

Development of a deep learning model to predict smoking status in patients with chronic obstructive pulmonary disease: A secondary analysis of cross-sectional national survey

Sudarshan Pant¹ , Hyung Jeong Yang¹, Sehyun Cho², EuiJeong Ryu³
and Ja Yun Choi^{2,4} 

Abstract

Objective: This study aims to develop and validate a deep learning model to predict smoking status in patients with chronic obstructive pulmonary disease (COPD) using data from a national survey.

Methods: Data from the Korea National Health and Nutrition Examination Survey (2007–2018) were used to extract 5466 COPD-eligible cases. The data collection involved demographic, behavioral, and clinical variables, including 21 predictors such as age, sex, and pulmonary function test results. The dependent variable, smoking status, was categorized as smoker or nonsmoker. A residual neural network (ResNN) model was developed and compared with five machine learning algorithms (random forest, decision tree, Gaussian Naive Bayes, K-nearest neighbor, and AdaBoost) and two deep learning models (multilayer perceptron and TabNet). Internal validation was performed using five-fold cross-validation, and model performance was evaluated using the area under the receiver operating characteristic (AUROC) curve, sensitivity, specificity, and F1-score.

Results: The ResNN achieved an AUROC, sensitivity, specificity, and F1-score of 0.73, 70.1%, 75.2%, and 0.67, respectively, outperforming previous machine learning and deep learning models in predicting smoking status in patients with COPD. Explainable artificial intelligence (Shapley additive explanations) identified key predictors, including sex, age, and perceived health status.

Conclusion: This deep learning model accurately predicts smoking status in patients with COPD, offering potential as a decision-support tool to detect high-risk persistent smokers for targeted interventions. Future studies should focus on external validation and incorporate additional behavioral and psychological variables to improve its generalizability and performance.

Keywords

Classification, deep learning, machine learning, chronic obstructive pulmonary disease, smoking

Received: 11 September 2024; accepted: 24 March 2025

Introduction

Chronic obstructive pulmonary disease (COPD), a major public health concern, is projected to cost \$4.33 trillion globally from 2020 to 2050.¹ Its global burden increased by 25.7% in disability-adjusted life years from 1990 to 2019.² In the United States, continued smoking is expected to significantly boost medical costs and disability-adjusted life years losses from 2019 to 2038.³ Often referred to as a “smoker’s disease,” COPD is primarily caused by smoking, with approximately 75% of patients having a smoking history.⁴ COPD severity positively correlates with age and

¹Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Republic of Korea

²College of Nursing, Chonnam National University, Gwangju, Republic of Korea

³Department of Nursing, Dongshin University, Naju, Republic of Korea

⁴College of Nursing, Chonnam National University, Chonnam Research Institute of Nursing Science, Gwangju, Republic of Korea

Corresponding author:

Ja Yun Choi, College of Nursing, Chonnam National University, Chonnam Research Institute of Nursing Science, Gwangju 61469, Republic of Korea.
Email: choijy@jnu.ac.kr



smoking duration.⁵ Given these findings, smoking cessation is essential for self-management and a key consideration in developing COPD treatment programs.⁶ Moreover, as smoking contributes to disease progression in patients with COPD, quitting is the only effective way to stop its progression.⁷ However, 33.6% to 69.5% of patients with COPD continue smoking despite their diagnosis.^{8,9} Thus, understanding and predicting their smoking status is essential for effective disease management, personalized treatment strategies, and targeted smoking cessation interventions.¹⁰

An accurate model is vital for predicting persistent smoking in patients with COPD. Machine learning techniques demonstrate effectiveness in various medical applications, including disease diagnosis, prognosis, and treatment prediction.¹¹ These techniques enable the effective analysis of large healthcare datasets to extract meaningful insights and improve predictive accuracy. Moreover, deep learning models show excellent performance across various healthcare domains,¹² and several studies have investigated their use in predicting smoking behavior. For instance, a real-time multiclass classification model using data obtained from wearable devices achieved 93.1% accuracy in detecting smoking activity.¹³ In another study, machine learning models were used to predict smoking behavior based on genomic data.¹⁴ Similarly, Wang et al.¹⁵ leveraged isoform-level ribonucleic acid sequencing data to improve smoking status predictions. Additionally, natural language processing was used to determine smoking status from unstructured electronic health records of patients with suspected lung cancer in Denmark, demonstrating its ability to convert unstructured text into structured data.¹⁶

Healthcare data sources include electronic health records, wearable devices, and genetic databases.¹⁷ Healthcare data analytics can minimize treatment costs, prevent diseases, slow disease progression, and improve quality of life, ultimately leading to life-saving outcomes. To achieve these benefits, the data should incorporate modifiable lifestyles relevant to health and wellness.¹⁸ Although smoking habits and biomarkers, such as nicotine levels, help determine current smoking status, identifying patterns linked to a higher likelihood of smoking cessation nonadherence remains challenging for targeted management and policy interventions. Nonadherence to self-management in patients with COPD includes failure to quit smoking, inconsistent medication use, and irregular clinical follow-ups—behaviors influenced by personal, socioeconomic, disease-related, functional, treatment-related, health system-related, and environmental factors.¹⁹ However, research on predicting persistent smoking following disease onset using individual, environmental, socioeconomic, and behavioral data remains limited. Since smoking cessation is the primary recommendation, accurately predicting persistent smoking after a COPD diagnosis can identify high-risk patients who may also struggle with other essential self-management strategies.

Exploring the potential of machine learning and deep learning models in predicting smoking status in patients with COPD is crucial for identifying those at risk of self-management nonadherence. Therefore, this study aims to develop a model to predict smoking status in patients with COPD and to compare this model with machine learning and deep learning models. This model would enable targeted advice, financial assistance for smoking cessation, and enrollment in cessation programs at the point of COPD diagnosis, thereby addressing health disparities among vulnerable patients with COPD.

Methods

Study design

In this study, a secondary analysis of data from a nationally representative cross-sectional survey was conducted.

Data sources and participant selection

The Korea National Health and Nutrition Examination Survey (KNHANES)²⁰ is a cross-sectional national survey designed to investigate key aspects of health and nutrition in Korea, including environmental factors, health status, medical care, welfare, and dietary habits. Initiated in 1998 and conducted annually since 2007, its primary objective is to generate statistically representative data, providing a critical foundation for formulating health policies. These policies include setting and assessing objectives for national health promotion plans and designing targeted health initiatives.²⁰ Accordingly, data from the KNHANES were considered suitable for predicting persistent smoking behaviors relevant to disease management and quality of life in patients with COPD. Researchers (CS and REJ) extracted data from the KNHANES²⁰ between 2007 and 2018, the most recent available period. Data from 2019 were excluded due to the COVID-19 pandemic and modifications in quality-of-life assessment tools, which hindered the collection of critical measurements.

