Dovepress
Taylor & Francis Group

ORIGINAL RESEARCH

Single Cell Transcriptomics Genomics Based on Machine Learning Algorithm: Constructing and Validating Neutrophil Extracellular Trap Gene Model in COPD

Jia Yu [1],*, Tiantian Xiao [1],*, Yun Pan [2],*, Yangshen He [1], Jiaxiong Tan [3]

[1]Department of Internal Medicine, Dongguan Hospital of Integrated Chinese and Western Medicine, Dongguan, Guangdong Province, People's Republic of China; [2]Department of Infectious Disease, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong Province, People's Republic of China; [3]Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yangshen He, Dongguan Hospital of Integrated Chinese and Western Medicine, Dongguan, Guangdong Province, 523000, People's Republic of China, Email 13713193315@163.com; Jiaxiong Tan, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, 300202, People's Republic of China, Email gdydtjx@163.com

Background: Neutrophil trap (NET) is an important feature of chronic inflammatory diseases. At present, there are still few studies to explore the characteristics of NET in different chronic obstructive pulmonary disease (COPD) patients. This study aimed to identify NET signature genes in different COPD patients.

Methods: We analyzed single-cell RNA sequencing data from COPD and non-COPD individuals to identify differentially expressed neutrophil genes. Machine learning algorithms were applied to construct models A and B, specific to smoking and non-smoking COPD patients, respectively.

Results: Through single-cell cluster analysis, 165 neutrophil characteristic genes in COPD group were successfully identified. Model A, consisting of key genes CD63, RNASE2, ERAP2, and model B, consisting of GRIPAP1, NHS, EGFLAM, and GLUL, were validated internally and externally, showing significant risk scores and good diagnostic efficacy (AUC: 60.24–87.22). Alveolar lavage fluid in patients with COPD was studied and confirmed higher expression levels of RNASE2 and NHS in severe COPD patients.

Conclusion: The study successfully developed NET signature gene models for identifying smoking and non-smoking COPD respectively, with validated specificity and predictive power, offering a foundation for personalized treatment strategies.

Keywords: COPD, neutrophil extracellular traps, single-cell sequencing, transcriptomics

Introduction

Chronic obstructive pulmonary disease (COPD) is now recognized as a complex, multicomponent disease characterized by chronic systemic inflammation and is currently the third leading cause of death worldwide.[1] The mortality rate for male COPD patients is higher, with an increasing trend in mortality rates among those over 45 years of age.[2] In developing countries, mortality rates are also increasing in tandem with the rise in smoking rates. In China, tobacco-related deaths account for 12% of all deaths, and predictions suggest that this proportion could reach 33% by 2030.[3] Although the hazardous role of smoking in the development of COPD is well established, nearly half of COPD patients are non-smokers, particularly women exposed to biomass smoke in poorly ventilated homes.[4]

The high mortality rate associated with COPD is linked to its progressive, irreversible airway obstruction and complex comorbidities.[5,6] Each bacterial or respiratory viral infection experienced by this population exponentially increases the risk of adverse outcomes.[6] Currently, identified endogenous factors associated with COPD include alpha-1

antitrypsin (AAT) deficiency and geNETic mutations in the telomerase gene TERT, which can lead to early-onset emphysema and are also seen in idiopathic pulmonary fibrosis.[7–9]

The role of inflammatory mechanisms in COPD has been progressively investigated, characterized by increased and activated macrophages and neutrophils.[10] Sputum neutrophilia is a significant feature of COPD, and this neutrophilic inflammation is induced by cigarette smoke, bacteria, viruses, and oxidative stress.[10] The increase and local chemotactic activation of neutrophils are associated with resistance to inhaled corticosteroids, which becomes a critical factor in exacerbating COPD mortality and progression.[11] Although CXCR2 antagonists (such as navarixin) can block the chemotactic effects of CXCL8 and related chemokines to reduce the number of neutrophils in the sputum of COPD patients, they do not improve already deteriorated lung function.[12]

In recent years, the formation of neutrophil extracellular traps (NETs) has gained increasing attention. Neutrophils, when recruited, have a certain probability of undergoing programmed cell death, leading to the rupture of cell membranes and the release of web-like structures containing extracellular double-stranded DNA (dsDNA) combined with histones, myeloperoxidase, and neutrophil elastase.[13] In respiratory viral infections, NETs formation is considered one of the beneficial protective mechanisms.[14] The formation of NETs is advantageous for the entrapment and containment of pathogens, and it has been reported that NETs exhibit antiviral properties against poxviruses.[15] Conversely, persistent NETs formation is considered a driver of immune pathology, most extensively demonstrated in COVID-19, where systemic and airway NET accumulation induces high inflammation in both acute and chronic disease stages.[16,17] NETs, loaded with highly basic histones and degradative enzymes, possess cytotoxic effects. When they are formed in excess or not cleared by the still poorly understood mechanisms, they can directly lead to host cell death and chronic tissue damage.[18] The tissue damage caused by local overload of NETs is closely related to pulmonary fibrosis, acute respiratory distress syndrome (ARDS), and the progression of asthma.[18,19] Studies have found that the sputum of exacerbated COPD patients contains a large amount of NET.[20] Importantly, NETs formation is not limited to exacerbated COPD; it also exists in stable COPD patients and correlates with the severity of airflow limitation.[20] Pharmacologists have been committed to manipulating NETs formation and reducing the pathogenic effects of NETs by administering DNase to degrade the NETs scaffold.[21] In recent years, with the development of genomics, a series of gene signatures related to NETs formation have been gradually identified in various diseases and have been well applied in predicting ischemia-reperfusion injury after lung transplantation, sepsis, and thrombotic diseases.[22–25]

Hematological and immunological data (HID) are crucial for the diagnosis, prognosis, and monitoring of numerous diseases.[26] As a result, HID data have been widely employed in artificial intelligence research to diagnose and predict the course of diseases.[27–29] In recent years, various clinical studies have demonstrated that HID is an effective and cost-efficient alternative for the early diagnosis and prognosis of coronavirus disease 2019 (COVID-19) and other diseases.[29–31] For instance, Huyut et al successfully developed artificial intelligence models using HID data for the diagnosis and prognosis of COVID-19.[32–34] Yan et al effectively applied the Gradient Boosting Decision Tree method to HID data for the early diagnosis of multiple myeloma.[35] Soerensen et al developed an artificial intelligence model based on HID to predict cancer risk in primary care.[36] Wu et al achieved 95.7% accuracy in diagnosing lung cancer using the Random Forest algorithm based on 19 routine blood test parameters.[37] Haider and others used artificial neural NET works to analyze HID data for early differentiation among leukemias.[38]

Compared with existing studies, this study innovatively identified NET-related genes associated with poor COPD prognosis using three machine learning methods combined with single cell sequencing dataset and multiple large sample lung tissue transcriptome sequencing dataset. We also constructed prognostic prediction models suitable for both smoking and non-smoking patients and preliminarily confirmed their good efficacy. This research will provide a basis for understanding the pathogenesis of COPD driven by gene-mediated NETs formation and for targeting NETs to improve the prognosis of COPD patients.

Materials and Methods
Data Acquisition and Preprocessing
We downloaded single-cell RNA sequencing (scRNA-seq) data of COPD lung tissue (GSE173896) and transcriptome sequencing data of COPD (GSE37768, GSE54837, GSE57148), and ARDS (GSE76293) from the Gene Expression

Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo/). Ultimately, we obtained scRNA-seq profiles from five COPD patient lung tissues and three age-matched non-COPD lung tissues as controls from GSE173896. The remaining transcriptome data were used to subsequently identify differentially expressed genes. The whole research process is shown in Supplemental Figure 1. In the entire workflow, first, the neutrophil signature genes in COPD tissues were successfully identified in comparison with normal tissues. Meanwhile, the following intergroup comparisons were conducted on the COPD-RNA-Seq data: high GOLD stage vs low GOLD stage, smokers vs non-smokers, and ARDS occurrence vs non-occurrence groups, to identify DEGs for subsequent machine learning. Subsequently, the preliminary validation efficacy was assessed using alveolar lavage fluid samples from COPD patients.

