X, Du L, Xiao W, et al. Validation and application of the simplified Chinese version of the dyspnea-12 questionnaire: an independent predictor for hospitalization risk in chronic obstructive pulmonary disease. *Int J Chronic Obstr*. 2025;20:3559–3574. doi:10.2147/COPD.S547507
15. Paladini L, Hodder R, Cecchini I, Bellia V, Incalzi RA. The MRC dyspnoea scale by telephone interview to monitor health status in elderly COPD patients. *Respir Med*. 2010;104(7):1027–1034. doi:10.1016/j.rmed.2009.12.012
16. Leidy NK, Schmier JK, Jones MK, Lloyd J, Rocchiccioli K. Evaluating symptoms in chronic obstructive pulmonary disease: validation of the breathlessness, cough and sputum scale. *Respir Med*. 2003; 97 Suppl A: S59–70.
17. Rodriguez-Galiano VF, Luque-Espinar JA, Chica-Olmo M, Mendes MP. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *Sci Total Environ*. 2018;624:661–672. doi:10.1016/j.scitotenv.2017.12.152
18. Wan KR, Maszczyk T, See AAQ, Dauwels J, King NKK. A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease. *Clin Neurophysiol*. 2019;130(1):145–154. doi:10.1016/j.clinph.2018.09.018
19. Kim V, Zhao H, Regan E, et al. The St.George's respiratory questionnaire definition of chronic bronchitis may be a better predictor of COPD exacerbations compared with the classic definition. *Chest*. 2019;156(4):685–695. doi:10.1016/j.chest.2019.03.041
20. Wang Z, Maschera B, Lea S, et al. Airway host-microbiome interactions in chronic obstructive pulmonary disease. *Respir Res*. 2019;20(1):113. doi:10.1186/s12931-019-1085-z
21. Leitao Filho FS, Alotaibi NM, Ngan D, et al. Sputum microbiome is associated with 1-year mortality after chronic obstructive pulmonary disease hospitalizations. *Am J Respir Crit Care Med*. 2019;199(10):1205–1213. doi:10.1164/rccm.201806-1135OC
22. Sartor RB. Therapeutic correction of bacterial dysbiosis discovered by molecular techniques. *Proc Natl Acad Sci USA*. 2008;105(43):16413–16414. doi:10.1073/pnas.0809363105
23. Manichanh C, Rigottier-Gois L, Bonnaud E, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*. 2006;55(2):205–211. doi:10.1136/gut.2005.073817
24. Mikami Y, Grubb BR, Rogers TD, et al. Chronic airway epithelial hypoxia exacerbates injury in muco-obstructive lung disease through mucus hyperconcentration. *Sci Trans Med*. 2023;15(699):eabo7728. doi:10.1126/scitranslmed.abo7728
25. Tanabe N, Matsumoto H. From mucus plugging to airway dilatation in chronic airway diseases: a perspective on the contribution of the airway microbiome and inflammation. *Allergol Int*. 2026;75(1):32–41. doi:10.1016/j.alit.2025.07.003
26. Abdel-Aziz MI, Brinkman P, Vijverberg SJH, et al. Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12 to 18 months. *J Allergy Clin Immunol*. 2021;147(1):123–134. doi:10.1016/j.jaci.2020.04.018

27. Huang YJ, Nariya S, Harris JM, et al. The airway microbiome in patients with severe asthma: associations with disease features and severity. *J Allergy Clin Immunol*. 2015;136(4):874–884. doi:10.1016/j.jaci.2015.05.044
28. Jia S, Liu P, Zhang H, Zeng H, Chen G, Zhao L. Why the COPD microbiome matters: how airway microbes shape disease severity and treatment response. *Int J Chronic Obstr*. 2026;21:531521. doi:10.2147/COPD.S531521
29. Tanabe N, Matsumoto H, Morimoto C, Hirai T. Sputum short-chain fatty acids, microbiome, inflammation, and mucus plugging in obstructive airway disease. *J Allergy Clin Immunol*. 2025;155(5):1675–1680. doi:10.1016/j.jaci.2025.01.031
30. Li X, Gao X, Wang Q, Wang C, Li B, Wan K. Feature analysis network: an interpretable idea in deep learning. *Cogni Comput*. 2024;16(3):803–826. doi:10.1007/s12559-023-10238-0
31. Fernández-Granero MA, Sánchez-Morillo D, León-Jiménez A, Crespo LF. Automatic prediction of chronic obstructive pulmonary disease exacerbations through home telemonitoring of symptoms. *Bio-Med Mater Eng*. 2014;24(6):3825–3832. doi:10.3233/BME-141212
32. Kocks JWH, Cao H, Holzhauer B, et al. Diagnostic performance of a machine learning algorithm (Asthma/Chronic Obstructive Pulmonary Disease [COPD] differentiation classification) tool versus primary care physicians and pulmonologists in asthma, COPD, and Asthma/COPD overlap. *J Allergy clin Immunol Pract*. 2023;11(5):1463–74.e3. doi:10.1016/j.jaip.2023.01.017
33. Mayhew D, Devos N, Lambert C, et al. Longitudinal profiling of the lung microbiome in the AERIS study demonstrates repeatability of bacterial and eosinophilic COPD exacerbations. *Thorax*. 2018;73(5):422–430. doi:10.1136/thoraxjnl-2017-210408