COPD cases were identified based on two criteria: self-reported physician diagnosis and pulmonary function test results. The pulmonary function criterion was defined as a forced expiratory volume in 1 second (FEV1) to forced vital capacity (FVC) ratio of $\leq 0.7\%$. Respondents who reported a physician-diagnosed COPD also exhibited pulmonary function test results consistent with those meeting the COPD diagnostic criteria. Cases were excluded if data on the dependent variable, current smoking status, were missing.

Variables

Dependent variable. In the KNHANES, smoking status was originally categorized into four groups: daily smokers,

occasional smokers, ex-smokers, and nonsmokers. Daily smokers were defined as individuals who smoked ≥ 1 cigarette per day, while occasional smokers were those who smoked intermittently but did not meet the criteria for daily smoking. Based on a previous study²¹ using KNHANES data, daily and occasional smokers were grouped as “smokers,” while individuals who had previously smoked or had never smoked were classified as “nonsmokers.” Using this binary classification system, the values of “1” and “0” were assigned to the “smoker” and “nonsmoker” groups, respectively.

Predictor variables. Predictor variables were selected based on factors associated with nonadherence to COPD self-management identified in a previous study.¹⁹ The predictor variables included 21 features categorized as continuous, categorical, and ordinal variables. Continuous variables included age, number of household members, quality of life, average sleep time, and walking days per week. Categorical variables included sex, occupation, educational level, economic activity, marital status, pulmonary function test results, influenza vaccination, diabetes prevalence, physician-diagnosed lung cancer, physician-diagnosed depression, physician-diagnosed hypertension, depressive symptoms for > 2 consecutive weeks, body weight control for a year, and body weight changes for a year. Additionally, ordinal variables included perceived stress level and perceived health status (Supplementary Table S1).

Research process

This study involved various steps, including feature selection (classification and optimization), data preprocessing, model development, model validation and evaluation, and result interpretation (Figure 1).

Feature selection process

The RandomForest feature selection method was used as a foundation for identifying the most important features and enhancing the interpretability of the results. This approach enabled a focus on the most influential variables and provided deeper insights into the underlying relationships within the data.²² In detail, we transformed continuous variables into categorical ones and simplified the original feature categories to assess whether the original or modified variables enhanced explanatory power (Supplementary Figure S1). Thus, the RandomForest classifier was used to identify the most relevant features by excluding two original and five modified features—occupation, education level, categorical average sleep duration, modified perceived health status, modified walking frequency per week, modified perceived stress level, and modified number of household members—based on their importance scores (Supplementary Figure S1).

Data preprocessing

Before training the models, data preprocessing was conducted, including variable modification, handling missing values, one-hot encoding of categorical variables, and normalizing continuous variables to ensure consistency and comparability across features. Missing values were imputed by replacing them with the most frequent value of each feature. Since the proportion of missing entries ranged from 0.1% to 18.9%—well below the 30% threshold considered negligible²³—this imputation method was chosen for its simplicity, minimal computational requirements, and robustness. This approach ensured consistency while preventing unnecessary complexity in the preprocessing pipeline. We employed the scikit-learn library for one-hot encoding and normalization.²⁴

The dataset was randomly divided into training, validation, and test sets with ratios of 70%, 15%, and 15%, respectively, using the “train_test_split” function from scikit-learn.²⁴

Model development

To predict smoking status, we developed a residual neural network (ResNN), a deep learning model incorporating residual connections. The feedforward neural network architecture, enhanced with residual connections, improved information flow and reduced the vanishing gradient problem. This design choice enhanced the ability of the model to capture complex patterns more effectively.²⁵ The proposed model comprised an input layer with 21 nodes, three fully connected hidden layers with 512 nodes each, and an output layer utilizing a Softmax function for classification. The rectified linear unit (ReLU) served as the activation function in the hidden layers. A residual connection was incorporated into the first hidden layer to bypass specific learned features directly from the input, further alleviating the vanishing gradient problem and enhancing the information flow. For hyperparameter tuning, we selected the AdamW optimizer with a learning rate of 0.0001 due to its effectiveness in handling sparse gradients and ensuring stable optimization. Additionally, to address the class imbalance, we applied focal loss and fine-tuned the learning rate and optimizer settings during the model training process to maximize performance. Figure 2 displays the architecture of the proposed classification model. To improve reproducibility, the trained model weights have been made publicly available on GitHub.²⁶

Additionally, we conducted experiments using five machine learning and two deep learning models. The machine learning models included random forest,²⁷ decision trees,²⁸ Gaussian Naive Bayes,²⁹ K-nearest neighbor,³⁰ and AdaBoost,³¹ selected for their performances in previous healthcare prediction studies. The deep learning models were selected for their ability to capture complex patterns in high-dimensional data.³² The multilayer perceptron

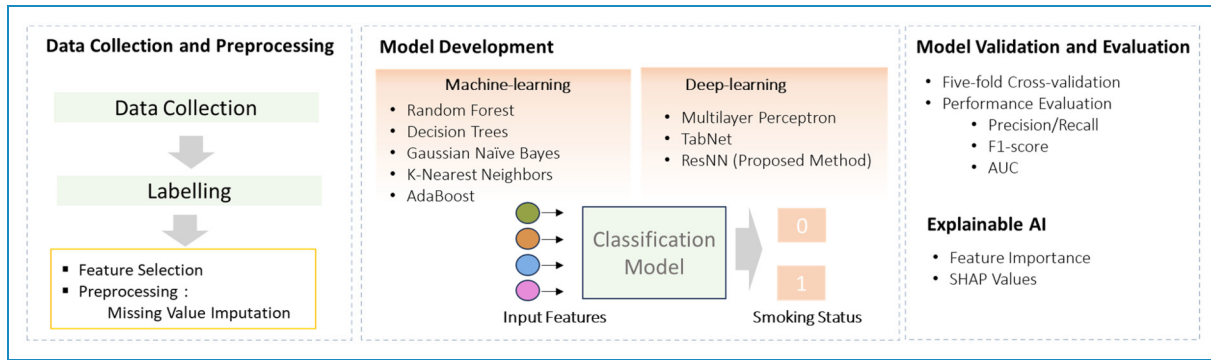


Figure 1. Flowchart illustrating the prediction of smoking status in patients with chronic obstructive pulmonary disease. *Abbreviations:* AI, artificial intelligence; AUC, area under the curve; ResNN, residual neural network; SHAP, Shapley additive explanations; TabNet, tabular neural network.