Acquisition of COPD Neutrophil Differential Gene Set

In order to obtain the genes associated with the neutrophil extranuclear trap, we first need to perform cluster analysis of single-cell sequencing samples of COPD and successfully annotate the neutrophil subsets. First, we performed quality control on the downloaded GSE173896 dataset to ensure the accuracy and reliability of the data. This included checking sequencing depth, excluding low-quality cells, and removing cells with a high proportion of mitochondrial genes (Supplementary Figure 2). We created Seurat objects based on the following filtering criteria: genes expressed in at least 3 cells and cells with at least 200 genes expressed. At the same time, samples with a total UMI number less than 1000 and samples with a gene number less than 500 were filtered out. The first 2000 highly variable genes were then selected for subsequent analysis. We utilized the RunPCA function from the Seurat package to perform principal component analysis (PCA) on the preprocessed data, selected high variable genes (HVGs) for dimensionality reduction using PC4, and employed the Run-TSNE function for t-SNE clustering analysis. The results of t-SNE cell subsets clustering in all samples of COPD group and normal control group are shown in Figure 1A and B. Clustered cell populations were annotated using Cell Marker 2.0 to identify components.[39] Differential analysis with control group data identified 165 highly expressed neutrophil differential genes.
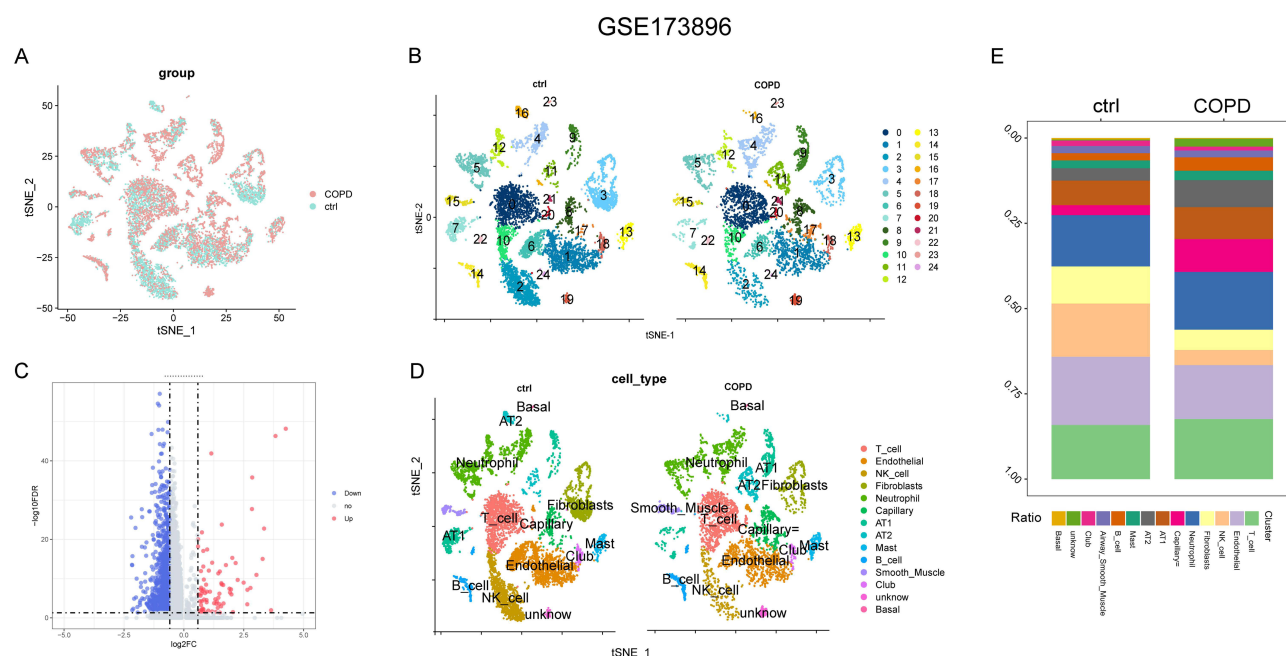

Figure 1 Single cell sequencing analysis of COPD and Control samples. (**A**) Preliminary results of tSNE clustering using high-mutation genes from PC4 in dimensionality-reduced data obtained from five COPD samples and three control samples. (**B**) Twenty-four cellular clusters identified through clustering at the same resolution. (**C**) Volcano plot of differential genes within the neutrophil cluster, with blue representing downregulated differential genes and red representing upregulated differential genes. (**D**) tSNE plot after cell cluster annotation. (**E**) Stacked plot of annotated cellular components, with different colors representing distinct cell clusters.

Identification of Differentially Expressed Genes (DEGs) Between Different Subgroups of COPD

Having successfully identified a range of neutrophil subpopulation specific genes from single-cell cluster analysis, we then need to identify which genes are specific to COPD disease, which genes are associated with poor disease prognosis, and even which genes may be associated with smoking as a risk factor. After obtaining gene mRNA expression matrices and clinical information for 18 COPD samples, 11 smoking healthy controls, and 9 non-smoking healthy controls from GSE37768; 68 GOLD stage 3/4 (stage high), 68 GOLD <3 (stage low), 84 smoker controls, and 6 non-smoker controls from GSE54837; and 12 ARDS samples and 12 normal controls from GSE76293, we preprocessed the data, including gene matching, removal of blank information, and standardization. We used the R package "limma" for intergroup differential analysis and "ggplot2" for visualizing DEGs, setting the criteria to |log2 fold change (FC)| > 1.0 and adjusted p-value < 0.05. Additionally, we employed the R package "cluster Profiler" for KEGG and GO enrichment analysis of differential genes. Furthermore, we used the MCP counter estimate function to analyze the infiltration of eight common immune cells in the GSE37768, GSE54837, and GSE76293 datasets.

RF-LASSO-SVM RFE Machine Learning for Feature Gene Selection and Model Construction

Machine learning can better identify the weight of different factors in different disease states. Therefore, we carry out machine learning on the differential genes obtained in the previous step to identify which genes are the most critical. We intersected the COPD smoking DEGs and COPD nonsmoking DEGs obtained above with COPD stage high DEGs, ARDS DEGs, and scRNA-NDEGs. Finally, we extracted 23 smoking COPD-NDEGs and 33 nonsmoking COPD-NDEGs and included them in the random forest algorithm, LASSO regression analysis, and SVM RFE machine learning. The three machine learning methods were completed using the R packages "randomForest", "glmnet", and "e1071", respectively. All variables were used for SVM model construction, with AvgRank sorted by the average ranking of 10-fold cross-validation. Ultimately, we obtained 3 COPD smoking NDEGs (CD63, RNASE2, and ERAP2) and 4 COPD nonsmoking NDEGs (GRIPAP1, NHS, EGFLAM, and GLUL) for the construction of models A and B through intersection.

Single-Gene Efficacy Assessment and Model Efficacy Validation

We used the GSE37768 dataset to perform group expression analysis and obtain ROC diagnostic dependency curves for the 7 feature genes selected by machine learning to assess the efficacy of single genes in identifying COPD. Subsequently, we used the R package "clusterProfiler" for GSEA analysis of highly expressed differential genes. Additionally, we studied the protein molecular interaction patterns of the two groups of genes using protein-protein interaction (PPI) NETworks and STRING database-based protein interaction NETwork analysis. Furthermore, we constructed gene diagnostic nomograms for the two groups of NDEGs. To validate the efficacy of the machine learning-built models A-B, we used GSE37768 as the internal training set and GSE54837 and GSE57148 as external validation sets for risk score grouping validation and construction of corresponding ROC curves.

Distribution and Protein Structure Prediction of RNASE2 and NHS

We investigated the distribution of RNASE2 and NHS genes through the website Human Protein Atlas (proteinatlas.org), including mRNA distribution in different human organs and various immune cell subsets, as well as subcellular localization prediction. Finally, we constructed protein structure models for RNASE2 and NHS using Alphafold 2.0.

qRT-PCR and ELISA Experiments on COPD Bronchoalveolar Lavage Fluid Samples

We collected bronchoalveolar lavage fluid from ten COPD patients for the detection of RNASE2 and NHS genes and MPO protein ELISA. Among them, five patients requiring ventilator support were classified as stage high, and the other five patients not requiring ventilator support were classified as stage low. The sequence of primers associated with each gene for qPCR is provided in Supplemental Table 1. mRNA was isolated using a commercial Total RNA

Purification Kit from NORGEN, and its concentration was determined by a Thermo Fisher Scientific spectrophot-ometer. The mRNA was then converted to cDNA using the High-Capacity RNA-to-cDNA Kit from Applied Biosystems. Quantitative PCR (qPCR) was performed with SYBR green PCR master mix, also from Applied Biosystems, on a Bio-Rad CFX Manager system. Finally, Δct values were used as the relative expression of RNASE2 and NHS genes compared to the GAPDH reference gene. A commercial MPO ELISA kit was used, which includes a pre-coated microplate with specific anti-MPO antibodies, enzyme-labeled secondary antibodies, washing buffer, substrate solution, and stopping solution. According to the kit instructions, the supernatant samples obtained after centrifugation of bronchoalveolar lavage fluid were appropriately diluted, and standard curves were constructed using standard samples at different gradient concentrations. Then, experimental samples were added, incubated, washed, and developed according to the instructions, and finally, the absorbance (OD values) of each well was measured at 450 nm using an enzyme-linked immunosorbent assay reader. The obtained sample OD values were converted into relative concentrations using the standard curve.

Statistical Methods

Data analysis and visualization were conducted using R software (version 4.3.3) along with GraphPad Prism 8. For the group difference analysis of RT-PCR and ELISA data, we employed both the two-tailed unpaired Student's *t*-test and the Wilcoxon rank-sum test. For the differential analysis of GEO transcriptome data, we utilized the limma-voom method to assess differences between groups. A statistically significant result was considered when the P-value was less than 0.05.

