J Korean Med Sci. 2025 Dec 8;40(47):e311
https://doi.org/10.3346/jkms.2025.40.e311
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Imaging

A Novel Machine Learning Approach to Assist Early Diagnosis of Diffuse Panbronchiolitis

Hwan Jin Lee,[1,2*] **Kyung Joon Heo**,[3*] **Yeon Seok You**,[2,4,5] **Kum Ju Chae**,[2,6] **Jong Seung Kim**,[2,4,5,7] **Jae Seok Jeong**,[1,2,7,8] **and Yong Chul Lee**[1,2,7]

[1]Department of Internal Medicine and Research Center for Pulmonary Disorders, Jeonbuk National University Medical School, Jeonju, Korea
[2]Research Institute of Clinical Medicine of Jeonbuk National University-Biomedical Research Institute of Jeonbuk National University Hospital, Jeonju, Korea
[3]Department of Internal Medicine, Sinchon Severance Hospital, Seoul, Korea
[4]Department of Otorhinolaryngology-Head and Neck Surgery, Jeonbuk National University Medical School, Jeonju, Korea
[5]Department of Medical Informatics, Jeonbuk National University Medical School, Jeonju, Korea
[6]Department of Radiology, Jeonbuk National University Medical School, Jeonju, Korea
[7]Respiratory Drug Development Research Institute, Jeonbuk National University Medical School, Jeonju, Korea
[8]Laboratory of Respiratory Immunology and Infectious Diseases, Korea Zoonosis Research Institute, Jeonbuk National University, Iksan, Korea

Received: Jan 14, 2025
Accepted: Apr 22, 2025
Published online: Nov 3, 2025

Address for Correspondence:
Jong Seung Kim, MD, PhD
Department of Internal Medicine, Jeonbuk National University Medical School, 20 Geonji-ro, Deokjin-gu, Deokjin-gu, Jeonju 54907, Republic of Korea.
Email: kjsjdk@gmail.com

Jae Seok Jeong, MD, PhD
Department of Internal Medicine, Jeonbuk National University Medical School, 20 Geonji-ro, Deokjin-gu, Deokjin-gu, Jeonju 54907, Republic of Korea.
Email: jeongjs@jbnu.ac.kr

Yong Chul Lee, MD, PhD
Department of Internal Medicine, Jeonbuk National University Medical School, 20 Geonji-ro, Deokjin-gu, Deokjin-gu, Jeonju 54907, Republic of Korea.
Email: leeyc@jbnu.ac.kr

*Hwan Jin Lee and Kyung Joon Heo contributed equally to this work.

ABSTRACT

Background: Diffuse panbronchiolitis (DPB) is a rare and progressive inflammatory lung disease affecting the small airways; however, it is often misdiagnosed as other respiratory conditions, such as nontuberculous mycobacterial infection or bronchiectasis. This study aimed to apply machine learning (ML) algorithms to improve early diagnostic accuracy for DPB.
Methods: ML algorithms were applied using clinical, laboratory, and radiological data from 99 patients with suspected DPB. Patients were categorized into two groups based on established diagnostic criteria and major diagnostic criteria for DPB without impaired lung function. Seven ML models were evaluated.
Results: The least absolute shrinkage and selection operator regression model demonstrated the highest predictive accuracy. The analysis identified two key diagnostic factors, allergic rhinitis and the presence of macronodules on computed tomography scans, both of which were strongly associated with DPB.
Conclusion: These results highlight the first application of ML in diagnosing DPB and underscore the significance of allergic rhinitis and macronodules as critical indicators for early detection. Incorporating ML techniques into clinical practice could improve the diagnostic accuracy and efficiency for rare diseases such as DPB. Further research involving larger patient datasets is recommended to validate these results and refine the diagnostic criteria for DPB.

Keywords: Diffuse Panbronchiolitis; Machine Learning; Macronodule; Allergic Rhinitis

INTRODUCTION

Diffuse panbronchiolitis (DPB) is a severe and progressive inflammatory disease of unknown etiology that affects small airways, such as the respiratory bronchioles.[1] If left untreated,

ORCID iDs

Hwan Jin Lee
https://orcid.org/0000-0003-2702-7863
Yeon Seok You
https://orcid.org/0000-0003-3956-4412
Kum Ju Chae
https://orcid.org/0000-0003-3012-3530
Jong Seung Kim
https://orcid.org/0000-0002-1384-6799
Jae Seok Jeong
https://orcid.org/0000-0002-4635-8302
Yong Chul Lee
https://orcid.org/0000-0002-0433-509X

Funding

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.RS-2024-00356349; JSJ) and the Special Operating Subsidy of the Jeonbuk National University Industrial Cooperation Foundation; a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: RS-2024-00440408; JSJ); and funds from the Biomedical Research Institute, Jeonbuk National University Hospital.

Disclosure

The authors have no potential conflicts of interest to disclose.