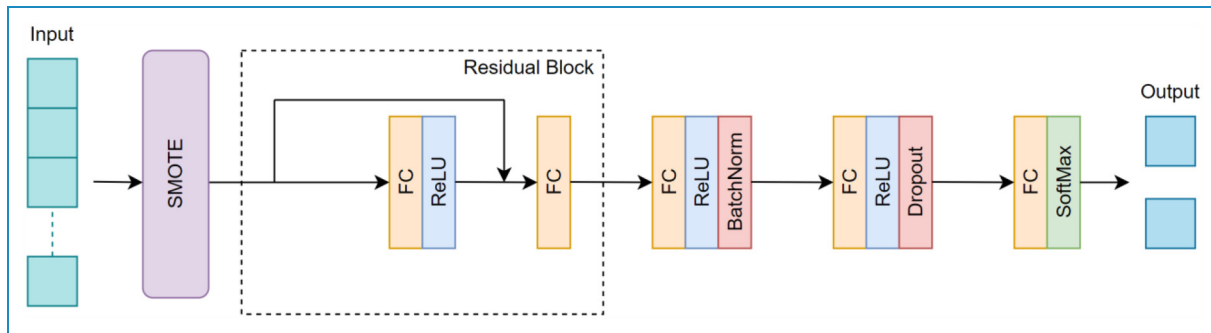


Figure 2. Overview of the proposed classification model for predicting smoking status. *Abbreviations:* BatchNorm, batch normalization; FC, fully connected layer; ReLU, rectified linear unit; SMOTE, synthetic minority over-sampling technique.

(MLP) model comprised diverse hidden layers with ReLU activation functions and a sigmoidal activation function in the output layer. Dropout regularization was applied to mitigate overfitting. The MLP model was trained using the same train-test split and hyperparameter tuning approach as the traditional machine learning models. For comparison, we employed TabNet, a deep learning architecture optimized for tabular data.³³ The TabNet model was configured with multiple decision steps, feature selection per step was determined, and the learning rate was optimized using early stopping.

Model validation and performance evaluation

Model validation. Five-fold cross-validation was utilized to evaluate the performance of the proposed model in predicting smoking status after a COPD diagnosis. Supplementary Figure S2 illustrates that the model showed consistent performance across all five folds, indicating strong generalizability. Supplementary Table S2 presents the hyperparameters for classifiers, selected using the Grid Search method,³⁴ which resulted in the highest area under the receiver operating characteristic curve (AUROC). These hyperparameters include RandomForest (n_estimators = 500), DecisionTree

(max_depth = 5), Gaussian Naïve Bayes (regularization parameter C = 0.1, gamma = 1), K-nearest neighbors (n_neighbors = 7), AdaBoost (n_estimators = 100), and MLP (regularization parameter alpha = 1, max_iterations = 2000).

Model performance evaluation. The receiver operating characteristic (ROC) curve was generated by plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) across various thresholds.³⁵ The AUROC was computed to evaluate the predictive power of the model, which achieved a high AUROC score of 0.72 during five-fold cross-validation. To minimize the influence of random sample partitions, we calculated the arithmetic mean of the performance metrics across all five folds.

Interpretable artificial intelligence

To enhance the interpretability of the proposed model, we used the Shapley additive explanations (SHAPs) framework, a game-theoretic approach for attributing the contribution of each feature to the predicted outcome.³⁶ SHAP values were computed for each feature in our proposed model, which was used to predict smoking status in patients with COPD. These values were used to quantify the marginal

contribution of individual features while accounting for interactions and dependencies, enhancing the interpretability of the deep learning model and offering a clearer understanding of the factors influencing smoking status predictions in patients with COPD. In the SHAP summary plot, the x-axis represents the SHAP values, indicating the direction and magnitude of the contribution of each feature to the prediction. Positive SHAP values (blue) increase the output of the model, while negative values (red) decrease it. Features are ranked by significance along the y-axis. Each point represents an observation, with its color reflecting the feature value across the x-axis, illustrating the variability in feature effects.

Ethical consideration

The KNHANES was approved annually by the Research Ethics Review Committee of the Korea Disease Control and Prevention Agency. However, the surveys conducted in 2015, 2016, and 2017 were exempt from ethical review, as they were directly conducted by the Korean government for public welfare purposes. Additionally, this study was approved by the Institutional Review Board of C National University Hospital (CNUH-2022-232).

Results

Participants and data flow characteristics

Among the 89,512 cases, 5503 responders met the COPD criteria from the 84,009 eligible participants. After excluding cases with missing data on current smoking status ($n=37$), the final sample included 5466 responders (Supplementary Figure S3).

The 5466 responders had a mean age of 64.81 ± 10.72 years, with 3921 being male (71.7%). Of the responders, 2100 (38.4%) were current smokers, 2426 (44.4%) had an education level of elementary school or lower, and >50% were employed. Most responders were married (98.0%) and lived with family members (88.0%). Influenza vaccination had been administered to 3495 (63.9%) responders, and 3141 (57.5%) had normal blood glucose levels (<126 mg/dL). Additionally, the prevalence of physician-diagnosed conditions among responders varied: 99.9%, 96.9%, and 66.8% did not have lung cancer, depression, and hypertension, respectively. High-stress levels were reported by 1444 (26.4%), while 537 (9.8%) experienced depressive symptoms over 2 weeks. Approximately 1589 (29.1%) responders rated their health status as good. The average EQ-5D score was 0.89 ± 0.12 . Additionally, 3678 (67.8%) engaged in daily exercise, while 2719 (49.7%) had never attempted to control their weight. Overall, 4026 (73.7%) responders indicated no weight change over the past year (Supplementary Table S3).

Significant differences were observed between smokers and nonsmokers across various variables: age ($t = -11.61, p < .001$),

sex ($X^2 = 697.24, p < .001$), occupation ($X^2 = 279.74, p < .001$), education level (elementary school or below) ($X^2 = 64.92, p < .001$), economic activity ($X^2 = 71.25, p < .001$), marital status ($X^2 = 12.90, p < .001$), household size ($X^2 = 34.34, p < .001$), influenza vaccination status ($X^2 = 96.65, p < .001$), pulmonary function test results ($X^2 = 94.61, p < .001$), physician-diagnosed lung cancer ($X^2 = 6.85, p = .009$), physician-diagnosed hypertension ($X^2 = 50.25, p < .001$), walking frequency per week ($X^2 = 213.23, p < .001$), body weight control for a year ($X^2 = 68.16, p < .001$), body weight change over a year ($X^2 = 19.77, p < .001$), depressive symptoms for > 2 consecutive weeks ($X^2 = 5.76, p = .016$), and perceived health status ($X^2 = 8.80, p = .066$). However, no significant differences were observed in diabetes prevalence ($X^2 = 3.43, p = .180$), physician-diagnosed depression ($X^2 = 2.75, p = .097$), quality of life scores ($t = 0.58, p = .559$), perceived stress levels ($X^2 = 8.80, p = .066$), and average sleep duration ($t = -0.06, p = .957$).