Results

Successful Identification of COPD Neutrophil Differential Genes via scRNA-Seq

We analyzed the scRNA-seq profiles of lung tissues from five COPD patients in the GSE173896 dataset, with three age-matched non-COPD lung tissues as controls. As shown in Supplementary Figure 2, after rigorous quality control analysis, data standardization and scaling, and selection and filtering of highly variable feature-expressing cells, four principal component analysis modules were successfully constructed (Supplementary Figure 2F), with the top 20 genes of each PCA displayed in Supplementary Figure 2G. In Figure 1A and B, following tSNE nonlinear dimensionality reduction of PC4, 24 cellular clusters were preliminarily identified. Neutrophil populations in both COPD and control samples were successfully identified after cellular annotation, yielding 165 overexpressed neutrophil differential genes (scRNA-NDEGs) (Figure 1C). The composition and proportion distribution of cells are shown in Figure 1D and E, with a significantly higher neutrophil ratio in COPD samples compared to controls.

Acquisition of COPD and ARDS Differential Genes from Large-Scale Transcriptome Datasets

To obtain differential genes between COPD and smoking and non-smoking controls, we performed group differential analysis on the mRNA data of 18 COPD samples, 11 smoking healthy controls, and 9 non-smoking healthy controls from the GSE37768 dataset, yielding 178 COPD-Smoking-DEGs and 511 COPD-Non-smoking-DEGs. To assess differences among COPD patients of different GOLD stages, we conducted group differential analysis on 68 GOLD stage 3/4 (stage high), 68 GOLD <3 (stage low), 84 smoker, and 6 non-smoker controls from the GSE54837 dataset. Finally, 708 COPD-stage high DEGs, 2076 COPD stage high-smoking-DGEs, and 904 COPD stage high-Nonsmoking-DEGs, were identi-fied. To further analyze gene differences in lung tissue samples as COPD progresses to ARDS, we performed differential analysis on 12 ARDS samples and 12 normal control samples from the GSE76293 dataset. We successfully identified 3320 DEGs in the transcriptome sets of ARDS and normal control samples. We selected the top 30 differential genes from the above subgroup analysis and displayed them in the form of heat map (Supplementary Figure 3A–F). Meanwhile, the results of gene expression difference analysis for each subgroup were displayed in the form of volcano map (Supplementary Figure 4A–F).

Candidate Gene Selection for Smokers and Non-Smokers and KEGG and GO Enrichment

To explore the impact of smoking and non-smoking on neutrophil proliferation, activation, and disease progression in COPD patients, we intersected the multiple transcriptome differential genes obtained with scRNA-NDEGs. Results in Supplementary Figure 5G and H show that we identified 23 smoking COPD-NDEGs and 33 non-smoking COPD-NDEGs for subsequent machine learning. Additionally, we performed KEGG and GO enrichment analyses on the differential genes obtained from the transcriptome samples, finding that genes highly expressed in COPD stage high were enriched in membrane potential regulation, passive transmembrane transporter activity, and ion channel activation pathways (Supplementary Figure 5E and F). COPD stage high differential genes are closely related to the formation of cell membranes and ion channel complexes (Supplementary Figure 5E and F). Compared to the non-smoking group, COPD-Smoking-DEGs were mainly enriched in DNA and cell signal transduction pathways, phosphatidylcholine metabolism pathways, and DNA damage repair checkpoint pathways (Supplementary Figure 5A and B). In contrast, COPD-NonSmoking-DEGs were primarily enriched in pathways related to endothelial cell proliferation and migration (Supplementary Figure 5C and D).

MCPcounter Immune Infiltration Analysis

To assess differences in immune cell infiltration among different COPD subgroup samples, we used the "MCPcounter" R package to quantify and score the distribution differences of eight immune cells in the GSE37768, GSE54837, and GSE76293 datasets. Results in Supplementary Figure 6A and B show that the COPD group had higher infiltration of T cells, monocytes, and neutrophils, with the most significant difference in neutrophil infiltration. Compared to the non-smoking group, smoking group samples showed a trend of increased neutrophil and monocyte infiltration. The COPD stage high group had significantly higher neutrophil infiltration than the COPD stage low and other control groups. Interestingly, higher neutrophil infiltration scores were also observed in ARDS samples (Supplementary Figure 6C).

Machine Learning Model Construction Using RandomForest, LASSO, and SVM-REF

To evaluate the disease-related effects of the 23 smoking COPD-NDEGs and 33 nonsmoking COPD-NDEGs obtained in previous steps and to construct an efficient model, we employed three machine learning methods: RandomForest, LASSO, and SVM-REF, to jointly filter key feature genes. For the smoking group, distinct outcomes were observed: The SVM-REF model prioritized 5 genes to optimize performance metrics (AvgRank < 10), as shown in Figure 2A and B. LASSO regression analysis, validated through 10-fold cross-validation, proposed 3 genes (CD63, RNASE2, and ERAP2) for Model A construction (Figure 2C and D). Gene importance rankings derived from the RandomForest algorithm are presented in Figure 2E and F. The intersection of gene selections across methods was visualized in Figure 2G, ultimately defining the final candidate genes for downstream modeling. For the nonsmoking group, the SVM-REF model identified 7 genes required to achieve maximum accuracy and minimum error rate under the condition of AvgRank < 10, as illustrated in Figure 3A and B. Following 10-fold cross-validation, LASSO regression further refined the selection, recommending 4 genes (GRIPAP1, NHS, EGFLAM, and GLUL) for constructing Model B (Figure 3C and D). The importance rankings of these genes, evaluated by the RandomForest algorithm, are displayed in Figure 3E and F. A Venn diagram analysis integrating results from the three methods (Figure 3G) highlighted candidate genes through intersection screening, which were subsequently used to build gene feature models. ALL the candidate genes after intersection screening were constructed by following gene feature models:

Model A (smoker): risk score = (0.126 * expression of CD63) + (0.535 * expression of RNASE2) + (0.019 * expression of ERAP2)

Model B (nonsmoker): risk score = (0.353 * expression of GRIPAP1) + (1.497 * expression of NHS) + (0.0726 * expression of EGFLAM) + (0.627 * expression of GLUL)

Single-Gene Evaluation

To assess the feature genes selected by machine learning, we used the GSE37768 dataset as a training set for single-gene efficacy validation. Evaluation results of the 3 candidate genes for Model A (Figure 4A–C) showed that COPD group
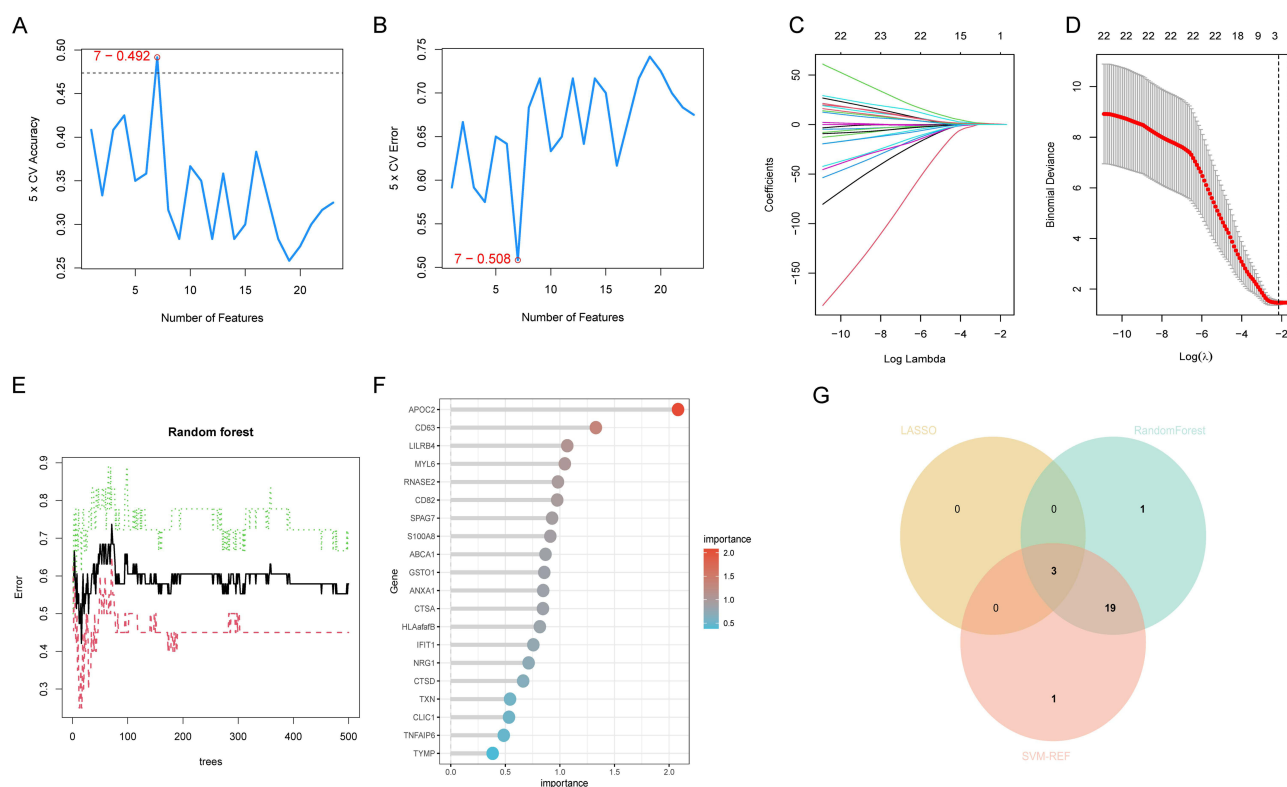
Figure 2 Machine Learning Analysis and Model Construction with Smoking COPD-NDEGs. (**A** and **B**) display the results obtained using the SVM-RFE algorithm, illustrating the number of feature genes recommended for achieving the highest accuracy and the lowest error rate, respectively. (**C** and **D**) present the outcomes of incorporating feature genes into LASSO regression analysis. The y-axis in (**C**) represents the coefficients assigned to each feature gene, while the dashed line in (**D**) indicates the optimal number of factors recommended. (**E** and **F**) showcase the results derived from stepwise random forest analysis. The lollipop chart in E is a sorted display based on the importance coefficients assigned. (**G**) is a Venn diagram representing the intersection of genes recommended by three machine learning methods. The machine learning above uses the smoking COPD-NDEGs obtained in the previous step.

