Received: 25 July 2023 | Revised: 27 August 2023 | Accepted: 11 September 2023 | Published online: 14 September 2023

RESEARCH ARTICLE

Artificial Intelligence and Applications
2024, Vol. 2(2) 97–106
DOI: 10.47852/bonviewAIA32021406

BON VIEW PUBLISHING

Implications of Classification Models for Patients with Chronic Obstructive Pulmonary Disease

Mengyao Kang[1], Jiawei Zhao[1] and Farnaz Farid[2,*]

[1]*Faculty of Engineering, The University of Sydney, Australia*

[2]*School of Social Sciences, Western Sydney University, Australia*

Abstract: Machine learning (ML)-based prediction models have the potential to revamp various industries, and one such promising area is healthcare. This study demonstrates the potential impact of ML on healthcare, particularly in managing patients with chronic obstructive pulmonary disease (COPD). The experimental results showcase the remarkable performance of ML models, surpassing doctors' predictions for COPD patients. Among the evaluated models, the gradient-boosted decision tree classifier emerges as the top performer, displaying exceptional classification accuracy, precision, recall, and F1-score compared to doctors' experience. Notably, the comparison between the best ML model and doctors' predictions reveals an interesting pattern: ML models tend to be more conservative, resulting in an increased probability of patient recovery.

Keywords: machine learning, prediction models, chronic obstructive pulmonary disease (COPD), healthcare, gradient-boosted decision tree classifier

1. Introduction

The research aims to investigate the impact of machine learning (ML) methods on optimizing the classification of patients with chronic obstructive pulmonary disease (COPD). COPD is a highly significant and preventable chronic lung disease that substantially burdens patients' daily lives and healthcare systems worldwide (Safiri et al., 2022). Patients with COPD often experience exorbitant medical costs and endure a significant physical perception of pain. Presently, there is a lack of standardized clinical methods to improve patients' self-examination and cognitive abilities to mitigate the disease's progression.

It is necessary to create links between the patient's clinical characteristics and the desired outcomes to accomplish early identification of COPD and significantly reduce patient hospitalization rates (Fromer, 2011). Due to time and resource limitations, medical professionals may find it challenging to track and monitor patients' health status promptly and effectively. The vast amount of information about COPD patients, including medical records, clinical measurements, diagnostic tests, patient histories, and other pertinent data, primarily relies on understanding and managing scarce human resources (López-Campos et al., 2016). Therefore, the development of tools capable of accurately predicting disease progression holds immense value as it can guide patients toward appropriate care within a home setting, ultimately reducing the strain on the healthcare system by minimizing unnecessary hospitalizations.

The application of personalized medicine, with ML as its core technology, is considered a crucial direction in medical advancement. It can facilitate COPD patients in self-examination and assess the likelihood of disease exacerbations, providing medical staff with more accurate decision support (Bertens et al., 2013). This approach alleviates the burden on the healthcare system and decreases dependency on human intervention. By enhancing work efficiency, expanding the scope of health monitoring, and providing personalized services to each patient, this study explores the feasibility of a novel model using a well-established COPD dataset. The paper is organized as follows: Section 2 provides an overview of the previous work in this field. Section 3 presents the detailed experiments, while Section 4 discusses the results and findings. Finally, Section 5 concludes the work and outlines potential future research directions.

2. Previous Work

In recent years, the development of computer operations and the rapid advancement of "big data" have greatly facilitated the application of artificial intelligence (AI) and ML technologies in various fields, including healthcare (Esteva et al., 2019). The utilization of AI techniques, such as the construction of "expert systems" and diagnostic software, has been shown to achieve accuracies surpassing those of pulmonary physicians in diagnosing and testing for COPD (Braido et al., 2018; Topalovic et al., 2019). Random forest (RF) and support vector machine

*Corresponding author: Farnaz Farid, School of Social Sciences, Western Sydney University, Australia. Email: Farnaz.Farid@westernsydney.edu.au

(SVM) algorithms have been used to explore and analyze patients' genetic data to identify dependencies between COPD and lung function, enabling screening, diagnosis, classification, and assessment of COPD (Matsumura & Ito, 2020).

ML models have been employed to determine factors contributing to exacerbation, hospitalization, and risk of readmission in COPD patients. Logistic regression (LR) models have been utilized to estimate the risk of COPD exacerbation within 2 years (Bertens et al., 2013). Wang et al. (2023) conducted an analysis and trained ML models using CT scan images to predict the risk of COPD.

Swaminathan et al. (2017) presented a ML strategy for the early detection and classification of COPD deterioration. The study evaluates nine classification algorithms, the most effective ones being LR and gradient-boosted decision trees (GBDTs). The experimental results show that ML performs well in predicting COPD deterioration and triage. However, the authors acknowledge the limitations of applying ML to real-world scenarios. Further research is needed to enhance its practical implementation.

Cavailles et al. (2020) explored factors influencing rehospitalization risk in COPD patients using a decision tree (DT) model. Their analysis highlighted the significance of the patients' initial 2-year hospitalization period in predicting readmission probability. Factors like age, gender, hospitalization frequency, and anxiety were identified as influential contributors to COPD patient readmissions. Their work guides feature engineering in ML models for accurate readmission risk prediction. This study underscores the importance of these factors in predictive modeling, offering valuable insights for COPD patient management.

Hussain et al. (2021) developed a diagnostic system for predicting COPD severity using ML techniques, including RF, SVM, K-nearest neighbors (KNNs), gradient boosting, and extreme gradient boosting (XGB). They addressed overfitting and underfitting issues by employing soft polling integration and used synthetic minority oversampling technique (SMOTE) to solve the data imbalance problem, which improved the system's accuracy by 4.73%. These findings demonstrate the effectiveness of integration, RFE, and SMOTE for accurate prediction. However, further experiments are required to validate the model's real-world performance.