Performance across different models

Table 1 lists the AUROCs, true positive rate (sensitivity), false positive rate (1-specificity), and F1 scores for various models. The proposed ResNN outperformed previously established machine learning and deep learning models across four key metrics: AUROC (.73), sensitivity (.70), specificity (.75), and F1-score (.67), in predicting smoking status following a COPD diagnosis. The adjusted accuracy, which accounts for class prevalence (38.4%), provides a more reliable measure of classifier performance in an imbalanced dataset. Table 1 shows that, among traditional classifiers, the Random Forest and AdaBoost classifiers achieved the highest adjusted accuracy (0.70). The proposed ResNN model achieved an adjusted accuracy of 0.73, indicating its competitive performance compared to those of traditional models and other deep learning approaches, such as TabNet (0.70) and MLPClassifier (0.72). These findings underscore the robustness of the ResNN model in mitigating the challenges associated with the imbalance of the dataset.

Figure 3 shows the ROC curves generated for each classifier. Despite the class imbalance, our proposed model, which incorporates advanced imbalance-handling techniques (focal loss), outperformed the traditional classifiers that rely on less effective class-weighting strategies. This highlights its superior predictive accuracy and ability to handle class imbalance.

Figure 4 presents the confusion matrices for all the classifiers used in this study, providing a comparison of their performance on the held-out test set. The results showed that the proposed ResNN method outperformed other classifiers.

Features importance

SHAP values were visualized using summary plots, which displayed the feature importance and their influence on the

Table 1. Comparison of classification performance metrics with 95% bootstrap confidence intervals (1000 iterations) for different models.

Classifier	AUROC (95% CI)	Recall (sensitivity)	Specificity	F1-score (95% CI)	ACC _{Prev} (95% CI)
RandomForestClassifier	.66 [0.63, 0.70]	.51 [0.46, 0.56]	.82 [0.79, 0.85]	.57 [0.52, 0.62]	.70 [0.67, 0.73]
DecisionTreeClassifier	.67 [0.64, 0.70]	.62 [0.57, 0.67]	.72 [0.68, 0.76]	.60 [0.56, 0.65]	.68 [0.65, 0.72]
GaussianNB	.68 [0.64, 0.71]	.81 [0.77, 0.86]	.54 [0.50, 0.58]	.64 [0.59, 0.67]	.64 [0.61, 0.67]
KNeighborsClassifier	.58 [0.55, 0.62]	.42 [0.36, 0.47]	.75 [0.71, 0.78]	.46 [0.40, 0.51]	.62 [0.59, 0.65]
AdaBoostClassifier	.68 [0.64, 0.71]	.55 [0.49, 0.61]	.80 [0.77, 0.84]	.59 [0.54, 0.64]	.70 [0.67, 0.73]
MLPClassifier	.70 [0.67, 0.73]	.63 [0.58, 0.69]	.77 [0.73, 0.80]	.63 [0.59, 0.68]	.72 [0.69, 0.75]
TabNet	.69 [0.65, 0.72]	.62 [0.56, 0.67]	.76 [0.72, 0.79]	.61 [0.57, 0.66]	.70 [0.67, 0.73]
ResNN (Proposed)	.73 [0.70, 0.76]	.70 [0.65, 0.75]	.75 [0.71, 0.79]	.67 [0.63, 0.71]	.73 [0.70, 0.76]

Abbreviations: AUROC: area under the receiver operating characteristic curve; GaussianNB: Gaussian Naive Bayes; MLPClassifier: multilayer perceptron classifier; ResNN: residual neural network; TabNet: tabular neural network; ACC_{Prev}: adjusted accuracy for prevalence 38.4%.

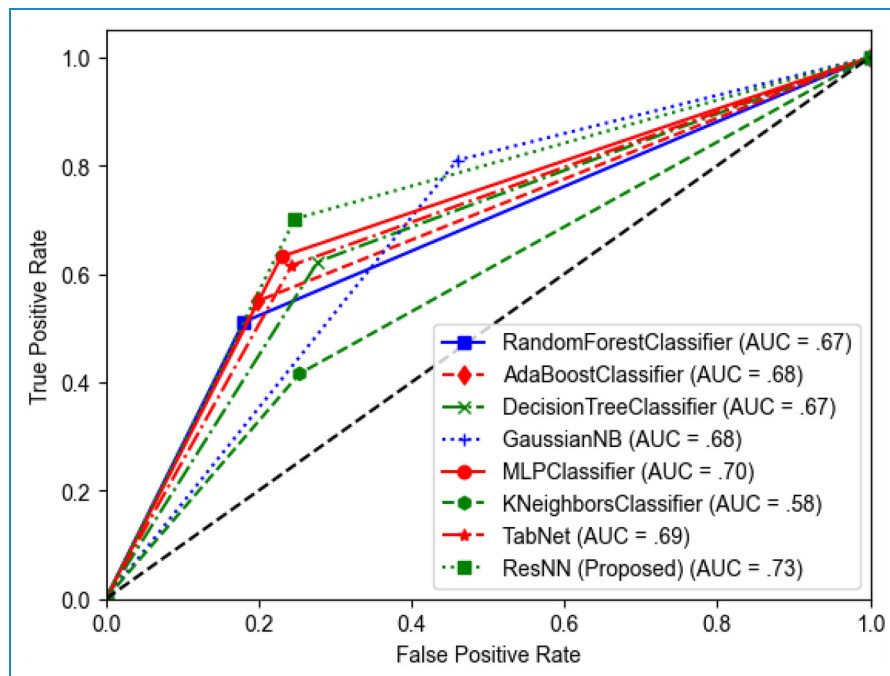


Figure 3. Comparison of receiver operating characteristic curves across various classification models. Abbreviations: AUC, area under the curve; GaussianNB, Gaussian Naïve Bayes; KNeighborsClassifier, K-nearest neighbors classifier; MLPClassifier, multilayer perceptron classifier; ResNN, residual neural network; TabNet, tabular neural network.

predicted smoking status (Figure 5). The most influential factors included sex, age, body weight control for a year, and walking frequency per week. The plot showed that male patients with COPD were more inclined to be classified as smokers. In this study, age exhibited a wider spread of SHAP values on both sides of the plot, suggesting a bidirectional influence on smoking status predictions.