samples have higher single-gene mRNA expression levels, with RNASE2 mRNA levels significantly higher than those in the normal control group (p=0.032). However, CD63 and ERAP2 did not show a difference between the two groups. Single-gene GSEA enrichment analysis showed that the 3 highly expressed differential genes in the COPD-smoking group are enriched in pathways related to allogeneic graft rejection, cobalamin transport and metabolism, and virus-related immune activation (Figure 4D–4F). Subsequent diagnostic ROC analysis revealed that the above 3 feature genes have high efficacy in diagnosing COPD (AUC: 0.586–0.727) (Figure 4G–4I). We then performed protein-protein interaction (PPI) analysis and using the STRING database to investigate the molecular interaction patterns of these genes. The results (Figure 4J and K) indicated that the primary interacting partners of CD63, RNASE2, and ERAP2 were CTSD, CTSA, and HLA-B, respectively. We also developed a clinical diagnostic array based on the three feature genes (Figure 4L).

In the COPD-nonsmoking group, GRIPAP1, NHS, EGFLAM, and GLUL expression levels are all significantly elevated compared to the control group (P=0.0083, p=0.015, p=0.0024, and p=0.0016) (Figure 5A–D). The 4 highly expressed differential genes in the COPD-nons moking group were enriched in pathways related to IgA production, allogeneic graft rejection, apoptosis, and IL-17 signaling (Figure 5E–H). Subsequent diagnostic ROC analysis revealed that the above 4 feature genes have high efficacy in diagnosing COPD (AUC: 0.735–0.870) (Figure 5I-L). The protein interaction network for the four feature genes in the COPD-nonsmoking group comprises key players such as ENO1, GAPDH, and CD63, with comprehensive interaction maps presented in Figure 5M and N. A diagnostic array diagram for clinical application was developed based on the four feature genes (Figure 5O).

Systematic Validation of Machine Learning Model Efficacy

Next, we conducted efficacy validation of the machine learning-constructed Models A and B across multiple datasets. Initially, analysis of the training set GSE37768 dataset showed that COPD group samples had significantly higher risk scores
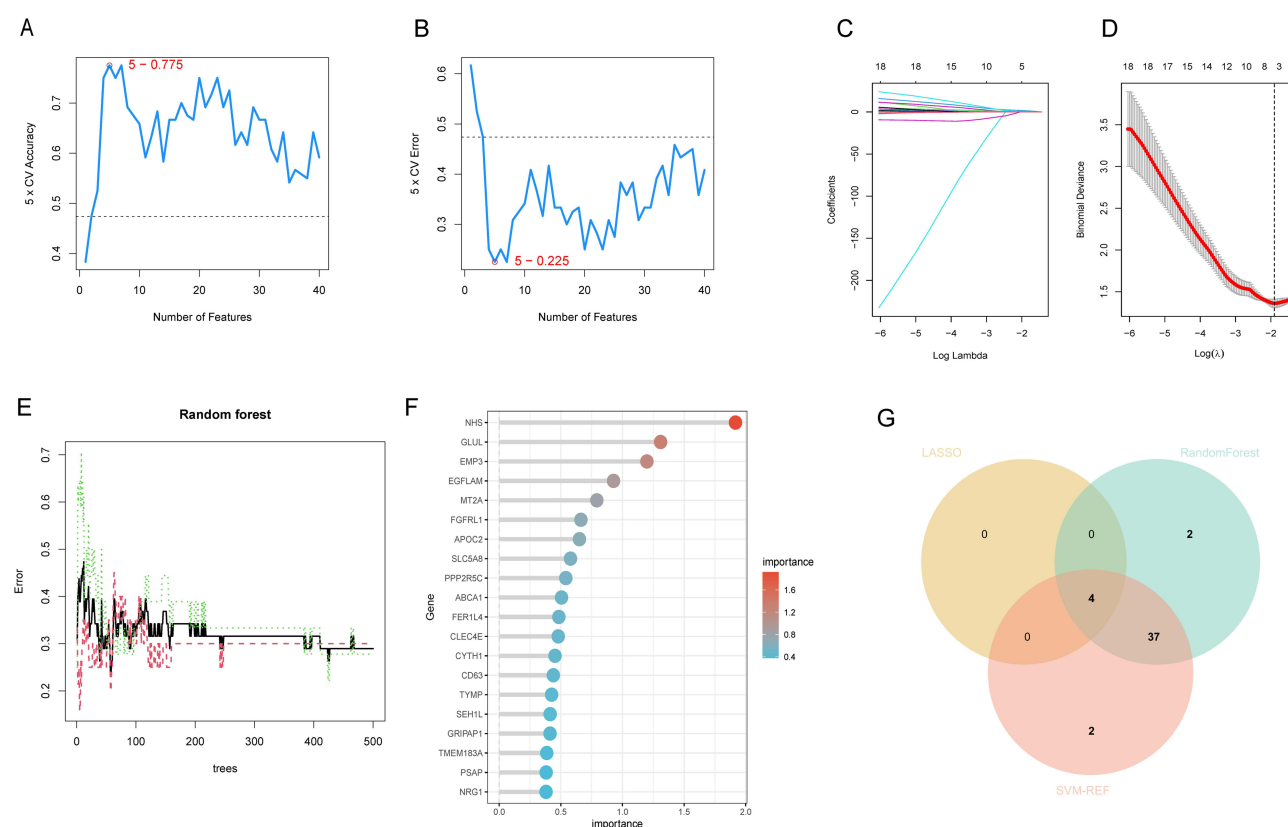
Figure 3 Machine Learning Analysis and Model Construction with nonsmoking COPD-NDEGs. (**A** and **B**) display the results obtained using the SVM-RFE algorithm, illustrating the number of feature genes recommended for achieving the highest accuracy and the lowest error rate, respectively. (**C** and **D**) present the outcomes of incorporating feature genes into LASSO regression analysis. The y-axis in (**C**) represents the coefficients assigned to each feature gene, while the dashed line in (**D**) indicates the optimal number of factors recommended. (**E** and **F**) showcase the results derived from stepwise random forest analysis. The lollipop chart in E is a sorted display based on the importance coefficients assigned. (**G**) is a Venn diagram representing the intersection of genes recommended by three machine learning methods. The machine learning above uses the nonsmoking COPD-NDEGs obtained in the previous step.

for both Model A and Model B (p<0.05); while the ROC curves indicated good diagnostic efficacy (AUC=71.388 and AUC=87.222), with Model B outperforming Model A, further distinguishing between smoking and nonsmoking samples (Figure 6A–D). Additionally, we used the external dataset GSE54837 for efficacy assessment. The results, as shown in Figure 6E–H, indicate that in Model A, both COPD stage high and stage low groups had higher scores than the smoking control group (p=0.044, p=0.043); while in Model B, only the COPD stage high score was significantly higher than the smoking control group (p=0.0097); Models A and B showed similar diagnostic efficacy (AUC=60.245 and AUC=62.287). Similarly, analysis of another independent COPD dataset GSE57148 revealed that, compared to the normal control group, COPD group samples had significantly higher risk scores in both Model A and Model B (p<0.001); while the ROC curves indicated that Model A had better diagnostic efficacy than Model B (AUC=70.228 and AUC=62.839) (Figure 6I–L).