Dhar (2021) proposed the multistage ensemble model (MSEN) as a solution for detecting COPD patients. The MSEN model, also known as the voting model, combines the outputs of eight trained classifiers through weighted voting. This strategy allows the MSEN model to capitalize on the unique strengths of each classifier, facilitating accurate predictions for samples that pose difficulties for individual classifiers. Each classifier presents distinct advantages. LR enables quick and precise data classification, KNN demonstrates resilience against outliers, and boosting algorithms such as GBDT effectively handle anomalous and challenging data by leveraging nonlinear transformations. The amalgamation of these diverse classifiers offers a promising avenue to enhance the accuracy of prediction tasks.

Preprocessing techniques have enhanced data quality and accuracy, including classification, clustering, and data augmentation. Data augmentation generates additional data for ML models, reducing reliance on training data and improving model performance (Maharana et al., 2022). Data augmentation is widely applied in computer vision, where researchers perform a series of operations on the original images, such as rotation, cropping, and brightness adjustment, to obtain more data (Asperti & Mastronardo, 2017). In natural language processing, researchers enhance the data by employing methods such as paraphrasing, adding appropriate noise, and sampling while ensuring the data's effectiveness, aiming to improve ML efficiency (Li et al., 2022).

Even though many studies have looked at COPD using ML models, this study used a wide variety of classifiers and applied approaches for data augmentation for a more thorough analysis. The study broadens classifiers' scope and uses data augmentation to acquire a more in-depth understanding of COPD prediction and management.

3. Experimental Details

The study investigates two research questions for people with COPD: identifying patient deterioration and determining the required level of care. The study utilizes rigorous dataset analysis, preprocessing, multiple classifier models, and label refining to maximize accuracy and achieve the most insightful results.

3.1. Dataset description

This study uses the dataset from published literature on COPD (Swaminathan et al., 2017). It is obtained through resource integration, expert evaluation, linear modeling, and Monte Carlo simulation. Multiple experts jointly review the dataset's features and labels to ensure objectivity and minimize personal cognitive bias. Moreover, machine-generated patient cases, closely aligned with real-world scenarios as specified by relevant literature and experts, are included.

The dataset is organized into two folders, encompassing training and testing data. The training folder consists of 39 Excel documents containing patient health data and labels assigned by different doctors. To address issues of repetition and disorderly distribution, a subset of 24 Excel tables is selected as the final training set for this project, encompassing a total of 2,400 samples labeled by five doctors. The testing folder comprises nine Excel tables with 101 identical sample health data. Nine doctors independently labeled these samples, and all 101 samples from the test set for this project.

3.2. Data preprocessing

Data preprocessing can correct specific problems in the dataset to improve the model's performance and make the data easier to process by the data model. It selects and processes features by browsing and experimenting with different combinations of columns from the dataset. Due to the limited 2400 sample size of the dataset, it is expanded through data augmentation after data preprocessing.

3.2.1. Feature selection

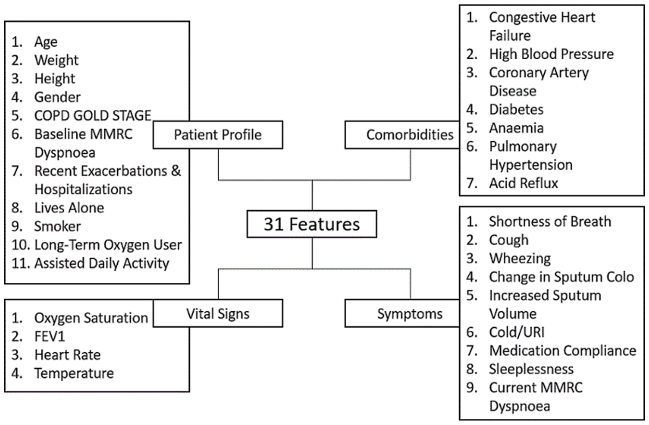
Through feature selection, the experiment will eliminate irrelevant or redundant features to improve the model's accuracy and reduce the running time. Selecting the truly relevant features and simplifying the model will also help understand the data generation process. In this experiment, by browsing the dataset to view data, searching for appropriate research cases, and constructing a classifier model experiment, features are selected based on their practical significance, code feasibility, and classifier processing ability (Gunen et al., 2005; Matkovic et al., 2012; Piquet et al., 2013; Wang et al., 2014).

By analyzing the data contained in each feature, the features with similar content are merged, and the complex features are split. For instance, the two columns of data that use feet and

inches to represent the height of the sample jointly are combined into one column by inch according to the conversion between the two units. The text data description of the "Sputum" column includes whether the color changes and whether the sputum volume increases. According to the extraction of specific words in the text, it is replaced by two features, which are "Change in Sputum Color" and "Increased Sputum Volume".

This study used the wrappers feature selection method, which has high classification accuracy but slow speed (Hsu et al., 2011). Based on the analysis of factors influencing the classification of COPD labels from a healthcare perspective, a total of 31 features were extracted, encompassing four aspects: patient profile, comorbidities, symptoms, and vital signs (Figure 1).

Figure 1
The 31 features



3.2.2. Feature processing

Based on data exploration and analysis, the original variables consist of continuous and categorical features. Variations in features are addressed due to their different characteristics. Analyzing the data revealed that the baseline features in the original dataset contain numerous invalid values, and Gellish et al. (2007) indicated that various factors affect the accuracy of these measurements. Hence, the continuous features were directly processed, and more appropriate features were selected from related original variables such as oxygen saturation, FEV1, and heart rate.

Categorical features include ordinal and unordered categories. Ordinal categorical features were encoded as "1," "2,", and "3" based on their rankings. For the unordered categorical features, the experiment primarily utilized the one-hot (or one-of-k) method to convert each unordered feature into a numerical vector. To make the dataset more comprehensible and easier for the classifier to handle, the feature columns were reshaped, with each risk factor represented as a separate feature and using a binary classification of the one-hot method to indicate whether the current patient possesses that particular risk factor. In addition, the features' null values and "Unknown" data were imputed with the mean or mode of the respective features.