Specifically, younger individuals were associated with a higher probability of being classified as smokers, while older individuals had an increased probability of being classified as nonsmokers. Body weight control over a year also exhibited a bidirectional influence, but with red and blue colors appearing oppositely in age. This suggests that individuals who did not manage their body weight were more

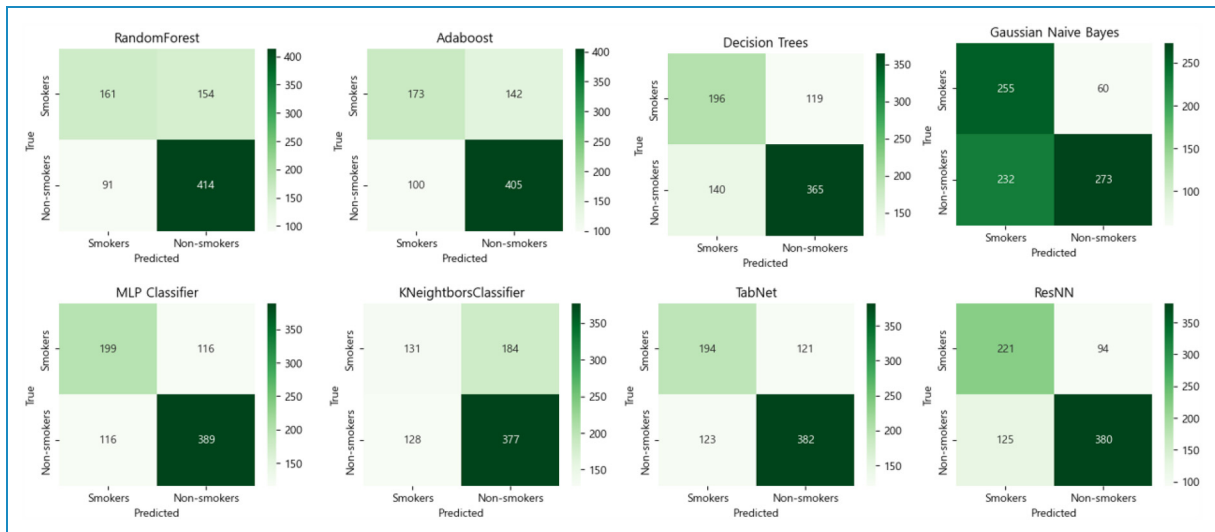


Figure 4. Confusion matrices for the evaluated classifiers on the held-out test set.

likely to be classified as smokers. Additionally, pulmonary function test results played a critical role in the decision-making process of our model. The longer tail extending toward the positive side (colored in red) in the SHAP value plot indicates that individuals with inconclusive test results tend to be classified as smokers.

Discussion

A report by the United States Centers for Disease Control and Prevention on adult smoking cessation behaviors shows that most adult cigarette smokers express a desire to quit. However, only 7.5% successfully achieved smoking cessation in 2018.³⁷ Moreover, individuals with COPD have historically attempted to quit at higher rates than those without COPD. Despite this, their recent successful cessation rates remain similar, and they exhibit a lower lifetime quit ratio.³⁸ These findings highlight the significance of efficiently predicting persistent smokers with COPD and detecting future smoking behaviors based on similar patterns. Nonadherence to smoking cessation is a hazardous behavior for individuals with COPD. Although smoking habits and biomarker analyses can determine the current smoking status of an individual, adherence to other self-management domains—such as medication use and symptom monitoring—may also reflect the efforts to quit smoking, which is a primary aspect of COPD self-management.

This study aims to develop a model for predicting smoking status following a COPD diagnosis and compare its performance with five machine learning and two deep learning models. Predictive artificial intelligence is widely used in healthcare to forecast future events or behaviors. For instance, studies were conducted to predict future events, including COPD prevalence,³⁴ the progression of COPD,³⁹ and medication adherence in noncommunicable

diseases.⁴⁰ Key applications of predictive artificial intelligence include diagnosis and treatment recommendations, patient engagement and adherence, and administrative functions.⁴¹ However, predicting the adherence of patients to healthy behaviors remains challenging due to the complexity of these behaviors.⁴² A systematic literature review reports that machine and deep learning models were used to predict medication adherence in patients with noncommunicable diseases with 56–93% of AUC, with most studies showing confusion matrix-linked performance around 70%.⁴⁰ Therefore, studies on predicting adherence to healthy behaviors—including smoking cessation, regular exercise, and ideal body weight maintenance—remain limited. Conversely, adherence to self-management is a cost-effective strategy for reducing personal and societal burdens in patients with chronic diseases.¹⁰

In this study, the performances of our proposed and comparative models were relatively lower than those of models reported in previous studies on predicting healthy behaviors.^{43,44} However, models with a performance level of ≥ 0.7 are generally considered acceptable.⁴⁵ A previous study developed a predictive model for estimating healthy life years without activity limitations, achieving a significantly high AUROC of .85 using the extreme gradient boosting classifier.⁴⁵ Additionally, 14 features were employed to predict medical appointment no-shows, yielding an AUROC $> .8$ using AdaBoost and decision tree algorithms based on a dataset from the Kaggle database.⁴³ The random forest algorithm also demonstrates superior predictive performance, achieving an AUROC of 93% for medication adherence and 96% for healthcare-seeking behaviors among mothers.⁴⁶

In this study, AdaBoost and random forest demonstrated lower predictive performance. This may be due to model limitations and the variables used. Despite efforts to address the imbalance between smokers and nonsmokers, these