Data Availability Statement

The datasets generated and/or analyzed are not publicly available for ethical and legal reasons. Nevertheless, they can be made available from the corresponding authors, Kim JS, Jeong JS, and Lee YC, upon reasonable request.

Author Contributions

Conceptualization: Lee YC. Data curation: You YS. Formal analysis: Heo KJ. Funding acquisition: Jeong JS. Investigation: Heo KJ, Chae KJ. Project administration: Lee YC. Software: Kim JS. Supervision: Kim JS, Jeong JS, Lee YC. Validation: Kim JS. Visualization: Lee HJ. Writing - original draft: Lee HJ, You YS. Writing - review & editing: Lee HJ, Jeong JS.

it ultimately progresses to irreversible structural changes leading to respiratory failure and death.[2] Long-term treatment with macrolides significantly improves clinical outcomes and enables disease control.[3] Therefore, early diagnosis of DPB is essential. The current diagnostic criteria proposed by a working group of the Ministry of Health and Welfare of Japan (1999) may be insufficient for early diagnosis of DPB; the titer of cold hemagglutinin in patients with DPB is a minor diagnostic criterion and is not widely used in clinical practice. Therefore, methods that are more efficient are required.

Machine learning (ML) algorithms are well-known in clinical applications because of their capability to integrate various clinical data for highly accurate early diagnosis.[4] Although successful applications of these algorithms have frequently been reported for early diagnosis of various diseases, there have been no studies on the application of ML to DPB to date. This study aimed to evaluate the effectiveness of ML algorithms for diagnosing DPB using clinical, laboratory, and radiological variables that reflect different aspects of the disease process.

METHODS

Study design and setting

This retrospective observational study was conducted in 2022 using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (**Supplementary Data 1**).

Dataset description and participants

We included a retrospective cohort of 274 patients with suspected DPB who underwent chest computed tomography (CT) between March 1998 and June 2019 at a single tertiary medical center. A total of 175 patients were excluded because of the absence of essential data such as spirometry results or medical records indicating upper airway disease. Considering that a substantial number of bronchioles should be affected before the onset of clinically evident obstructive symptoms,[5] we divided the remaining 99 patients with suspected DPB into groups A (n = 34) and B (n = 65) based on the assessment of lung physiology using spirometry. Patients in Group A fulfilled all diagnostic criteria for DPB. Patients in Group B fulfilled only the major diagnostic criteria for DPB without disturbed lung physiology, which manifested as forced expiratory volume in 1 second (FEV1)/forced vital capacity (FVC) values of less than 70% (suspected DPB).

Variables

We selected 30 variables to identify those involved in the development of DPB. The selected variables comprised 11 categorical and 19 continuous variables. Categorical variables consisted of one demographic variable (sex), four related ENT profiles (nasal polyp, rhinitis, chronic rhinitis syndrome [CRS], and ENT clinic), five CT characteristics (bronchiectasis, cystic bronchiectasis, cavity, macronodule, and consolidation), and one Akira type. Continuous variables included one demographic variable (age), six pulmonary function prediction variables (FEV1_percent, FVC_percent, TLC_percent, VC_percent, RV_percent, and DLCO_percent), one blood profile (hsCRP_liter_per_mg), three BAL fluid flow cytometry-related variables (Lym_percent, Neu_percent, and T-lymphocyte_HS_Ratio), eight image variables, divided into lobar (RUL, RML, RLL, LUL, and LLL) and zonal (ULZ, MLZ, and LLZ) distributions.

Performance measurements

Before analyzing several models, a performance measure for model comparison should be selected. In this study, area under the receiver operating characteristic (ROC) curve (AUC) and binomial deviation were adopted. The AUC is the calculated value of the area under the ROC curve, and the binomial deviation is defined as

$$D = 2\sum\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right)\right].$$

where yi and $\hat{\mu}_i$ are the i-th observed response and the probability of DPB estimated from the model, respectively. The smaller this measurement, the closer the predicted value is to the actual value. Further, in this study, cross-validation was used to compare statistical and ML models. The performance of the models was compared using the average of two measurements in five groups using 5-fold cross-validation.

Statistical modeling

Based on the collected data, DPB development was analyzed using several ML modeling techniques to identify variables that affect DPB. First, we applied the logistic regression-based generalized linear models that were most suitable for modeling probability.
The relationship between the dependent variable DPB and the 30 independent variables was analyzed using the data given in the ridge regularization with alpha = 0, elastic net (EN) = 0.5, and least absolute shrinkage and selection operator (LASSO) regularization = 1 by adjusting the alpha value.