3.2.3. Label processing

The dataset was annotated with two labels. They consist of a binary classification problem to determine whether COPD patients have "exacerbation" (Label1) and a multi-class classification problem to determine "4 nursing levels" (Label2). As the data lack real value as labels, the training set labels were independently annotated by five doctors, while the validation set labels were annotated by nine doctors. Relying on the predictions of a single doctor as labels could introduce bias in the test results. Therefore, in the experiment, the label patterns for each annotated sample were determined by the consensus of the nine doctors as the labels for the validation set to mitigate individual doctor biases. Additionally, after summarizing statistical data, the sample sizes for each classification in the training and testing datasets are imbalanced, with the ratio of samples for the exacerbation label being 2:1 and the nursing levels being approximately 1:2:4:3.

3.2.4. Data augmentation

Data augmentation techniques are applied to expand the dataset. It obtains more data by making minor changes to the existing dataset, which will help optimize the model (Moreno-Barea et al., 2020). The methods for data augmentation in the experiment include randomly changing the patient's age, height, and weight within a small range. Because COPD is a chronic disease, there is a possibility of latent or apparent symptoms when the patient's sample is collected; reducing or increasing their age by 1 or 2 years can be explained. Height measurement may have errors due to the thickness of the sole and hair. At the same time, weight may also have a range of errors due to the weight of clothing or the amount of food consumed during the day the sample was measured. Data augmentation can help experiments avoid being limited by a fixed sample size. In addition, combining SMOTE can also fill the problem of imbalanced sample sizes in various categories of the original dataset, so that each classification in the ML process is trained fairly.

3.3. Classifier selection

The 18 classifiers used in this experiment can be categorized based on their underlying principles of classification as follows:

- Linear Models: LR and stochastic gradient descent (SGD)
- Support Vector Machines: support vector machines with polynomial kernel, support vector machines with linear kernel, and support vector machines with Gaussian kernel
- Decision Trees: DT, RF, GBDTs, extra trees (ETs), XGBoost, AdaBoost, and CatBoost
- Naïve Bayes: Naïve Bayes with Gaussian distribution (GNB), Naïve Bayes with Gaussian and sigmoid distributions (GNB-S), multinomial Naïve Bayes with Laplace smoothing, and Bernoulli Naïve Bayes (BNB)
- Other Classifiers: KNN and voting classifier.

These classifiers were selected based on their specific algorithms and characteristics, which align with the requirements and objectives of the experimental study. In the experiment, K-fold model cross-validation technology was used to effectively utilize all raw data, and the optimal parameter set that met the target was selected through grid search combined with manual parameter adjustment.

3.4. Evaluation criteria

Experimental results are evaluated using five criteria: confusion matrix, accuracy, precision, recall, and F1-score. The confusion matrix visually displays predicted and actual results for comprehensive analysis (Deng et al., 2016). Accuracy measures the proportion of correct predictions among all samples, while

precision identifies the false positive rate for each category. Recall quantifies the classifier's ability to correctly predict samples in each category, and F1-score provides a comprehensive measure of classifier sensitivity by considering both false positive and false negative rates.

4. Results and Discussions

4.1. Results analysis

We expanded the dataset to 6000 samples in the experiment and observed a significant improvement in the accuracy of most of the models by 2–3%. The results of the experiments involving 18 ML classifiers were analyzed and evaluated based on two sets of labels. The classification results were compared with those of human doctors to verify the usability and potential application of the ML models. Additionally, another method for generating labels was experimented with, which aimed to mitigate the bias introduced by the influence of human doctors on the labels. The analysis of the results demonstrated the effectiveness of the ML models in accurately classifying the data. The classification performance of the models was found to be comparable, and in some cases even superior, to that of human doctors. This suggests that ML models have the potential to be utilized as valuable decision-support tools in medical applications.

Among all the classifiers tested, GBDT achieved the highest prediction accuracy. It combines weak learners into strong learners and calculates the negative gradient based on the residual between the current model's prediction results and the target value, following the direction of the negative gradient, in order to minimize the residual and ultimately optimize the prediction results (Rao et al., 2019). But while improving the prediction results, it usually comes at the cost of longer training time (Wu et al., 2021). When conducting local experiments based on the same hardware and software, GBDT used the longest train-test computational time.

4.1.1. Whether Exacerbation

In the evaluation of the Whether Exacerbation classification experiment with 18 ML classifiers, the results were recorded and analyzed. The findings are presented in a bar chart that summarizes the classifiers' accuracy along with the accuracy of nine doctors (Figure 2).

In the bar chart, the black bars represent the classifier models, while the white bars represent human doctors. From the color distribution in the figure, it is evident that most ML models perform similarly to human doctors in this classification problem. The best classifier, GBDT, achieves an accuracy of 96%, while the best doctor, Doctor 10, has an accuracy of 94%. GBDT has a slightly higher accuracy than the best doctor. The results indicate that the majority of classifiers have an accuracy of over 90%. Among these classifiers, models based on DT algorithms, such as GBDT, ET, RF, AdaBoost, CatBoost, XGBoost, and DT, exhibit higher accuracy. BNB has an accuracy lower than 80%. This may be due to the dataset containing multiple features that introduce interference factors for this classifier. This conjecture is supported by the improved accuracy of BNB when the dataset with reduced features was tested. The bar chart displays the best accuracy achieved by BNB with the reduced set of features.

To provide a more detailed analysis of the classification performance, the experiment compares GBDT (the best classifier) and Doctor 10 (the doctor with the highest accuracy) using evaluation metrics such as accuracy, precision, recall, and F1-score (Table 1). The comparison shows that GBDT outperforms Doctor 10 in all the listed evaluation metrics. However, for this simple binary classification problem, the difference in predictive performance between the best doctor and GBDT is not substantial.