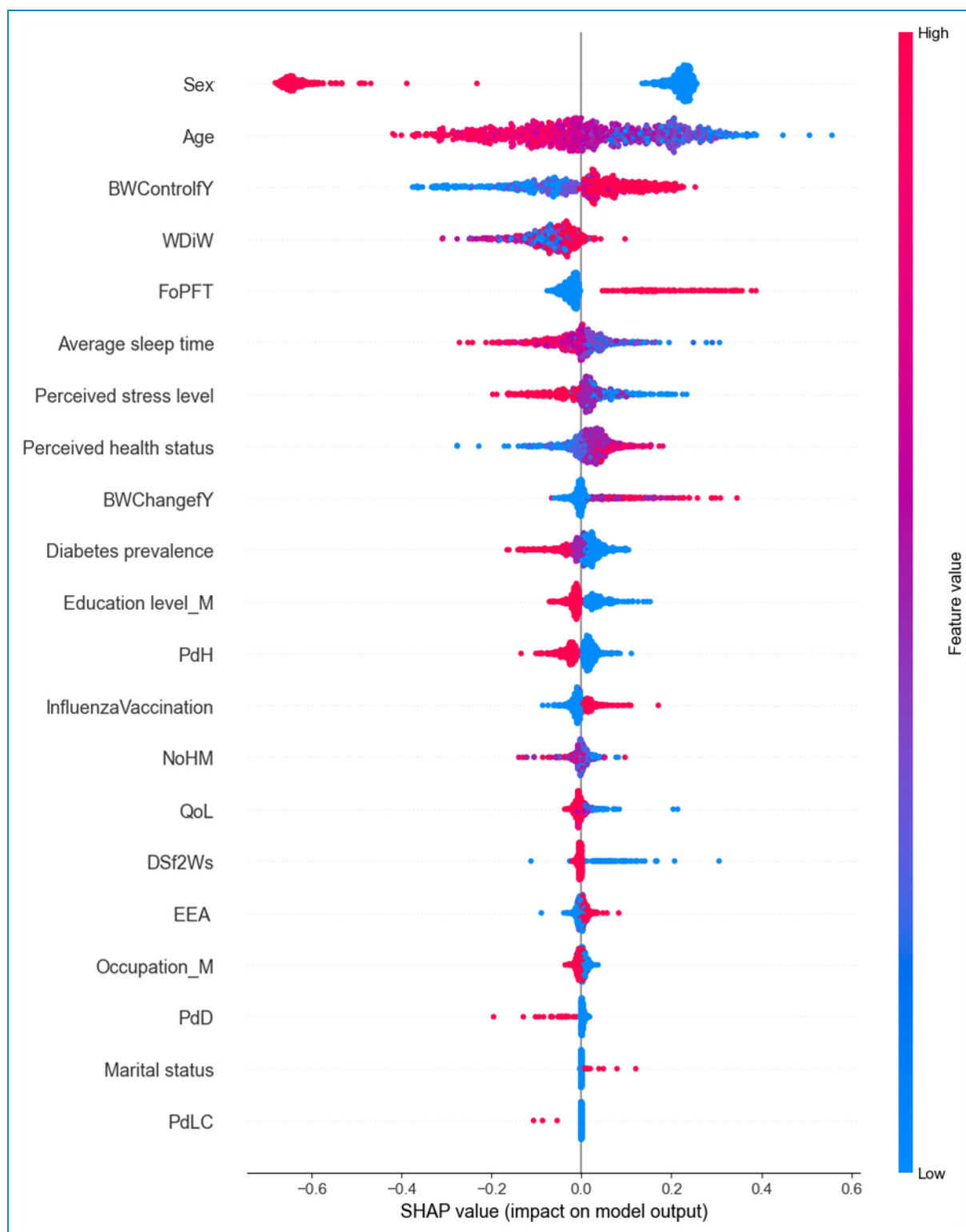


Figure 5. Shapley additive explanations analysis of the features based on the proposed model. *Abbreviations:* BWChangefY, body weight change for a year; BWControlfY, body weight control for a year; DSf2Ws, depressive symptoms for more than 2 consecutive weeks; Educationallevel_M, educationallevel_modified; EEA, economic engagement activity; FoPFT, findings of pulmonary function test; NoHM, number of household members; Occupation_M, occupation_modified; PdD, physician-diagnosed depression; PdH, physician-diagnosed hypertension; PdLC, physician-diagnosed lung cancer; QoL, quality of life; SHAP, Shapley additive explanations; WDiW, walking days in a week.

measures were insufficient to fully mitigate the imbalance in the dependent variable.⁴⁷ Additionally, the inherent difficulty of smoking cessation among patients with COPD probably contributed to the relatively low predictive accuracy of our model. For complex healthcare challenges such as smoking cessation, deep learning models may offer advantages over traditional machine learning models. A ResNN uses residual connections to overcome the vanishing gradient problem in deep neural networks, enabling efficient learning even in deeper architectures.⁴⁸ Therefore, the model can be effective at learning hidden patterns in complex data, making it well-suited for healthcare data predictions, where various variables interact. Although the ResNN achieved the highest performance in this study, its predictive accuracy for persistent smoking remains relatively low.

A gap exists between the intention of smokers to quit and their actual cessation outcomes.⁴⁹ Furthermore, one possible reason for the relatively low-model performances is the absence of key predictive features identified in previous studies, such as willpower and support systems for smoking cessation.⁵⁰ These factors were not included in the KNHANES. Further studies should identify features that can enhance the performance of models for prospective patients with COPD by exploring additional behavioral or psychological variables with data from electronic health records.

Smoking cessation is challenging owing to the complex factors involved, such as nicotine addiction, lifestyle habits, and unsuccessful cessation attempts.⁵¹ We estimated the effect of various features on model performances using permutation importance and evaluated their influence on model output through SHAP values. The utilization of the SHAP framework facilitated the interpretability of our model in predicting smoking status in patients with COPD, enabling the development of targeted interventions and personalized treatment strategies. Features such as sex, age, body weight control over a year, and walking frequency per week contributed to the classification of performances. These features can be divided into modifiable or nonmodifiable variables. For instance, nonmodifiable variables include age, sex, and pulmonary function test results, while modifiable variables encompass body weight control per year, walking frequency per week, average sleep duration, and perceived stress level. In this study, the SHAP revealed that male patients with COPD were more inclined to be smokers. Younger individuals also had a higher probability of being classified as smokers. Similarly, those with inconclusive pulmonary function test findings tended to be classified as smokers. These findings align with those from previous studies. A hospital-based prospective follow-up study reports that younger and milder patients with COPD are less likely to quit smoking.⁵² Additionally, diabetes prevalence emerged as the strongest comorbidity-related predictor of smoking status, consistent with findings from a previous study showing that diabetes negatively affects

smoking cessation in women.⁵³ However, in this study, persistent smoking was primarily associated with education level among socioeconomic variables. Previous studies show that socioeconomic variables, such as low income or marital status, influence smoking cessation.^{50,52} This finding suggests that, within Korean culture, education level may serve as a broader indicator of socioeconomic variables. Consequently, continuous counseling from healthcare providers and access to smoking cessation programs should be provided to patients with COPD with nonmodifiable risk variables. This support is particularly critical for younger patients with lower socioeconomic status and less severe pulmonary function test findings.

Modifiable lifestyle variables (such as body weight control for a year and weekly walking exercise) exhibited a relatively strong effect on smoking status in individuals with COPD. A randomized trial reports that quitters are more inclined to walk (as a form of exercise) and increase their fruit and vegetable intake compared to those who continue smoking.⁵⁴ This suggests that adopting one healthy behavior may encourage other positive lifestyle changes. However, the relationship between lifestyle changes and smoking status remains unclear, as does the direction of causality between these variables. Further studies are needed to clarify these associations. A similar association may exist between smoking status and psychological variables in our dataset. Among psychological factors, perceived stress levels had the highest effect on smoking status in individuals with COPD. A study that utilized data from the World Health Survey—a cross-sectional, community-based study conducted in 70 countries across the world—reports that perceived stress is associated with daily smoking in most countries.⁵⁵ Given the cross-sectional nature of the previous research and our study, prospective studies are needed to confirm the causal relationship between predictors and smoking status and to enhance the predictive performance of our model for patients with COPD.