Definition of elastic net for generalized linear models

For α strictly between 0 and 1 and nonnegative λ, EN solves the problem by

$$\left[\frac{(1-\alpha)}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right]$$

where $\beta = (\beta_1,\ldots,\beta_p)' \in R^p$ is the vector of regression coefficients, $l(\beta)$ is the log-likelihood function that one tries to maximize in the usual logistic regression without regularizing, ë > 0 is a penalty parameter, and α∈[0,1] is a mixing parameter that balances between the LASSO and ridge models. We noted that α = 0 corresponded to the ridge and α = 1 to the LASSO model. We further noted that α > 0 may yield a sparse model, which meant that some of the estimated coefficients shrank to exactly zero. Therefore, it was expected that regularization would not only improve the prediction performance, but also automatically select relevant predictors. For implementation, the "glmnet" package in R was used and λ was selected via fivefold cross-validation for each α∈{0,0.5,1}. The α = 0.5 case was denoted by EN for simplicity.

ML models

In addition to the three statistical models used above, four ML models were applied to investigate the prediction performance and major variables involved in the onset of DPB.

Tree-based models

A single decision tree generally has a low predictive power. Therefore, in this study, the random forest and bagging models, which are ensemble models of multi-decision trees, were applied. Bagging and random forest depend on the number of randomly generated trees selected. Bagging is a method used to select 30 equal to the number of variables, and a random forest is used to extract five with the root value (5≈√30).

https://doi.org/10.3346/jkms.2025.40.e311

Support vector machine (SVM)

SVMs can be used for almost any task, including classification and numerical prediction. This methodology can effectively solve the non-linear classification problem by transforming to a higher dimension in situations that are difficult to distinguish in the existing linear space. A method called the kernel trick is used to perform the above transform effectively. In this study, modeling was performed by automatically setting the kernel trick using R packages "rminer" and "e1071."

Neural network (NN)

A NN is a statistical learning algorithm inspired by the structures of NNs in biology. It is known as a methodology with high accuracy for linear and non-linear classification by linearly connecting multiple nodes. However, it is difficult to interpret the model as to what process the result was derived from inside, as it is called a black box. In this study, a model that learned a total of 65 weights by setting two nodes with a hidden layer on the first layer for 30 variables using R package "nnet" was used.

The seven ML models used were logistic regression-based models (LASSO, EN, and ridge), tree-based models (bagging, random forest, which are ensemble methods composed of multiple decision tree analyses), SVM, and NN. As we fitted multiple ML models, their comparison was performed using the AUC. Cross-validation approaches were used to calculate the AUC metric to select a more flexible model for the new data. The closer the AUC value is to 1, the higher the accuracy and predictive power of the model.

Code availability

Code and scripts to reproduce the analyses are presented in **Supplementary Data 2**.

Ethics statement

The Institutional Review Board of Jeonbuk National University Hospital approved the study (approval No. 2019-11-046) and waived the requirement for informed consent because the K-NHIS data were de-identified. The study was conducted in accordance with the principles of the Declaration of Helsinki. All procedures were performed in accordance with the relevant guidelines and regulations.

RESULTS

Comparisons of clinicoradiographic characteristics

The standard mean difference (SMD) was used to compare differences between the two groups for each of the 30 variables. As shown in **Table 1**, the mean age (standard deviation [SD]) of patients in groups A and B was 59.18 (14.33) and 52.06 (16.32) years, respectively. The proportion of males in the control and groups A and B was 52.9% and 50.8%, respectively. No significant differences were observed in the demographic data, upper airway symptoms, pulmonary function, inflammatory phenotypes, or CT variables between the two groups. FEV1 (SMD = 1.432) showed a significant difference between the two groups.

Comparative analysis of patient characteristics and model performance

The characteristics of the patients in groups A and B were analyzed using seven ML models (**Table 2**). The performances of the models were ranked based on the area under the receiver operating characteristic curve (AUC): LASSO (0.9498), EN (0.8801), ridge (0.8285), random