The confusion matrix (Figure 3) can be used to further analyze their classification results. The best doctor misclassified two samples

Table 1
Interpretation of the mean scale for belief, concern, and practice

	GBDT			Doctor 10		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
0	0.94	0.94	0.94	0.93	0.88	0.90
1	0.97	0.97	0.97	0.94	0.97	0.96
Average	0.95	0.95	0.95	0.94	0.92	0.93
Accuracy	96%			94%		

Figure 2
The bar chart results of Whether Exacerbation

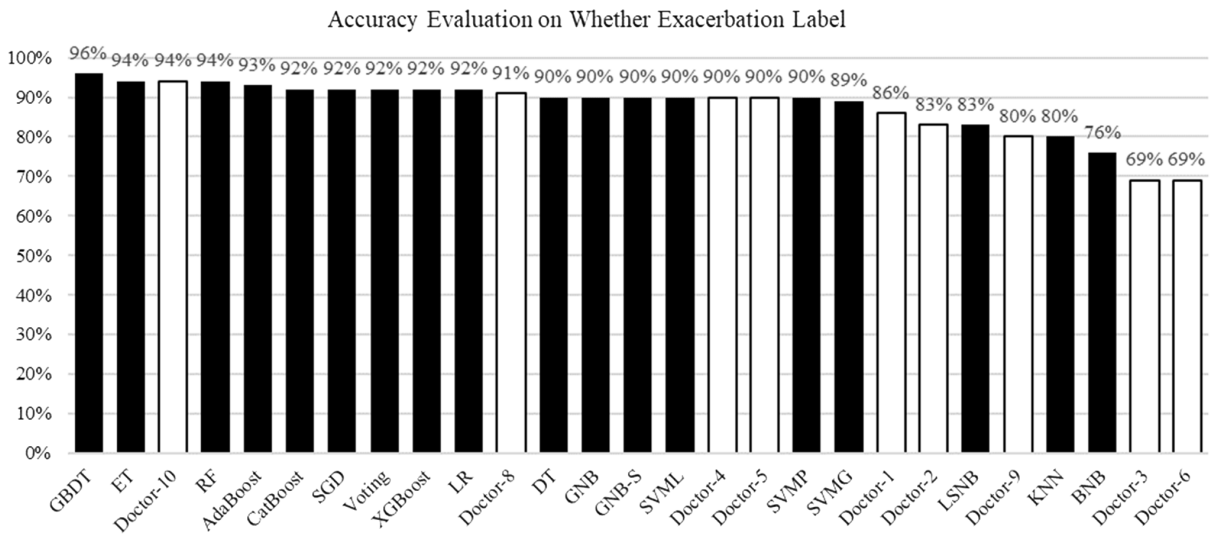


Figure 3
The confusion matrices of GBDT and Doctor 10 in Whether Exacerbation label

GBDT		Predicted Class	
		0	1
True Class	0	30	2
	1	2	67

Dorcter10		Predicted Class	
		0	1
True Class	0	28	4
	1	2	67

Figure 4
The learning curve of Whether Exacerbation label

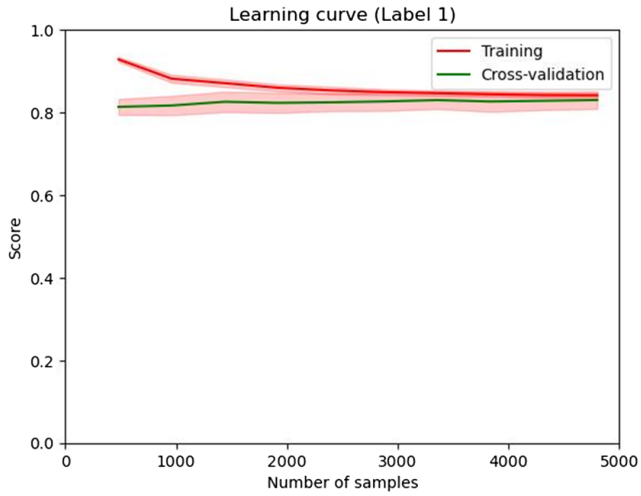
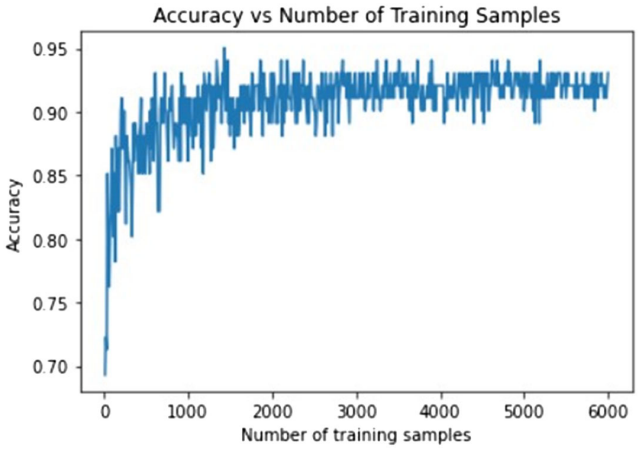


Figure 5
The test curve of Whether Exacerbation label



as class 1 (severe) instead of the true class 0 (non-severe) among the 101 validation samples. This suggests that the best ML model outperforms the best doctor in classifying whether it is severe or not. However, the overall comparison between the two is difficult to assert due to factors such as the small size of the validation set and the uncertainty of the label accuracy in the validation set. In this training and testing, the optimal GBDT model parameters were taken as learning_rate of 0.03, n_estimators of 100, max_depth of 4, and max_feature of "log2".

From the learning curve, it can be seen that on the Whether Exacerbation label, the training and validation curves of the model converge to a higher value and tend to stabilize (Figure 4). Simultaneously, according to the curve of the accuracy of the test set changing with the number of training samples, the model gradually becomes more accurate and stable (Figure 5).

4.1.2. Nursing grades

Based on a dataset of 6000 samples, 18 ML classifiers were tested for classifying Nursing Grades labels. The best classification results of each classifier and the classification accuracy of nine doctors were summarized and displayed in a bar chart (Figure 4).