Distribution Patterns of RNASE2 and NHS and Protein Structure Prediction Based on Alphafold

Since the three machine learning methods assigned the highest importance coefficients to the RNASE2 and NHS genes, we conducted a systematic investigation and analysis of their distribution. The results, as shown in Figures Supplementary 7–8A, indicate that RNASE2 and NHS genes are expressed to varying degrees in multiple human organs, with RNASE2 being highly expressed in the spleen and lungs, and NHS in the retina and cervix; further analysis of their distribution in immune cells revealed that, in addition to high expression in neutrophils, RNASE2 and NHS genes are also highly expressed in PBMCs, monocytes, macrophages, basophils, and myeloid DC cells to varying degrees (Supplementary Figures 7, 8B and C). Subcellular localization revealed that RNASE2 exists in a membrane secretion
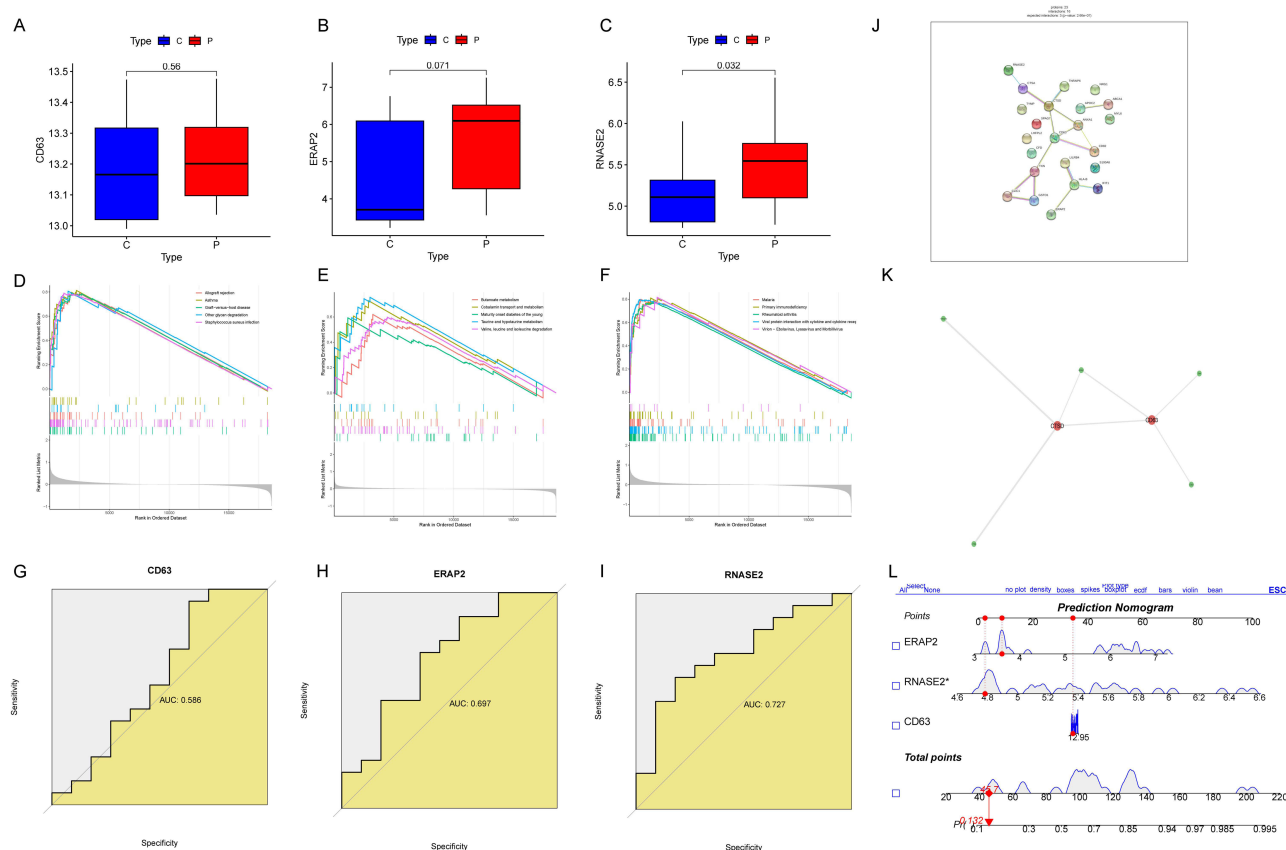
Figure 4 Single-Gene Assessment, Protein-Protein Interaction NETwork Analysis, and Nomogram Construction in smoking COPD group. (**A–C**) present the gene expression group statistics for three smoking COPD NDEGs derived from machine learning on the GSE37768 dataset. The blue bars signify the normal control group, and the red bars represent the COPD group, with the y-axis indicating the level of gene expression. (**D–F**) depict the GSEA enrichment of differential genes associated with the high expression of the three single genes. (**G–I**) showcase the diagnostic dependency ROC curves for assessing the efficacy of single genes, with the yellow area corresponding to the area under the curve. (**J** and **K**) display the PPI protein-protein interaction NETworks and protein interaction diagrams based on STRING database analysis. (**L**) is a gene diagnostic Nomogram constructed from the three genes.

manner, while NHS is involved in the construction of the cell membrane scaffold (Supplementary Figures 7–8D). Finally, we used Alphafold version 2.0 to construct protein structure models for RNASE2 and NHS (Supplementary 7–8E).

qRT-PCR and ELISA Validation of Alveolar Lavage Fluid Samples

Lastly, we extracted RNA from the sediment of alveolar lavage fluid samples from 10 COPD patients for RT-PCR validation and used the ELISA method to detect the relative concentration of MPO in the supernatant. The results, as shown in Figure 7A and B, indicate that the relative expression levels of RNASE2 and NHS genes in the alveolar lavage fluid of COPD stage-high patients are significantly higher than those in COPD stage-low patients. After constructing the MPO-ELISA standard curve, the results of converting sample OD values to relative concentrations using the formula $y = 1220.5x - 118.1$ showed that the supernatant of alveolar lavage fluid from COPD stage-high patients had a higher concentration of MPO (Figure 7C and D).

Discussion

This study innovatively combines single-cell sequencing with large-scale transcriptomics and employs three classic machine learning methods to identify NETs signature genes closely associated with poor prognosis in COPD. These genes were used to construct models for predicting outcomes in different COPD populations based on smoking differences. The effectiveness of these models was preliminarily validated using alveolar lavage fluid samples.