Table 1. Characteristics and SMD between groups A and B for 30 variables

Variables	Group A (n = 34)	Group B (n = 65)	*P* value	SMD
Demographic data				
Age, yr	59.18 (14.33)	52.06 (16.32)	0.034	0.463
Sex, male	18 (52.94)	33 (50.77)	1.000	0.043
Upper airway symptom				
Nasal polyps	4 (11.76)	7 (10.77)	1.000	0.031
Allergic rhinitis	17 (50.00)	11 (16.92)	0.001	0.748
CRS	11 (32.35)	13 (20.00)	0.265	0.284
ENT clinics	7 (20.59)	17 (26.15)	0.714	0.132
Pulmonary function				
FEV1 (% predicted)	54.06 (19.24)	79.68 (16.44)	< 0.001	1.432
FVC (% predicted)	66.41 (15.87)	90.51 (95.16)	0.147	0.353
TLC (% predicted)	114.43 (14.24)	111.58 (12.79)	0.314	0.211
VC (% predicted)	69.22 (13.74)	78.38 (12.02)	0.001	0.709
RV (% predicted)	194.90 (49.32)	168.57 (40.09)	0.005	0.586
DLCO (% predicted)	72.99 (22.96)	87.07 (14.71)	< 0.001	0.730
Inflammatory phenotype				
hsCRP, mg/L	31.44 (36.33)	27.49 (22.46)	0.506	0.131
Lymphocytes, %	20.14 (9.53)	17.50 (9.38)	0.189	0.279
Neutrophils, %	44.32 (10.44)	45.15 (16.75)	0.795	0.059
H/S ratio	1.29 (0.17)	1.41 (0.48)	0.142	0.350
CT Distribution				
RUL	0.65 (0.47)	0.33 (0.41)	0.001	0.720
RML	0.81 (0.51)	0.71 (0.44)	0.324	0.205
RLL	0.96 (0.48)	0.91 (0.28)	0.530	0.122
LUL	0.62 (0.49)	0.43 (0.43)	0.055	0.403
LLL	1.01 (0.48)	0.92 (0.27)	0.195	0.253
ULZ	0.56 (0.50)	0.38 (0.41)	0.058	0.394
MLZ	0.78 (0.45)	0.72 (0.36)	0.445	0.157
LLZ	1.03 (0.43)	0.98 (0.19)	0.395	0.160
Characteristics				
Bronchiectasis	19 (55.88)	52 (80.00)	0.022	0.535
Cystic bronchiectasis	9 (26.47)	17 (26.15)	1.000	0.007
Cavity	2 (5.88)	0 (0.00)	0.221	0.354
Macronodule[a]	0.09 (0.29)	0.00 (0.00)	0.015	0.433
Consolidation[b]	8 (23.53)	9 (13.85)	0.351	0.250
Akira type	2.24 (0.99)	2.25 (0.85)	0.955	0.012

Values are given as mean ± standard deviation or percentages.
SMD = standardized mean difference, CRS = chronic rhinitis syndrome, ENT clinics = outpatient visit to the otolaryngologist, FEV1 = forced expiratory volume in one second, FVC = forced vital capacity, TLC = total lung capacity, VC = vital capacity, RV = residual volume, DLCO = diffusing capacity of the lung for carbon monoxide, H/S ratio = the proportion of CD4+ to CD8+ T-cell sequences.
[a]Macronodule is defined as a nodule between 10 mm and 30 mm in diameter.
[b]Consolidation is defined as ground glass opacity greater than 30 mm in diameter.

forest (0.8054), bagging (0.7774), NN (0.7389), and SVM (0.5862). The LASSO model had the best performance (**Fig. 1A**). It is a representative linear model and a method of automatic variable selection. The results revealed a linear relationship between the main variables and DPB. **Fig. 1B** shows a heat map of the coefficients for the top ten variables of the absolute value of DPB in the three statistical models, including the LASSO model, which had the best performance. The most important variables selected in the lasso were allergic rhinitis (1.76), bronchiectasis (−1.65), FEV1 (%) (−1.41), and macronodules (1.23). EN showed the second-best performance, and the variables selected were allergic rhinitis (1.62), bronchiectasis (−1.45), macronodules (1.27), and FEV1 (%) (−1.05). Finally, the third-best performance was by the ridge model and the variables selected were macronodules (1.39), allergic rhinitis (1.28), bronchiectasis (−1.07), and FEV1 (%) (−0.65). Allergic rhinitis was ranked among the top two variables in all three models, and macronodules were consistently ranked within the top four for all three models

https://doi.org/10.3346/jkms.2025.40.e311

Table 2. The top ten variables based on their absolute values for DPB, identified using seven machine learning models

Model	Variables	Values
LASSO	Rhinitis	1.76
	Bronchiectasis	−1.65
	FEV1, %	−1.41
	Macronodule[a]	1.23
	Consolidation[b]	0.48
	Age	0.40
	CRS	0.36
	LUL	0.12
	RV, %	0.07
	DLCO, %	−0.06
Elastic	Rhinitis	1.62
	Bronchiectasis	−1.45
	Macronodule	1.27
	FEV1_percent	−1.05
	Consolidation	0.46
	CRS	0.46
	Age	0.43
	LUL	0.18
	DLCO, %	−0.17
	RV, %	0.13
Ridge	Macronodule	1.39
	Rhinitis	1.28
	Bronchiectasis	−1.07
	FEV1, %	−0.65
	Cavity	0.54
	Consolidation	0.49
	CRS	0.45
	Age	0.42
	LLL	0.36
	LUL	0.34
Random forest	FEV1, %	23.11
	FVC, %	3.91
	Age	2.80
	Rhinitis	1.94
	Bronchiectasis	1.60
	RV, %	1.15
	TLC, %	1.01
	DLCO, %	0.82
	hsCRP, mg/L	0.81
	Macronodule	0.81
Bagging	FEV1, %	10.65
	FVC, %	4.88
	Age	3.22
	VC, %	2.72
	DLCO, %	2.64
	Rhinitis	2.49
	RV, %	1.81
	Bronchiectasis	1.47
	TLC, %	1.38
	hsCRP, mg/L	1.30
Neural network	Bronchiectasis	0.10
	FVC, %	0.09
	FEV1, %	0.09
	Cystic bronchiectasis	0.07
	Rhinitis	0.06
	RLL	0.06
	CRS	0.06
	Age	0.05
	LLL	0.05
	RV, %	0.04