The results indicate that both ML models and human doctors exhibit lower accuracy in multi-class problems compared to binary classification tasks. The color distribution of the black bars (classifiers) and white bars (doctors) reveals a trend where most ML models outperform human doctors in the four-class problem. The best-performing classifier, GBDT, achieves an accuracy rate of 84%, while the highest-performing human doctor achieves a 77% accuracy rate (Figure 6). Boosting algorithms show more significant potential and research value in predicting the two labels of ML models. GBDT, CatBoost, AdaBoost, and voting classifiers using boosting algorithms demonstrate high accuracy in these four classification experiments, surpassing the best-performing doctor. DT-based classifiers perform exceptionally well. Linear classifiers SGD and LR results fall short of the highest accuracy. Still, their performance on this problem is on par with top-tier human doctors. SVM models and Naive Bayes also do not exhibit outstanding performance. Among the 11 ML models, more than two-thirds achieve a classification accuracy exceeding 70%. Among the doctors, three doctors achieve a classification accuracy of over 70%. ML generally provides more stable classification performance at the nursing level than doctors. BNB shows the lowest performance in predicting the Nursing Grades label. This result is consistent with the best results obtained using reduced feature data.

GBDT and Doctor 10, the classifiers with the highest accuracy rates and the human doctors, are analyzed and compared using evaluation metrics (Table 2). When faced with a multi-classification task, GBDT and Doctor 10 exhibit a more significant discrepancy than the binary classification results. GBDT outperforms Doctor 10 in accuracy, precision, recall, and F1-score, as shown in the confusion matrix (Figure 7).

Table 2
The evaluation value of GBDT and Doctor 10 in Nursing Grades label

	GBDT			Doctor 10		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
1	1.00	1.00	1.00	0.50	1.00	0.67
2	0.87	0.68	0.76	0.64	0.74	0.68
3	0.83	0.86	0.84	0.80	0.80	0.80
4	0.83	0.89	0.86	0.95	0.71	0.82
Average	0.88	0.71	0.87	0.72	0.81	0.74
Accuracy	84%			77%		

Figure 6
The bar chart results of Nursing Grades

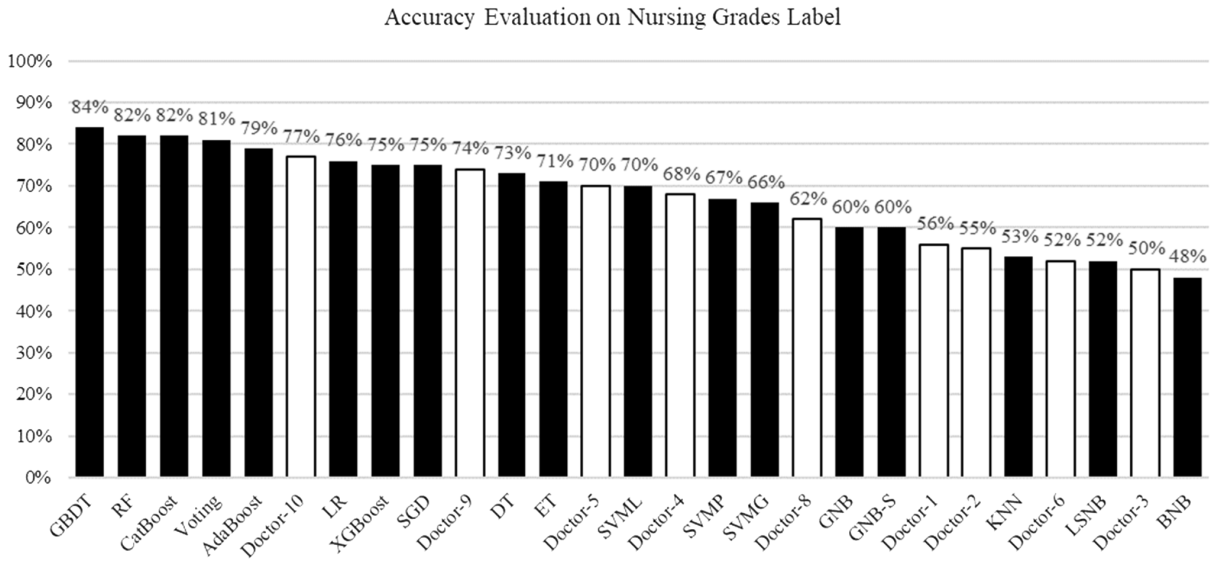
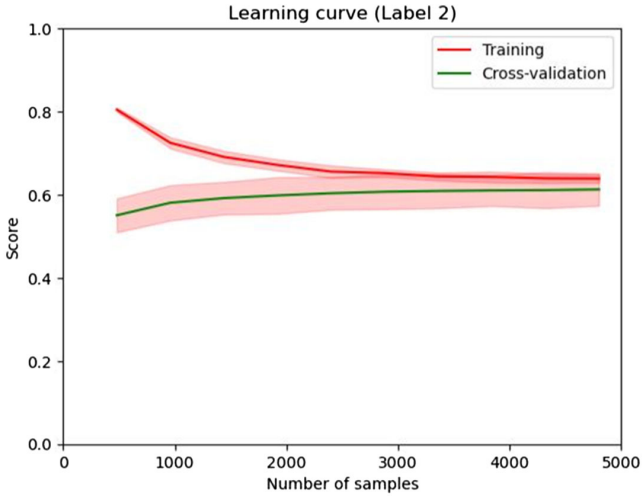


Figure 7
The confusion matrices of GBDT and Doctor 10 in Nursing Grades label

GBDT		Predicted Class			
		1	2	3	4
True Class	1	4	0	0	0
	2	0	13	6	0
	3	0	2	43	5
	4	0	0	3	25