Inflammatory mechanisms play a decisive role in the development of COPD, characterized by increased and activated macrophages and neutrophils.[14] An increase in sputum neutrophils is a significant feature of COPD, and this neutrophilic
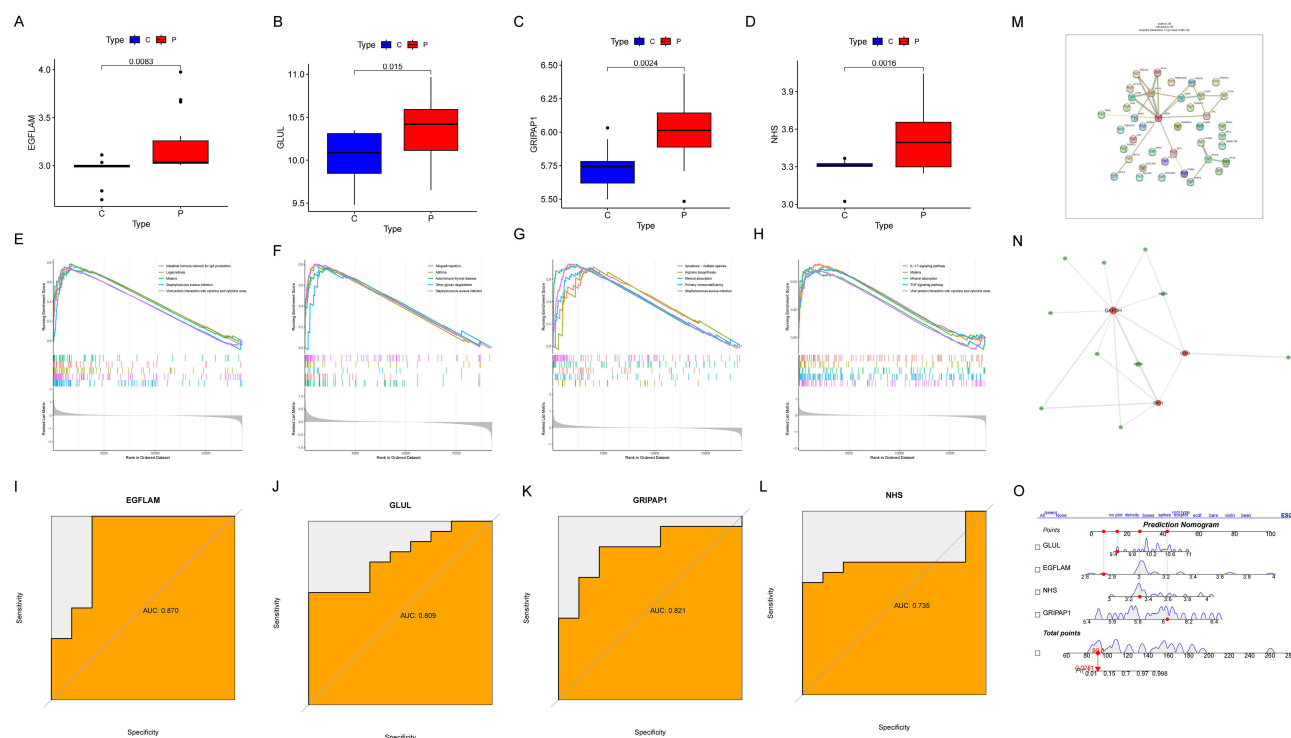
Figure 5 Single-Gene Assessment, Protein-Protein Interaction NETwork Analysis, and Nomogram Construction in nonsmoking COPD group. (**A–D**) present the gene expression group statistics for four nonsmoking COPD NDEGs derived from machine learning on the GSE37768 dataset. The blue bars signify the normal control group, and the red bars represent the COPD group, with the y-axis indicating the level of gene expression. (**E–H**) depict the GSEA enrichment of differential genes associated with the high expression of the three single genes. (**I–L**) showcase the diagnostic dependency ROC curves for assessing the efficacy of single genes, with the yellow area corresponding to the area under the curve. (**M** and **N**) display the PPI protein-protein interaction NETworks and protein interaction diagrams based on STRING database analysis. (**O**) is a gene diagnostic Nomogram constructed from the three genes.
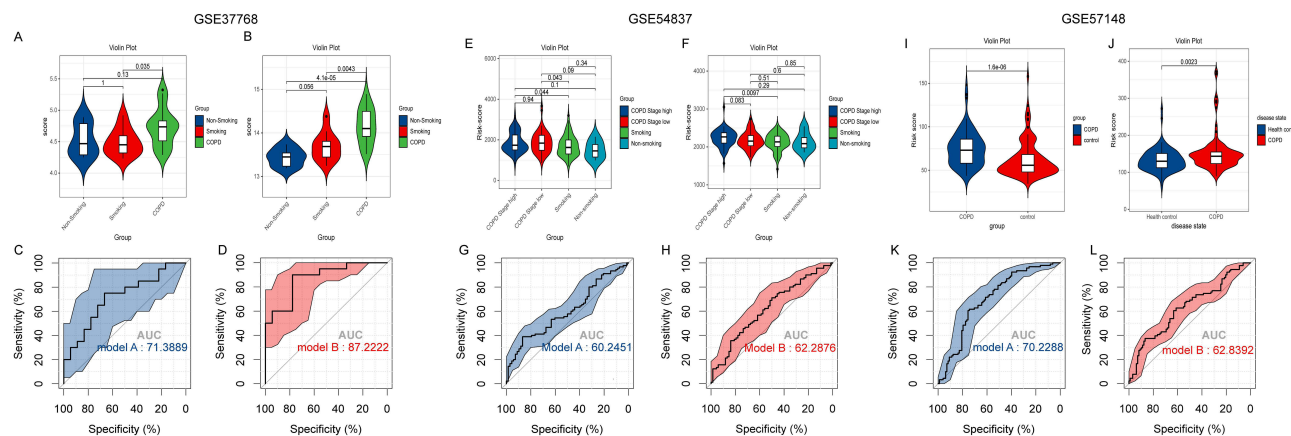

Figure 6 Validation of the Risk Scoring Model with Internal and External Datasets. (**A–D**) depict the validation outcomes of the internal dataset GSE37768 using Model A and Model B. In (**A** and **B**), the y-axis denotes the computed risk scores, with violin plots in various colors representing different groups. (**C** and **D**) exhibit the diagnostic dependency ROC curves for Model A and Model B, respectively, with the surrounding hues indicating the confidence intervals. (**E** and **F**) and (**I** and **J**) illustrate the disparities in group model scores for the external validation datasets GSE54837 and GSE57148. (**G** and **H**) and (**K** and **L**) present the diagnostic dependency ROC curves for Model A and Model B on the external validation datasets GSE54837 and GSE57148, respectively, with the surrounding hues signifying the confidence intervals.

inflammation is induced by cigarette smoke, bacteria, viruses, and oxidative stress.[10] Current research indicates that sputum from all levels of COPD patients, whether exacerbated or stable, is characterized by the presence of a large number of NETs and NET-forming neutrophils.[40,41] This is highly consistent with the results of our MCPcounter immune infiltration analysis study. Additionally, we detected varying concentrations of MPO in the alveolar lavage fluid of COPD patients. NETosis related to COPD is significantly associated with lung function impairment, which is considered one of the best indicators of disease
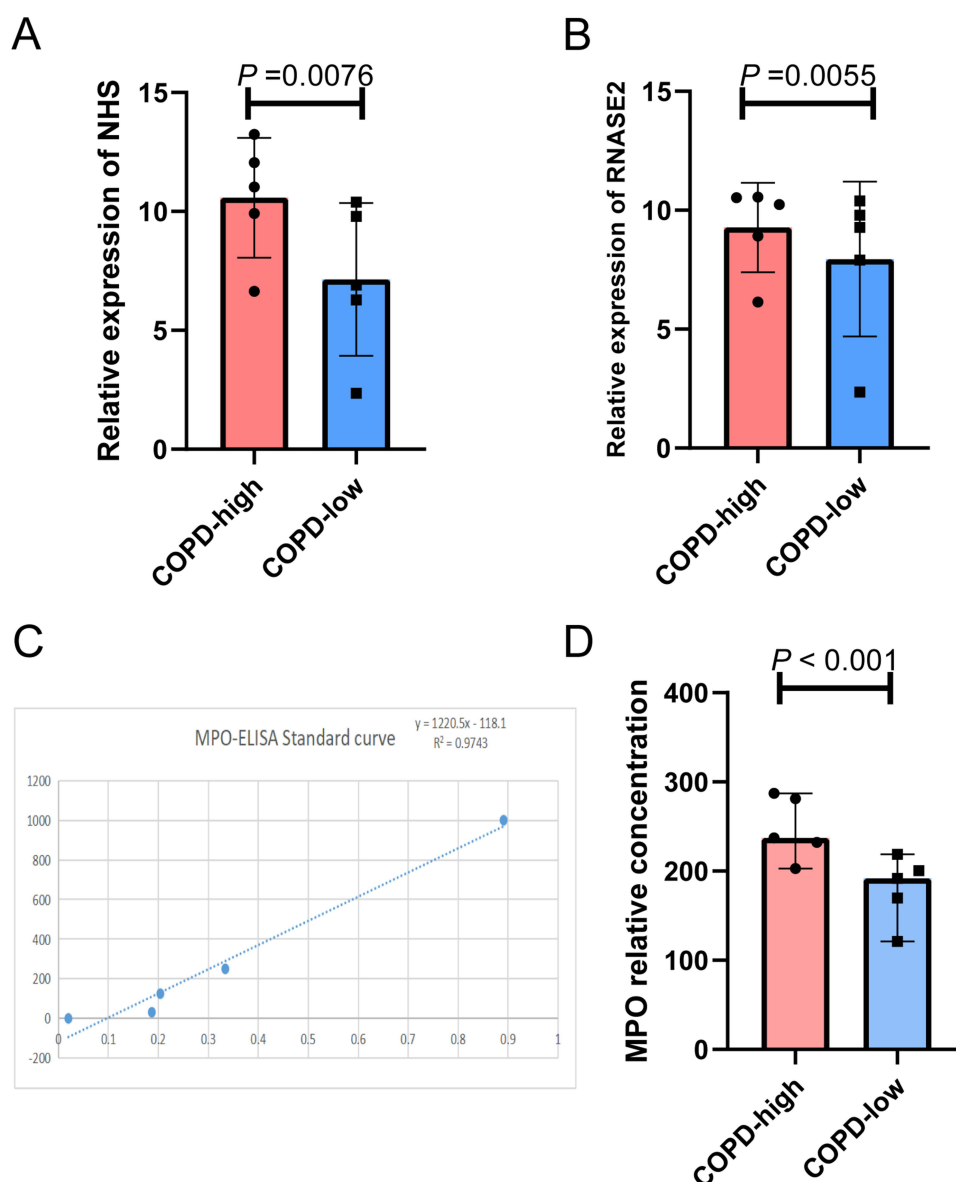
Figure 7 Quantitative RT-PCR and ELISA Analysis of RNASE2 and NHS Gene Expression and MPO Concentration in Clinical Samples. (**A** and **B**) display the outcomes of quantitative real-time polymerase chain reaction (qRT-PCR), where the red bars indicate COPD patients who need mechanical ventilation support (stage high), and the blue bars indicate those who do not (stage low). The y-axis denotes the relative expression levels of the target genes normalized against the GAPDH housekeeping gene, represented as the relative expression ratio (ΔCt value). (**C**) illustrates the ELISA standard curve that was established, while (**D**) shows the relative concentration of myeloperoxidase (MPO) calculated after calibration against the standard curve. The red bars in (**D**) represent COPD patients requiring mechanical ventilation support (stage high), and the blue bars represent those not requiring such support (stage low).