(continued to the next page)

https://doi.org/10.3346/jkms.2025.40.e311

Table 2. (Continued) The top ten variables based on their absolute values for DPB, identified using seven machine learning models

Model	Variables	Values
Support vector machine	FEV1, %	0.33
	TLC, %	0.17
	LLZ	0.16
	FVC, %	0.16
	Age	0.15
	Macronodule	0.14
	ULZ	0.13
	RLL	0.12
	Rhinitis	0.11
	Bronchiectasis	0.08

DPB = diffuse panbronchiolitis, FEV1 = forced expiratory volume in one second, CRS = chronic rhinitis syndrome, LUL = left upper lobe, RV = residual volume, DLCO = diffusing capacity of the lung for carbon monoxide, LLL = left lower lobe, FVC = forced vital capacity, TLC = total lung capacity, hsCRP = high-sensitivity C-reactive protein, VC = vital capacity.
[a]Macronodule us defined as a nodule between 10 mm and 30 mm in diameter.
[b]Consolidation is defined as ground glass opacity greater than 30 mm in diameter.

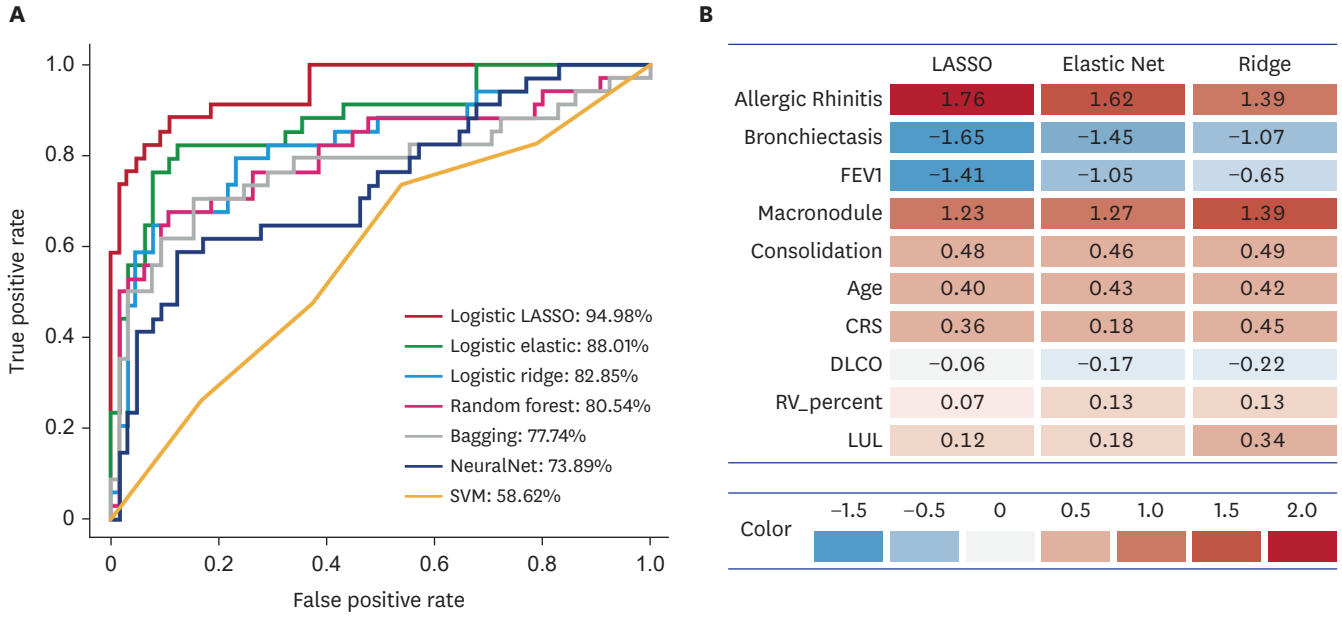

Fig. 1. Comparison of machine learning models and variable coefficients in predicting diffuse panbronchiolitis. (**A**) A comparison of AUCs among seven machine learning models using ROC curves shows that the LASSO model achieved the highest AUC. (**B**) Heat map showing the coefficients of ten variables for diffuse panbronchiolitis across three statistical models (LASSO, Elastic net and Ridge).
SVM = support vector machine, FEV1 = forced expiratory volume in one second, CRS = chronic rhinitis syndrome, DLCO = diffusing capacity of the lung for carbon monoxide, RV = residual volume, LUL = left upper lobe, AUC = area under the ROC curve, ROC = receiver operating characteristic, LASSO = least absolute shrinkage and selection operator.