Limitations

To our knowledge, this is the first experimental study to predict smoking status in individuals with COPD and contribution to the advancement of artificial intelligence-based prediction models in healthcare. However, the study has some limitations. First, this study relies on data from 2007–2018, constrained by the completeness of the available features. This led to a comparatively small training set after separating the test set for evaluation. Although cross-validation was employed, the lack of external validation remains a limitation. While internal validation offers valuable insights, external validation with comparable external datasets is necessary to enhance generalizability. Second, the causal relationships between smoking status and predictive variables cannot be established using cross-sectional data alone. Future studies incorporating cohort

data may better capture these trajectories. Third, imputing missing values using the mode was appropriate given the low proportion of missing data, but this approach may introduce bias in datasets with more complex missing data mechanisms. Future research should explore advanced techniques, such as multiple imputation or model-based methods, to enhance robustness and generalizability, particularly in more intricate scenarios. Furthermore, variables such as smoking cessation attempts and participation in smoking cessation programs exhibited an effect on smoking status in individuals with COPD but were not included in the features selected from our national dataset. Ultimately, cultural or sociodemographic characteristics from Korea may influence the global generalization of the model.


Clinical usability


The deep learning model developed in this study has the potential for clinical application, particularly in helping healthcare providers identify persistent smokers at higher risk of adverse outcomes. The model showed strong performance during internal validation. Future studies should focus on external validation and incorporating additional behavioral and psychological factors to enhance its accuracy and generalizability. With these advancements, this could become a valuable tool in personalized COPD management, particularly in optimizing smoking-related interventions. Its reliance on easily accessible clinical data, including demographic information and pulmonary function test results, supports its practical implementation in routine care. Furthermore, the explainability of the model, facilitated by SHAP, allows healthcare providers to identify the key predictors influencing the predictions, thereby supporting informed decision-making.

Conclusions

In this study, we developed a high-accuracy deep learning model for predicting smoking status in individuals with COPD. Our proposed model outperformed several machine learning and deep learning models in this predictive task. With further improvement of its performance measures, it could serve as a valuable decision-support tool, aiding healthcare providers and policymakers in identifying smokers with COPD and implementing targeted interventions to prevent disease progression.

ORCID iDs

Sudarshan Pant  <https://orcid.org/0000-0002-2385-9673>

Ja Yun Choi  <https://orcid.org/0000-0002-1284-250X>

Statements and declarations

Ethical considerations

This study received ethical approval from the C National University Hospital IRB (approval #CNUH-2022-232) on July

09, 2023. This is an IRB-approved retrospective study; all patient information was de-identified, and patient consent was not required. Patient data will not be shared with third parties.

Author contributions/CRedit

Sudarshan Pant did investigation, validation, visualization, writing: original draft, review and editing.

Hyung Jeong Yang did investigation, validation, visualization, writing: original draft, review and editing.

Sehyun Cho did data curation, methodology, writing: original draft.

Ryu Eui Jeong did formal analysis, data curation.

Ja Yun Choi did conceptualization or/and methodology, funding acquisition, investigation, project administration or/and supervision, writing, original draft, review and editing.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [grant number NRF-2022R1A2C1010364].

Conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

According to the Korea National Health and Nutrition Examination Survey (KNHANES) data policies, secondary processed data cannot be redistributed. However, raw data can be downloaded for research purposes at <https://knhanes.kdca.go.kr/knhanes/eng/index.do>. To facilitate reproducibility, the trained model weights are publicly available on GitHub (<https://github.com/sudarshanpant/SmokingStatusPrediction4COPD.git>).

Supplemental material

Supplemental material for this article is available online.

References

1. Chen S, Kuhn M, Prettner K, et al. The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories in 2020–50: a health-augmented macroeconomic modelling study. *Lancet Glob Health* 2023; 11: e1183–e1193.
2. Li H, Liang H, Wei L, et al. Health inequality in the global burden of chronic obstructive pulmonary disease: findings from the global burden of disease study 2019. *Int J Chron Obstruct Pulm Dis* 2022; 17: 1695–1702.
3. Zafari Z, Li S, Eakin MN, et al. Projecting long-term health and economic burden of COPD in the United States. *Chest* 2021; 159: 1400–1410.
4. American Lung Association [Internet]. COPD causes and risk factors. [cited 2023 Apr 28], <https://www.lung.org/lung-health-diseases/lung-disease-lookup/copd/what-causes-copd>