severity, further strengthening the relevance and accuracy of our results. All characteristics of NETosis are more commonly observed in the sputum of exacerbated patients and in GOLD stage 3 and 4 patients.[20] Nevertheless, NETosis is clearly detected in the sputum of stable COPD and GOLD stage 1 and 2 patients.[20]

The formation of NETosis can be understood as a result of neutrophil autophagy, an essential component of innate immunity.[13] This programmed cell death through self-destruction is observed in numerous inflammatory diseases and cancers.[42,43] Their formation not only helps to confine pathogens and cancer cells but also aids in clearing pathogenic antigens due to the highly basic histones and degradative enzymes within their structure.[43] However, excessive neutrophil infiltration and activation leading to a large amount of pathological NETosis formation and cytokine storms may exacerbate lung function deterioration in patients with respiratory diseases.[17,44] Studies have shown that NETs formation is not limited to exacerbations but also exists in stable COPD and is correlated with the severity of airflow

limitation.[20] NETs are major contributors to chronic inflammation and lung tissue damage in COPD most of the time. Pharmacologists have been working on manipulating NETs formation and administering DNase to degrade the NETs scaffold to reduce the pathogenic effects of NETs.[42] With the development of genomics in recent years, a series of gene signatures related to NETs formation have been gradually identified in various diseases and have been well applied in predicting ischemia-reperfusion injury after lung transplantation, sepsis, and thrombotic diseases.[22,23] In our study, we used single-cell sequencing clustering analysis to identify differentially expressed genes in neutrophils and successfully identified NETosis-related genes associated with poor prognosis in COPD using machine learning combined with high-level gene signatures in COPD and ARDS from transcriptomics. Similarly, we detected higher levels of NETosis-related genes and MPO concentrations in alveolar lavage fluid samples from severe COPD patients requiring ventilator support.

A substantial body of research indicates that the genetic impact of smoking behavior increases over time.[45] Long-term smoking suppresses DNA repair mechanisms and causes genomic changes, leading to the occurrence and development of respiratory diseases and even cancer.[46] Non-smoking respiratory disease patients often have unique genetic biological characteristics and genetic susceptibility.[47] Interestingly, we found differences in gene transcriptomes between smoking and non-smoking COPD, leading us to construct separate NETosis-related gene models for predicting outcomes in smoking and non-smoking COPD patients. These models were preliminarily validated for effectiveness in smoking and non-smoking subgroups using multiple external datasets. Feature importance evaluation by the three machine learning methods showed that RNASE2 and NHS genes play a key role in smoking and non-smoking groups, respectively. Human RNase2 is a secreted protein expressed in white blood cells and is reported to have antiviral activity against single-stranded RNA viruses.[48] Current research also finds that RNase2 specifically cleaves cellular ncRNA.[49] During oxidative stress responses, Schlafen 2 (SLFN2) protein can protect tRNA from RNase cleavage.[50] The immunomodulatory role and potential targeting of RNase2 in human macrophage cell RNA populations have been preliminarily validated.[49] However, whether the high expression of the RNASE2 gene mediates neutrophil autophagy and further promotes NETosis formation during the inflammatory state of COPD exacerbations remains unconfirmed. NHS gene mutations are found to be associated with the X-linked hereditary disease Nance-Horan syndrome (NHS).[51,52] Early studies found extra expression of NHS in the kidneys, lungs, and thymus.[51] Our subcellular localization found that NHS seems to be closely related to the scaffold structure of the cell membrane. There is currently very limited research on the NHS gene in immune cells.

In summary, the innovation of this study lies in the use of machine learning combined with single-cell omics and transcriptomics to screen for NET signature genes in smoking COPD and non-smoking COPD and to construct two gene models suitable for predicting outcomes in different COPD groups. However, the limitation of this study is the small number of clinical samples used for validation, which may cause some bias in the data. Additionally, there are no fixed indicators for the detection of neutrophil extracellular traps, often requiring multiple indicators and various detection techniques for identification. This study preliminarily uses MPO alone as a detection indicator for neutrophil extracellular traps, which can only serve as a preliminary reference, and the above shortcomings will be gradually improved in our subsequent research.

Conclusion

In conclusion, this study identified NET signature genes in smoking COPD and non-smoking COPD using various machine learning methods combined with single-cell omics and transcriptomics, and constructed COPD NDEGs Model A and Model B. The specificity of the feature genes RNASE2 and NHS and the good predictive efficacy of the scoring models were validated through internal training sets, external datasets, and clinical samples.

Data Sharing Statement

The datasets utilized and analyzed in this study were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

Ethical Approval

The research project has been reviewed and approved by the Ethics Committee of the Guangdong Dongguan Integrated Traditional Chinese and Western Medicine Hospital. This study adheres to the "Ethical Review Methods for Biomedical Research Involving Humans" and the Declaration of Helsinki.

Informed Consent

We have explained the purpose, process, risks, and benefits of the research to all individuals participating in the study verbally and have obtained their informed consent.

Author Contributions

Jia Yu, Tiantian Xiao and Yun Pan are co-first authors with a joint contribution to the article. 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.

Disclosure

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

References

1. Venkatesan P. GOLD COPD report: 2024 update. *Lancet Respir Med*. 2024;12(1):15–16. doi:10.1016/S2213-2600(23)00461-7
2. Agusti A, Bohm M, Celli B, et al. GOLD COPD DOCUMENT 2023: a brief update for practicing cardiologists. *Clin Res Cardiol*. 2024;113 (2):195–204. doi:10.1007/s00392-023-02217-0
3. Niu SR, Yang GH, Chen ZM, et al. Emerging tobacco hazards in China: 2. Early mortality results from a prospective study. *BMJ*. 1998;317 (7170):1423–1424. doi:10.1136/bmj.317.7170.1423
4. Yang IA, Jenkins CR, Salvi SS. Chronic obstructive pulmonary disease in never-smokers: risk factors, pathogenesis, and implications for prevention and treatment. *Lancet Respir Med*. 2022;10(5):497–511. doi:10.1016/S2213-2600(21)00506-3
5. Suissa S, Dell'Aniello S, Ernst P. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. *Thorax*. 2012;67(11):957–963. doi:10.1136/thoraxjnl-2011-201518
6. Ferrera MC, Labaki WW, Han MK. Advances in chronic obstructive pulmonary disease. *Annu Rev Med*. 2021;72(1):119–134. doi:10.1146/ annurev-med-080919-112707
7. Dasi F. Alpha-1 antitrypsin deficiency. *Med Clin*. 2024;162(7):336–342. doi:10.1016/j.medcli.2023.10.014
8. Cordoba-Lanus E, Montuenga LM, Dominguez-de-Barros A, et al. Oxidative damage and telomere length as markers of lung cancer development among chronic obstructive pulmonary disease (COPD) smokers. *Antioxidants*. 2024;13(2). doi:10.3390/antiox13020156.
9. Guzman-Vargas J, Ambrocio-Ortiz E, Perez-Rubio G, et al. Differential genomic profile in TERT, DSP, and FAM13A between COPD patients with emphysema, IPF, and CPFE syndrome. *Front Med Lausanne*. 2021;8:725144. doi:10.3389/fmed.2021.725144
10. Barnes PJ. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2016;138(1):16–27. doi:10.1016/j.jaci.2016.05.011
11. Ernst P, Saad N, Suissa S. Inhaled corticosteroids in COPD: the clinical evidence. *Eur Respir J*. 2015;45(2):525–537. doi:10.1183/ 09031936.00128914
12. Rennard SI, Dale DC, Donohue JF, et al. CXCR2 antagonist MK-7123. A Phase 2 proof-of-concept trial for chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2015;191(9):1001–1011. doi:10.1164/rccm.201405-0992OC
13. Papayannopoulos V. Neutrophil extracellular traps in immunity and disease. *Nat Rev Immunol*. 2018;18(2):134–147. doi:10.1038/nri.2017.105

14. Jenne CN, Wong CH, Zemp FJ, et al. Neutrophils recruited to sites of infection protect from virus challenge by releasing neutrophil extracellular traps. *Cell Host Microbe*. 2013;13(2):169–180. doi:10.1016/j.chom.2013.01.005

15. Hemmers S, Teijaro JR, Arandjelovic S, Mowen KA. PAD4-mediated neutrophil extracellular trap formation is not required for immunity against influenza infection. *PLoS One*. 2011;6(7):e22043. doi:10.1371/journal.pone.0022043

16. George PM, Reed A, Desai SR, et al. A persistent neutrophil-associated immune signature characterizes post-COVID-19 pulmonary sequelae. *Sci Transl Med*. 2022;14(671):eabo5795. doi:10.1126/scitranslmed.abo5795