DISCUSSION

To our knowledge, this is the first study that uses an ML algorithm to assist in the diagnosis of DPB. Our data indicate that allergic rhinitis and macronodules on CT, can be helpful for the early diagnosis of DPB in patients with suspected disease. If a patient shows evidence of allergic rhinitis and macronodules on chest CT, the possibility of DPB should be considered.

DPB research has historically been limited to Western countries because of its genetic predisposition, which is predominantly observed among East Asian populations.[6] Given

the rarity and diagnostic challenges of DPB, minimal progress has been made in research, and the diagnostic criteria have been updated over the decades. Our findings demonstrate that ML analysis of multiple variables can significantly improve early diagnostic performance beyond traditional criteria, addressing this long-standing clinical challenge.

ML approaches play a critical role in various aspects of medicine, including diagnosis, classification, treatment planning, and predictive analytics.[7] Recent studies have demonstrated the increasing utility of ML in various medical fields.

In terms of diagnosis and classification, several recent studies have highlighted the potential of ML. Kim et al.[8] used deep learning to identify atrial fibrillation from single-lead mobile ECGs. Ahn et al.[9] employed ML to identify diagnostic biomarkers for sarcopenia through DNA methylation analysis. Huang et al.[10] utilized ML to classify interstitial lung disease based on plasma proteomics. Further, Kim and Tagkopoulos[11] applied ML in rheumatic disease research for improved diagnosis and classification. Kang et al.[12] used deep learning to differentiate COVID-19 from bacterial pneumonia using chest CT pattern analysis. Additionally, Mun and Cho[13] discussed AI applications for monitoring indoor air quality and its health effects, indirectly aiding respiratory issue diagnosis.

For prognostic evaluation, Park et al.[14] used explainable ML to predict mortality in septic patients. Jeong et al.[15] explored ML approaches for understanding acute kidney injury outcomes. Kwon et al.[16] employed ML models to predict postoperative lung function in patients with lung cancer. Ryu et al.[17] applied ML methods to identify the factors associated with vasomotor symptoms in women, with random forest models yielding the best predictive performance and providing an important decision support system for predicting and managing these symptoms. Currently, ML-based research is being increasingly integrated into routine clinical practice.

We tested ML models, including random forest, EN, ridge regression, bagging, SVM, and NN.[18] Each model has unique characteristics: random forest excels in handling non-linear relationships and interactions between variables, EN combines L1 and L2 regularization for feature selection and multicollinearity management, ridge regression focuses on L2 regularization to reduce overfitting, bagging improves stability by combining predictions from multiple models, SVM works well with high-dimensional data and complex boundaries,[19] and NNs are powerful for capturing non-linear patterns but require large datasets to avoid overfitting. Despite these strengths, the LASSO regression model demonstrated superior predictive accuracy in our study.

Most (> 80%) patients with DPB experience chronic sinusitis, an important factor in the diagnosis of DPB.[1,2] This study showed that allergic rhinitis is a prominent parameter in patients with DPB. Rhinosinusitis is the inflammation and swelling of the nasal lining, with the production of thick mucus that obstructs the paranasal sinuses and eventually allows secondary bacterial overgrowth. It is well known that there is a close relationship between allergic rhinitis and chronic sinusitis.[20] MUC5B, which is closely linked to the pathophysiology of DPB,[21] is also associated with the genetic pathophysiology of allergic rhinitis.[22] IL-8 and TNF-α are also thought to be possible clinical mediators of allergic rhinitis and DPB.[23-26] Therefore, considering the results of this study and the involvement of shared pro-inflammatory mediators, allergic rhinitis may contribute to the diagnosis of DPB. Further studies are required to understand the common immunopathological mechanisms of the upper and lower respiratory tracts.

https://doi.org/10.3346/jkms.2025.40.e311

CT plays a key role in the diagnosis of DPB.[27] Typical high-resolution CT findings of DPB include centrilobular micronodules, sometimes referred to as tree-in-bud pattern.[1,2] If the inflammation of the respiratory bronchioles persists, yellowish sputum and fever become prominent, with respiratory bronchiolar narrowing and intraluminal mucosal plugs leading to the formation of macronodules. Ultimately, the dilatation of the terminal conducting bronchioles resembles that in bronchiectasis with copious purulent sputum and obstructive respiratory functional impairment.[28] Based on the CT imaging features that reflect disease progression, we analyzed various specific image variables such as micronodules, macronodules, cavities, consolidations, and cystic bronchiectasis and used high-resolution CT grading as described by Akira et al.[28] In this study, macronodules were a significant finding in the early diagnosis of DPB. In terms of pathological features, peribronchioles beyond the respiratory bronchioles are observed to be surrounded by intense lymphoplasmacytic inflammatory infiltration, and the proliferation of lymphoid follicles is observed along the airways.[29] These findings may be related to the development of macronodules, which seem to be significant for diagnosis.