Dorcter10		Predicted Class			
		1	2	3	4
True Class	1	4	0	0	0
	2	3	14	2	0
	3	1	8	40	1
	4	0	0	8	20

Figure 8
The learning curve of Nursing Grades label

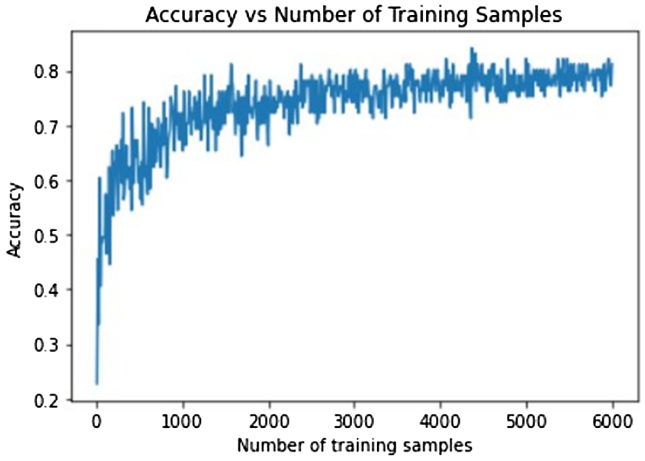


The confusion matrix reveals that GBDT tends to make more mistakes in classifying the second and third levels. The GBDT classifier incorrectly classifies five samples as lower-level classes and misclassifies 11 samples as higher-level classes. On the other hand, Doctor 10 is more prone to errors in classifying the third and fourth levels. He misclassifies 20 samples as lower-level classes and three as higher-level classes. These results suggest that the misclassification of Grades labels may be attributed to the ambiguity in the boundaries between nursing levels, making it challenging to define them clearly. Compared to Doctor 10, GBDT tends to be more conservative when assigning samples to higher-level categories, considering this ambiguity. The optimal GBDT model parameters were taken as learning_rate of 0.08, n_estimators of 140, max_depth of 2, and max_feature of "log2".

From the learning curve, it can be seen that on Nursing Grades label, the training and validation curves of the model converge to a relatively stable value, but the accuracy has decreased (Figure 8). This may be due to the use of data augmentation, resulting in more false data appearing in the data. The label is more complex than the first, which makes this impact more obvious. However, according to the curve of the accuracy of the test set changing with the number of training samples, the model gradually becomes more accurate and stable (Figure 9). This indicates that the model has a certain degree of robustness to noise and has a certain degree of generalization ability when facing complex real data.

According to the comparison of feature importance of the best models of the two labeling studies, the "Vital Signs" features mentioned in Section 3.2.1 generally have a significant impact on GBDT, followed by the "Symptoms" (Figure 1). Among them, oxygen saturation has the highest weight among the 31 features, accounting for about a quarter. Shortness of breath, wheeling, and cough in "symptoms" are also important influencing factors that deserve attention.

Figure 9
The test curve of Nursing Grades label



4.1.3. Second labeling method

The previous results were based on a validation set with labels annotated by nine doctors, which served as reference labels for prediction and comparison. This labeling method reduces individual biases among doctors. However, this labeling method gives an advantage to doctors when comparing them with classifiers. In this case, ML models still outperform doctors, effectively demonstrating the practicality of ML models. Therefore, to further compare them on an equal basis with human doctors, another labeling method was employed, which involved using the mode between the best predictions of each classifier and the results of the nine doctors as the labels.

With this labeling method, the performance of the best classifier, GBDT, and the best doctor remains unchanged for the "Whether Exacerbation" problem (Table 1). GBDT and Doctor 10 remain the best classifiers and doctors, respectively. This result is because the final labels determined from the validation set of 101 samples did not change, i.e., the mode agreement between doctors and classifiers remains consistent with the previous labeling method. Additionally, their original accuracies were already at a high value, and the difference between them is essentially only two samples, resulting in the minimal impact of the new labeling method on the final results.

With this labeling method, GBDT achieves an accuracy rate of 91% on the "Nursing Grades" label, while the best doctor becomes Doctor 4 with an accuracy rate of 75% (Table 3). This demonstrates that with the inclusion of GBDT, the correctness of some human-annotated labels in the validation set is called into question, thereby subtly showcasing how ML models can collaborate when human doctors make decisions.

Table 3
The evaluation value of GBDT and Doctor 4 in Nursing Grade label with second labeling method

	GBDT			Doctor 4		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
1	1.00	1.00	1.00	1.00	0.75	0.86
2	0.93	0.78	0.85	0.62	0.28	0.38
3	0.87	0.96	0.91	0.71	0.83	0.76
4	0.97	0.91	0.94	0.83	0.91	0.87
Average	0.94	0.91	0.92	0.79	0.69	0.72
Accuracy	91%			75%		

4.1.4. Balanced dataset

After multiple and varied data augmentation experiments and incorporating the results of various classifiers, it was found that the accuracy of the balanced dataset was slightly lower than the sample proportion of each class in the original dataset during testing. And the result is even higher in model cross-validation. This may be due to the uneven expansion of data samples to balance the dataset, resulting in spurious data that affect the quality of the dataset and model to a certain extent. Another possible reason is the inconsistent distribution of the training data and the used test data after balancing, as the distribution of each class of samples in the test dataset is close to that of the original dataset.

However, when testing on imbalanced data, various ensemble algorithms, DTs, RF, and SVMs have better predictive performance, especially GBDT, due to their inherent noise immunity for imbalanced data. The experiments in the study of Sun et al. (2021) were also conducted on imbalanced data, and GBDT performed better. In real situations, collected data often cannot achieve complete balance. Using algorithms with stronger anti-interference properties can enhance the generalization ability of ML models, making it easier to apply them to a wide range of real-world problems.