5. Toghyani A and Sadeghi S. Association of demographic variables and smoking habits with the severity of lung function in adult smokers. *J Res Med Sci* 2022; 27: 18.
6. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: 2025 report. Global Initiative for Chronic Obstructive Lung Disease, 2025, <https://goldcopd.org/2025-gold-report/>
7. Li X, Wu Z, Xue M et al. An observational study of the effects of smoking cessation earlier on the clinical characteristics and course of acute exacerbations of chronic obstructive pulmonary disease. *BMC Pulm Med* 2022; 22: 21.
8. Au DH, Bryson CL, Chien JW, et al. The effects of smoking cessation on the risk of chronic obstructive pulmonary disease exacerbations. *J Gen Intern Med* 2009; 24: 457–463.
9. Doo JH, Kim SM, Park YJ, et al. Smoking cessation after diagnosis of COPD is associated with lower all-cause and cause-specific mortality: a nationwide population-based cohort study of south Korean men. *BMC Pulm Med* 2023; 23: 237.
10. van Eerd EA, van der Meer RM, van Schayck OC et al. Smoking cessation for people with chronic obstructive pulmonary disease. *Cochrane Database Syst Rev* 2016; 2016: CD010744.
11. Rajkomar A, Dean J and Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380: 1347–1358.
12. Hu S. Deep learning in healthcare. *Highl Sci Eng Technol* 2023; 57: 279–285.
13. Thakur SS, Poddar P and Roy RB. Real-time prediction of smoking activity using machine learning-based multiclass classification model. *Multimed Tools Appl* 2022; 81: 14529–14551.
14. Xu Y, Cao L, Zhao X, et al. Prediction of smoking behavior from single nucleotide polymorphisms with machine learning approaches. *Front Psychiatry* 2020; 11: 16.
15. Wang Z, Masoomi A, Xu Z, et al. Improved prediction of smoking status via isoform-aware RNA-Seq deep learning models. *PLoS Comput Biol* 2021; 17: e1009433.
16. Ebrahimi A, Henriksen MBH, Brasen CL, et al. Identification of patients' smoking status using an explainable AI approach: a Danish electronic health records case study. *BMC Med Res Methodol* 2024; 24: 114.
17. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. *J Big Data* 2019; 6: 54.
18. Troiano A. Wearables and personal health data: putting a premium on your privacy. *Brooklyn Law Rev* 2016; 82: 1715–1753.
19. Choi JY, Ryu EJ, Yun SY, et al. Development of a conceptual framework for non-adherence to self-management in patients with chronic obstructive pulmonary disease: an exploratory study. *Kor J Adult Nurs* 2024; 36: 126–135.
20. Korea National Health and Nutrition Examination Survey [Internet]. Korea National Health and Nutrition Examination Survey (KNHANES). [cited 2020 Feb 28], https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do
21. Kang SN, Kim HW, Lim J, et al. Characteristics of intermittent smokers in Korean adults: comparison with daily smokers. *J Kor Soc Res Nicotine Tob* 2019; 8: 58–64.
22. Cai J, Luo J, Wang S, et al. Feature selection in machine learning: a new perspective. *Neurocomputing* 2019; 300: 70–79.
23. Lin WC and Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev* 2020; 53: 1487–1509.
24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
25. Perrusquía A and Yu W. Identification and optimal control of nonlinear systems using recurrent neural networks and reinforcement learning: an overview. *Neurocomputing* 2021; 438: 145–154.
26. Pant S. Smoking status prediction for COPD [Internet], <https://github.com/sudarshanpant/SmokingStatusPrediction4COPD.git>
27. Loeff B, Wong A, Janssen NAH, et al. Using random forest to identify longitudinal predictors of health in a 30-year cohort study. *Sci Rep* 2022; 12: 10372.
28. Banihashem SY and Shishehchi S. Ontology-based decision tree model for prediction of fatty liver diseases. *Comput Methods Biomech Biomed Eng* 2023; 26: 639–649.
29. Vedaraj M, Anita CS, Muralidhar A, et al. Early prediction of lung cancer using Gaussian naive Bayes classification algorithm. *Int J Intell Syst Appl Eng* 2023; 11: 838–848.
30. Jabbar MA, Deekshatulu BL and Chandra P. Classification of heart disease using K-nearest neighbor and genetic algorithm. *Procedia Technol* 2013; 10: 85–94.
31. Hatwell J, Gaber MM and Azad ARM. Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences. *BMC Med Inform Decis Mak* 2020; 20: 50.
32. Taud H and Mas JF. Multilayer perceptron (MLP). In: Camacho Olmedo MT, Paegelow M, Mas JF and Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Cham: Springer International Publishing, 2018, pp.451–455.
33. Arik SO and Pfister T. TabNet: attentive interpretable tabular learning [oral presentation]. In: Association for the Advancement of Artificial Intelligence Conference-21, 2021 Feb 2–9; Online, <https://ojs.aaai.org/index.php/AAAI/article/view/16826>
34. Zeng S, Arjomandi M, Tong Y, et al. Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: retrospective cohort study. *J Med Internet Res* 2022; 24: e28953.
35. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010; 5: 1315–1316.
36. Lundberg SM and Lee SI. A unified approach to interpreting model predictions [oral presentation]. In: NIPS 2017; 2017 December 4–9; Long Beach, CA, USA. <https://api.semanticscholar.org/CorpusID:21889700>
37. Creamer MR, Wang TW, Babb S, et al. Tobacco product use and cessation indicators among adults - United States, 2018. *Morb Mortal Wkly Rep* 2019; 68: 1013–1019.

38. Liu Y, Greenlund KJ, VanFrank B, et al. Smoking cessation among U.S. adult smokers with and without chronic obstructive pulmonary disease, 2018. *Am J Prev Med* 2022; 62: 492–502.
39. Smith LA, Johnson B, Lee T, et al. Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Lancet Digit Health* 2023; 5: e872–e881.
40. Kanyongo W and Ezugwu AE. Machine learning approaches to medication adherence amongst NCD patients: a systematic literature review. *Inform Med Unlocked* 2023; 38: 101210.
41. Davenport T and Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019; 6: 94–98.
42. Thompson T and Haskard-Zolnieriek K. Adherence and communication. *Oxford Research Encyclopedia of Communication [Internet]*, 27 August 2020. <https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-23>
43. Liu D, Shin WY, Sprecher E, et al. Machine learning approaches to predicting no-shows in pediatric medical appointment. *npj Digit Med* 2022; 5: 50.
44. Nishi M, Nagamitsu R and Matoba S. Development of a prediction model for healthy life years without activity limitation: national cross-sectional study. *JMIR Public Health Surveill* 2023; 9: e46634.
45. Simundic AM. Diagnostic accuracy—part 1: basic concepts: sensitivity and specificity, ROC analysis, STARD statement. *Point Care* 2012; 11: 6–8.
46. Yehuala TZ, Agimas MC, Derseh NM, et al. Machine learning algorithms to predict healthcare-seeking behaviors of mothers for acute respiratory infections and their determinants among children under five in sub-Saharan Africa. *Front Public Health* 2024; 12: 1362392.
47. Kang S. Model validation failure in class imbalance problems. *Expert Syst Appl* 2020; 146: 113190.
48. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
49. Horn K, Dearfield CT, Beth Johnson S, et al. Smoking cessation intentions and attempts one year after the federally mandated smoke-free housing rule. *Prev Med Rep* 2021; 24: 101600.
50. Martins RS, Junaid MU, Khan MS, et al. Factors motivating smoking cessation: a cross-sectional study in a lower-middle-income country. *BMC Public Health* 2021; 21: 1419.
51. Chean KY, Goh LG, Liew KW, et al. Barriers to smoking cessation: a qualitative study from the perspective of primary care in Malaysia. *BMJ Open* 2019; 9: e025491.
52. Tøttenborg SS, Thomsen RW, Johnsen SP, et al. Determinants of smoking cessation in patients with COPD treated in the outpatient setting. *Chest* 2016; 150: 554–562.
53. Allagbé I, Nicolas R, Airagnes G, et al. Clinical factors associated with smoking cessation among smokers with chronic obstructive pulmonary disease by sex: longitudinal analyses from French smoking cessation services. *Heliyon* 2024; 10: e30920.
54. Berg CJ, Thomas JL, An LC, et al. Change in smoking, diet, and walking for exercise in Blacks. *Health Educ Behav* 2012; 39: 191–197.
55. Stubbs B, Veronese N, Vancampfort D, et al. Perceived stress and smoking across 41 countries: a global perspective across Europe, Africa, Asia and the Americas. *Sci Rep* 2017; 7: 7597.