The reason for better performance of LASSO regression model is the nature of our dataset. Our sample size was relatively small (n = 100), while the number of independent variables was large (n = 30). This increases the risk of multicollinearity and overfitting, which can compromise the model reliability. LASSO regression addresses these challenges effectively by incorporating L1 regularization, which not only minimizes overfitting but also performs feature selection by shrinking the less important coefficients to zero. Its ability to identify the most relevant predictors while reducing noise likely contributed to its superior performance.

Furthermore, our research team previously applied ML models to datasets of varying sizes and complexities. For example, we developed ML models to predict sleep surgery outcomes using a small sample size (n = 29).[30] We also applied these models to larger datasets, such as COVID-19 (n = 8,070)[31] and diabetes (n = 5,120)[32] data, where the LASSO regression model consistently demonstrated high performance across different sample sizes and contexts. This suggests that the proposed model is robust and adaptable.

Taken together, the application of ML algorithms is effective for early diagnosis of DPB. In this study, we identified two factors that may be important for the early diagnosis: allergic rhinitis and macronodules. However, further studies with larger datasets are warranted.

SUPPLEMENTARY MATERIALS

Supplementary Data 1

TRIPOD checklist for reporting

Supplementary Data 2

Code availability

REFERENCES

1. Homma H, Yamanaka A, Tanimoto S, Tamura M, Chijimatsu Y, Kira S, et al. Diffuse panbronchiolitis. A disease of the transitional zone of the lung. *Chest* 1983;83(1):63-9. PUBMED | CROSSREF

2. Yamanaka A, Saiki S, Tamura S, Saito K. Problems in chronic obstructive bronchial diseases, with special reference to diffuse panbronchiolitis. *Naika* 1969;23(3):442-51. PUBMED
3. Chuang MC, Chou YT, Lin YC, Hsieh MJ, Tsai YH. Diffuse panbronchiolitis-The response and recurrence after erythromycin therapy. *J Formos Med Assoc* 2016;115(10):876-82. PUBMED | CROSSREF
4. Van Calster B, Wynants L. Machine learning in medicine. *N Engl J Med* 2019;380(26):2588-90. PUBMED | CROSSREF
5. Grover S, Mathew J, Bansal A, Singhi SC. Approach to a child with lower airway obstruction and bronchiolitis. *Indian J Pediatr* 2011;78(11):1396-400. PUBMED | CROSSREF
6. Keicho N, Hijikata M. Genetic predisposition to diffuse panbronchiolitis. *Respirology* 2011;16(4):581-8. PUBMED | CROSSREF
7. Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, et al. Artificial intelligence in health care: current applications and issues. *J Korean Med Sci* 2020;35(42):e379. PUBMED | CROSSREF
8. Kim J, Lee SJ, Ko B, Lee M, Lee YS, Lee KH. Identification of atrial fibrillation with single-lead mobile ECG during normal sinus rhythm using deep learning. *J Korean Med Sci* 2024;39(5):e56. PUBMED | CROSSREF
9. Ahn S, Sung Y, Song W. Machine learning-based identification of diagnostic biomarkers for Korean male sarcopenia through integrative DNA methylation and methylation risk score: from the Korean genomic epidemiology study (KoGES). *J Korean Med Sci* 2024;39(26):e200. PUBMED | CROSSREF
10. Huang Y, Ma SF, Oldham JM, Adegunsoye A, Zhu D, Murray S, et al. Machine learning of plasma proteomics classifies diagnosis of interstitial lung disease. *Am J Respir Crit Care Med* 2024;210(4):444-54. PUBMED | CROSSREF
11. Kim KJ, Tagkopoulos I. Application of machine learning in rheumatic disease research. *Korean J Intern Med* 2019;34(4):708-22. PUBMED | CROSSREF
12. Kang M, Hong KS, Chikontwe P, Luna M, Jang JG, Park J, et al. Quantitative assessment of chest CT patterns in COVID-19 and bacterial pneumonia patients: a deep learning perspective. *J Korean Med Sci* 2021;36(5):e46. PUBMED | CROSSREF
13. Mun E, Cho J. Review of internet of things-based artificial intelligence analysis method through real-time indoor air quality and health effect monitoring: focusing on indoor air pollution that are harmful to the respiratory organ. *Tuberc Respir Dis (Seoul)* 2023;86(1):23-32. PUBMED | CROSSREF
14. Park SW, Yeo NY, Kang S, Ha T, Kim TH, Lee D, et al. Early prediction of mortality for septic patients visiting emergency room based on explainable machine learning: a real-world multicenter study. *J Korean Med Sci* 2024;39(5):e53. PUBMED | CROSSREF
15. Jeong I, Cho NJ, Ahn SJ, Lee H, Gil HW. Machine learning approaches toward an understanding of acute kidney injury: current trends and future directions. *Korean J Intern Med* 2024;39(6):882-97. PUBMED | CROSSREF
16. Kwon OB, Han S, Lee HY, Kang HS, Kim SK, Kim JS, et al. Prediction of postoperative lung function in lung cancer patients using machine learning models. *Tuberc Respir Dis (Seoul)* 2023;86(3):203-15. PUBMED | CROSSREF
17. Ryu KJ, Yi KW, Kim YJ, Shin JH, Hur JY, Kim T, et al. Machine learning approaches to identify factors associated with women's vasomotor symptoms using general hospital Data. *J Korean Med Sci* 2021;36(17):e122. PUBMED | CROSSREF
18. Wu Y, Zhu W, Wang J, Liu L, Zhang W, Wang Y, et al. Using machine learning for mortality prediction and risk stratification in atezolizumab-treated cancer patients: Integrative analysis of eight clinical trials. *Cancer Med* 2023;12(3):3744-57. PUBMED | CROSSREF
19. Guido R, Ferrisi S, Lofaro D, Conforti D. An overview on the advancements of support vector machine models in healthcare applications: a review. *Information* 2024;15(4):235. CROSSREF
20. Li S, Zhao CJ, Hua HL, Deng YQ, Tao ZZ. The association between allergy and sinusitis: a cross-sectional study based on NHANES 2005-2006. *Allergy Asthma Clin Immunol* 2021;17(1):135. PUBMED | CROSSREF
21. Kamio K, Matsushita I, Hijikata M, Kobashi Y, Tanaka G, Nakata K, et al. Promoter analysis and aberrant expression of the MUC5B gene in diffuse panbronchiolitis. *Am J Respir Crit Care Med* 2005;171(9):949-57. PUBMED | CROSSREF
22. Zhang Y, Huang Y, Chen WX, Xu ZM. Identification of key genes in allergic rhinitis by bioinformatics analysis. *J Int Med Res* 2021;49(7):3000605211029521. PUBMED | CROSSREF
23. Emi M, Keicho N, Tokunaga K, Katsumata H, Souma S, Nakata K, et al. Association of diffuse panbronchiolitis with microsatellite polymorphism of the human interleukin 8 (IL-8) gene. *J Hum Genet* 1999;44(3):169-72. PUBMED | CROSSREF
24. Keicho N, Emi M, Nakata K, Taguchi Y, Azuma A, Tokunaga K, et al. Promoter variation of tumour necrosis factor-alpha gene: possible high risk for chronic bronchitis but not diffuse panbronchiolitis. *Respir Med* 1999;93(10):752-3. PUBMED | CROSSREF