4.2. Discussions

4.2.1. Practical significance

The study investigation demonstrates that ML models exhibit predictive performance that surpasses human predictions. Introducing ML models into the healthcare industry can significantly assist in reducing hospital readmission rates and anticipating future health conditions. Compared to doctors' predictive effectiveness for COPD, most ML models demonstrate superior performance. In prediction experiments for different questions, such as patient deterioration and required care level, most classifiers outperform multiple doctors in accuracy. ML can better forecast patients' health conditions, adjust their physiological awareness and medical plans at an earlier stage, and prevent further deterioration of future illnesses.

Moreover, GBDT performs best among the classifiers, surpassing the highest accuracy, precision, recall, and F1-score. The confusion matrix of their predictions illustrates the conservative nature of GBDT's predictions. Although patients may not receive the most accurate treatment plans, receiving a higher level of care ensures that they are not subject to relapses due to missed effective treatments, increases the likelihood of recovery, and reduces the probability of readmission.

Higher accuracy allows patients to obtain more precise medical plans, avoiding additional healthcare expenses. With the development and application of ML models, they can aid in self-diagnosis for patients and provide references for doctor decisions, thereby relieving the pressure on medical resources and reducing costs to some extent. ML can also reduce some patient expenses; statistics show that 21% of COPD patients in the United States are readmitted, usually incurring 18% higher costs than their previous admissions (Min et al., 2019). Although the results indicate that ML models focusing on assigning patients to higher levels of care may incur additional costs during treatment, reducing readmission rates will save patients more costs. Additionally, with the current technological advancements,

training ML models to possess highly accurate judgment abilities will incur lower costs than training human doctors.

4.2.2. Advantages

This study investigates ML models, including the universality of classification results and the representativeness of selecting the best classification results. This study uses 18 different ML models. Compared with previous similar studies, introducing more classifiers can more rigorously obtain the wide availability of ML models in the healthcare field. After simple data augmentation of the dataset size, the accuracy of most classifiers has increased by 1–3%. Through rigorous analysis, this research selects the optimal classifier and representative doctor for conducting in-depth data analysis. By completing a comparative analysis of metrics such as accuracy, precision, recall, F1-score, and confusion matrix using two labeling methods, this study comprehensively evaluates the advantages of the ML models deployed, enhancing the persuasiveness of the experimental findings.

4.2.3. Gaps

A particular gap exists between the study's implementation and simulation and its actual application. Notably, data collection for chronic diseases poses difficulties. However, if the actual patient states can be accurately recorded as labels, it would reduce the noise on the labels and, to some extent, decrease the noise ratio during data augmentation. Training the classifiers with accurate labels would result in more precise training.

Furthermore, although the experimental results demonstrate the powerful predictive performance of the ML models, these results are generated on specific datasets. The characteristics of the datasets were specifically processed in the early stages based on the data's properties. It is uncertain whether ML models can produce the same results when faced with more complex data. Studying to determine the common and exact factors contributing to COPD would greatly facilitate the widespread application of ML models in daily life. Identifying fixed high correlational factors would be beneficial for training and maintaining the accuracy of the models.

Additionally, the interpretability of ML models is relatively low. It is challenging to clearly explain the influences that lead to the final classification results, interactions between datasets, and their impact on ML models. In order to comprehensively explain the effect of applying ML models in the healthcare domain, further research on the interpretability of these models is necessary. This is a crucial question for the widespread adoption of ML models.

4.2.4. Challenges

There are still some challenges that cannot be ignored in the application of ML in the real world. First, patients are unable to trust the prediction results provided by ML models unconditionally. The low interpretability of machine decision-making makes it difficult for machines to make a sound judgment of the condition and provide transparency in the process of judging the state. When the process is explained correctly, patients can gradually accept suggestions from the machine.

When using ML technology to predict patients, it is inevitable to use a large amount of actual data to train and test the model. Ensuring the privacy and security of patients is a necessary consideration. The loss or leakage of any critical data may lead to serious adverse consequences, which have significant negative impacts on patients and medical institutions using machine assistance.

From a moral and legal perspective, when using ML for medical assistance, prediction bias may lead to specific errors. Clarifying the division of responsibilities when errors occur and ensuring that the rights and interests of patients and medical institutions are protected is a must before ML is put into practical application.

5. Conclusion and Future Work

Following the modeling with 18 ML classifiers and two labeling methods for the "Whether Exacerbation" and "Nursing Grades" labels, the results of all models were recorded, and the accuracies of the classifiers and nine doctors were summarized and ranked. According to the benchmark, the best classifier in the experiment is GBDT. Using the first labeling method, the GBDT classifier achieved the highest result of 96% for the Exacerbation label. In comparison, doctors had the highest result of 94%. For the Nursing Grades label, the highest result for the classifier was 84%, while the highest result for doctors was 77%.

Following the principle of establishing validation set labels on an equal basis between classifiers and human doctors, the second labeling method was used. The results for the Exacerbation label remained unchanged. However, the GBDT classifier achieved a classification accuracy of 91% on the Grades label. There were significant improvements in precision, recall, and F1-scores. Additionally, due to the influence of GBDT on the validation set labels, the best doctor transitioned from Doctor 10 to Doctor 4. This validates the assistive role of ML in human decision-making.

The results demonstrate that with the COPD dataset, ML models can make relatively accurate predictions. Notably, not only the GBDT classifiers but also most of the other classifiers exhibit consistently reliable predictions, surpassing the summarization of doctors' experiences. This observation serves as compelling evidence that ML models can achieve predictive performance that exceeds human experiential predictions, highlighting the transformative potential of these models in various domains, including healthcare. Furthermore, the predictive results of ML models can also be enhanced with inexpensive training to increase their relatively conservative nature, i.e., even if the nursing grade classification is incorrect, it can still be assigned to a higher grade without delaying the patient's condition. If such a model could be practically applied in home testing or collaborative decision-making in hospitals, it would significantly control patient readmission rates and reduce healthcare costs.