17. Veras FP, Pontelli MC, Silva CM, et al. SARS-CoV-2-triggered neutrophil extracellular traps mediate COVID-19 pathology. *J Exp Med*. 2020;217 (12). doi:10.1084/jem.20201129

18. Villanueva E, Yalavarthi S, Berthier CC, et al. NETting neutrophils induce endothelial damage, infiltrate tissues, and expose immunostimulatory molecules in systemic lupus erythematosus. *J Immunol*. 2011;187(1):538–552. doi:10.4049/jimmunol.1100450

19. Dworski R, Simon HU, Hoskins A, Yousefi S. Eosinophil and neutrophil extracellular DNA traps in human allergic asthmatic airways. *J Allergy Clin Immunol*. 2011;127(5):1260–1266. doi:10.1016/j.jaci.2010.12.1103

20. Grabcanovic-Musija F, Obermayer A, Stoiber W, et al. Neutrophil extracellular trap (NET) formation characterises stable and exacerbated COPD and correlates with airflow limitation. *Respir Res*. 2015;16(1):59. doi:10.1186/s12931-015-0221-7

21. Ravindran M, Khan MA, Palaniyar N. Neutrophil extracellular trap formation: physiology, pathology, and pharmacology. *Biomolecules*. 2019;9 (8):365. doi:10.3390/biom9080365

22. Sheng M, Cui X. A machine learning-based diagnostic model for myocardial infarction patients: analysis of neutrophil extracellular traps-related genes and eQTL Mendelian randomization. *Medicine (Baltimore)*. 2024;103(12):e37363. doi:10.1097/MD.0000000000037363

23. Wang L, Wang Q, Li Y, Qi X, Fan X. A signature based on neutrophil extracellular trap-related genes for the assessment of prognosis, immunoinfiltration, mutation and therapeutic response in hepatocellular carcinoma. *J Gene Med*. 2024;26(1):e3588. doi:10.1002/jgm.3588

24. Huang Z, Zhang H, Fu X, et al. Autophagy-driven neutrophil extracellular traps: the Dawn of sepsis. *Pathol Res Pract*. 2022;234:153896. doi:10.1016/j.prp.2022.153896

25. Gao J, Zhang Z, Yu J, et al. Identification of neutrophil extracellular trap-related gene expression signatures in ischemia reperfusion injury during lung transplantation: a transcriptome analysis and clinical validation. *J Inflamm Res*. 2024;17:981–1001. doi:10.2147/JIR.S444774

26. Huyut MT. Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models. *Ing Rech Biomed*. 2023;44(1):100725. doi:10.1016/j.irbm.2022.05.006

27. Huyut MT, Huyut Z. Forecasting of oxidant/antioxidant levels of COVID-19 patients by using expert models with biomarkers used in the diagnosis/ prognosis of COVID-19. *Int Immunopharmacol*. 2021;100:108127. doi:10.1016/j.intimp.2021.108127

28. Huyut MT, Huyut Z. Effect of ferritin, INR, and D-dimer immunological parameters levels as predictors of COVID-19 mortality: a strong prediction with the decision trees. *Heliyon*. 2023;9(3):e14015. doi:10.1016/j.heliyon.2023.e14015

29. Valle A, Rodriguez J, Camina F, Rodriguez-Segade M, Ortola JB, Rodriguez-Segade S. The oxyhaemoglobin dissociation curve is generally left-shifted in COVID-19 patients at admission to hospital, and this is associated with lower mortality. *Br J Haematol*. 2022;199(3):332–338. doi:10.1111/bjh.18431

30. Mertoglu C, Huyut MT, Olmez H, Tosun M, Kantarci M, Coban TA. COVID-19 is more dangerous for older people and its severity is increasing: a case-control study. *Med Gas Res*. 2022;12(2):51–54. doi:10.4103/2045-9912.325992

31. Huyut MT, Velichko A. LogNNET model as a fast, simple and economical AI instrument in the diagnosis and prognosis of COVID-19. *MethodsX*. 2023;10:102194. doi:10.1016/j.mex.2023.102194

32. Huyut MT, Ustundag H. Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study. *Med Gas Res*. 2022;12(2):60–66. doi:10.4103/2045-9912.326002

33. Velichko A, Huyut MT, Belyaev M, Izotov Y, Korzun D. Machine learning sensors for diagnosis of COVID-19 disease using routine blood values for InterNET of things application. *Sensors*. 2022;22(20):7886. doi:10.3390/s22207886

34. Huyut MT, Velichko A. Diagnosis and prognosis of COVID-19 disease using routine blood values and LogNNET neural NETwork. *Sensors*. 2022;22(13):4820. doi:10.3390/s22134820

35. Yan W, Shi H, He T, et al. Employment of artificial intelligence based on routine laboratory results for the early diagnosis of multiple myeloma. *Front Oncol*. 2021;11:608191. doi:10.3389/fonc.2021.608191

36. Soerensen PD, Christensen H, Gray Worsoe Laursen S, Hardahl C, Brandslund I, Madsen JS. Using artificial intelligence in a primary care setting to identify patients at risk for cancer: a risk prediction model based on routine laboratory tests. *Clin Chem Lab Med*. 2022;60(12):2005–2016. doi:10.1515/cclm-2021-1015

37. Wu CC, Yeh WC, Hsu WD, et al. Prediction of fatty liver disease using machine learning algorithms. *Comput Methods Programs Biomed*. 2019;170:23–29. doi:10.1016/j.cmpb.2018.12.032

38. Haider RZ, Ujjan IU, Khan NA, Urrechaga E, Shamsi TS. Beyond the in-practice CBC: the research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias. *Diagnostics*. 2022;12(1). doi:10.3390/diagnostics12010138

39. Hu C, Li T, Xu Y, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res*. 2023;51(D1):D870–D876. doi:10.1093/nar/gkac947

40. Trivedi A, Khan MA, Bade G, Talwar A. Orchestration of neutrophil extracellular traps (NETs), a unique innate immune function during chronic obstructive pulmonary disease (COPD) development. *Biomedicines*. 2021;9(1):53. doi:10.3390/biomedicines9010053

41. Katsoulis O, Toussaint M, Jackson MM, et al. Neutrophil extracellular traps promote immunopathogenesis of virus-induced COPD exacerbations. *Nat Commun*. 2024;15(1):5766. doi:10.1038/s41467-024-50197-0

42. Mutua V, Gershwin LJ. A review of neutrophil extracellular traps (NETs) in disease: potential Anti-NETs therapeutics. *Clin Rev Allergy Immunol*. 2021;61(2):194–211. doi:10.1007/s12016-020-08804-7

43. Demkow U. Neutrophil extracellular traps (NETs) in cancer invasion, evasion and metastasis. *Cancers*. 2021;13(17):4495. doi:10.3390/cancers13174495

44. Szturmowicz M, Demkow U. Neutrophil extracellular traps (NETs) in severe SARS-CoV-2 lung disease. *Int J mol Sci*. 2021;22(16):8854. doi:10.3390/ijms22168854

45. Domingue BW, Conley D, Fletcher J, Boardman JD. Cohort Effects in the GeNETic Influence on Smoking. *Behav GeNET*. 2016;46(1):31–42. doi:10.1007/s10519-015-9731-9

46. Li Q, Wang T, Tang Y, et al. A novel prognostic signature based on smoking-associated genes for predicting prognosis and immune microenvironment in NSCLC smokers. *Cancer Cell Int.* 2024;24(1):171. doi:10.1186/s12935-024-03347-9

47. Smolle E, Pichler M. Non-smoking-associated lung cancer: a distinct entity in terms of tumor biology, patient characteristics and impact of hereditary cancer predisposition. *Cancers.* 2019;11(2):204. doi:10.3390/cancers11020204

48. Rosenberg HF, Dyer KD, Domachowske JB. Respiratory viruses and eosinophils: exploring the connections. *Antiviral Res.* 2009;83(1):1–9. doi:10.1016/j.antiviral.2009.04.005

49. Lu L, Li J, Wei R, et al. Selective cleavage of ncRNA and antiviral activity by RNase2/EDN in THP1-induced macrophages. *Cell mol Life Sci.* 2022;79(4):209. doi:10.1007/s00018-022-04229-x

50. Yue T, Zhan X, Zhang D, et al. SLFN2 protection of tRNAs from stress-induced cleavage is essential for T cell-mediated immunity. *Science.* 2021;372(6543). doi:10.1126/science.aba4220

51. Brooks SP, Ebenezer ND, Poopalasundaram S, Lehmann OJ, Moore AT, Hardcastle AJ. Identification of the gene for Nance-Horan syndrome (NHS). *J Med GeNET.* 2004;41(10):768–771. doi:10.1136/jmg.2004.022517

52. Huang Y, Ma L, Zhang Z, et al. Nance-Horan syndrome pedigree due to a novel microdeletion and skewed X chromosome inactivation. *mol GeNET Genomic Med.* 2023;11(2):e2100. doi:10.1002/mgg3.2100