25. Gosset P, Tillie-Leblond I, Malaquin F, Durieu J, Wallaert B, Tonnel AB. Interleukin-8 secretion in patients with allergic rhinitis after an allergen challenge: interleukin-8 is not the main chemotactic factor present in nasal lavages. *Clin Exp Allergy* 1997;27(4):379-88. PUBMED | CROSSREF

26. Iwasaki M, Saito K, Takemura M, Sekikawa K, Fujii H, Yamada Y, et al. TNF-alpha contributes to the development of allergic rhinitis in mice. *J Allergy Clin Immunol* 2003;112(1):134-40. PUBMED | CROSSREF

27. Hansell DM. Small airways diseases: detection and insights with computed tomography. *Eur Respir J* 2001;17(6):1294-313. PUBMED | CROSSREF

28. Akira M, Kitatani F, Lee YS, Kita N, Yamamoto S, Higashihara T, et al. Diffuse panbronchiolitis: evaluation with high-resolution CT. *Radiology* 1988;168(2):433-8. PUBMED | CROSSREF

29. Poletti V, Casoni G, Chilosi M, Zompatori M. Diffuse panbronchiolitis. *Eur Respir J* 2006;28(4):862-71. PUBMED | CROSSREF

30. Yang SJ, Kim JS, Chung SK, Song YY. Machine learning-based model for prediction of outcomes in palatal surgery for obstructive sleep apnoea. *Clin Otolaryngol* 2021;46(6):1242-6. PUBMED | CROSSREF

31. Kim DH, Kim MG, Yang SJ, Lee EJ, Yeom SW, You YS, et al. Influenza and anosmia: Important prediction factors for severity and death of COVID-19. *J Infect* 2021;83(5):e10-3. PUBMED | CROSSREF

32. Lee KA, Kim JS, Kim YJ, Goak IS, Jin HY, Park S, et al. A machine learning-based prediction model for diabetic kidney disease in Korean patients with type 2 diabetes mellitus. *J Clin Med* 2025;14(6):2065. PUBMED | CROSSREF