ML applications in the healthcare field are becoming increasingly mature. In the future, with the widespread use and deployment of ML, ML models may better assist doctors in making decisions, thereby increasing healthcare accuracy and helping reduce hospital costs. It also aids patients in self-testing at home, alleviating the pressure on COPD healthcare personnel and assisting patients in reducing consultation costs. Suppose the predicted probability of patient deterioration is high. In that case, patients can take timely self-management measures to control and alleviate their condition, avoiding readmission and reducing the cost of hospitalization. This indicates that ML can positively affect healthcare costs and readmissions.

Experiments are limited by sample size, with training samples only annotated by doctors based on their experience. The experiments only expanded the sample size through simple data augmentation. With the continuous improvement of modern health data and the development of cloud computing, there will be more opportunities to accumulate and obtain more patient condition data samples and train ML algorithms on larger real datasets. This will significantly improve the accuracy and stability of existing algorithms, laying the foundation for the application and deployment of future algorithms. Furthermore, the current algorithm model can still be improved by adding or generalizing more new classification models, integrating various algorithm models, or reshaping existing models. Various neural

network models may enhance classification performance, thereby enhancing the performance of ML methods in diagnosing COPD patient conditions and nursing levels. The participation of more data resources will also validate the improved models, which will be more conducive to scalable model development and may be promoted in a wider medical environment or different patient groups. Moreover, training the machine based on more balanced real data and continuously adjusting model parameters during this process may strike a balance between its conservative methods and the professional knowledge of doctors to obtain more appropriate decisions.

Currently, the ML algorithms of this study have not been deployed in mobile applications. In the future, the algorithms of this study can be deployed in mobile applications for use by patients, doctors, and nurses. Patients can use this mobile application to assess the severity of their condition, prevent anxiety among patients who are healthy and do not require hospitalization, and help alleviate the psychological pressure on patients. Doctors and nurses can understand patients' basic health conditions and assist doctors in evaluating patients' conditions. At the same time, through further training, the application of ML models can continuously track patients' diseases and assist doctors in providing personalized healthcare plans that meet different patient preferences based on the information provided by patients, including recommendations for drugs and certain lifestyle interventions. Additionally, the clinical effectiveness of mobile applications in real patient populations should be further explored in the future as they are introduced into the healthcare sector to improve patient decision-making and reduce the severity and frequency of COPD deterioration in a clinical environment.

Although at this stage, doctors' diagnostic opinions are the primary criterion for most clinical decisions, including the diagnosis of COPD exacerbation and the determination of nursing levels, the active cloud training that combines patient data in electronic medical records with ML and existing scientific knowledge will likely provide specific COPD condition predictions and nursing-level recommendations to support medical decision-making.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in PLOS ONE at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188532

Author Contribution Statement

Mengyao Kang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Jiawei Zhao:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Farnaz Farid:** Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration.

References

Asperti, A., & Mastronardo, C. (2017). The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images. *arXiv Preprint: 1712.03689*.

Bertens, L. C., Reitsma, J. B., Moons, K. G., van Mourik, Y., Lammers, J. W. J., Broekhuizen, B. D., . . . , & Rutten, F. H. (2013). Development and validation of a model to predict the risk of exacerbations in chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, *8*, 493–499. https://doi.org/10.2147/COPD.S49609

Braido, F., Santus, P., Corsico, A. G., Di Marco, F., Melioli, G., Scichilone, N., & Solidoro, P. (2018). Chronic obstructive lung disease "expert system": Validation of a predictive tool for assisting diagnosis. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 1747–1753. https://doi.org/10.2147/COPD.S165533

Cavailles, A., Melloni, B., Motola, S., Dayde, F., Laurent, M., Le Lay, K., . . . , & Flament, T. (2020). Identification of patient profiles with high risk of hospital re-admissions for acute COPD exacerbations (AECOPD) in France using a machine learning model. *International Journal of Chronic Obstructive Pulmonary Disease*, *15*, 949–962. https://doi.org/10.2147/COPD.S236787

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340-341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Dhar, J. (2021). Multistage ensemble learning model with weighted voting and genetic algorithm optimization strategy for detecting chronic obstructive pulmonary disease. *IEEE Access*, *9*, 48640–48657. https://doi.org/10.1109/ACCESS.2021.3067949

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., . . . , & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fromer, L. (2011). Diagnosing and treating COPD: Understanding the challenges and finding solutions. *International Journal of General Medicine*, *4*, 729–739. https://doi.org/10.2147/IJGM.S21387

Gellish, R. L., Goslin, B. R., Olson, R. E., McDonald, A., Russi, G. D., & Moudgil, V. K. (2007). Longitudinal modeling of the relationship between age and maximal heart rate. *Medicine & Science in Sports and Exercise*, *39*(5), 822–829. https://doi.org/10.1097/mss.0b013e31803349c6

Gunen, H., Hacievliyagil, S. S., Kosar, F., Mutlu, L. C., Gulbas, G., Pehlivan, E., . . . , & Kizkin, O. (2005). Factors affecting survival of hospitalised patients with COPD. *European Respiratory Journal*, *26*(2), 234–241. https://doi.org/10.1183/09031936.05.00024804

Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, *38*(7), 8144–8150. https://doi.org/10.1016/j.eswa.2010.12.156

Hussain, A., Choi, H. E., Kim, H. J., Aich, S., Saqlain, M., & Kim, H. C. (2021). Forecast the exacerbation in patients of chronic obstructive pulmonary disease with clinical indicators using machine learning techniques. *Diagnostics*, *11*(5), 829. https://doi.org/10.3390/diagnostics11050829

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

López-Campos, J. L., Tan, W., & Soriano, J. B. (2016). Global burden of COPD. *Respirology*, *21*(1), 14–23. https://doi.org/10.1111/resp.12660

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global*

Transitions Proceedings, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Matkovic, Z., Huerta, A., Soler, N., Domingo, R., Gabarrús, A., Torres, A., & Miravitlles, M. (2012). Predictors of adverse outcome in patients hospitalised for exacerbation of chronic obstructive pulmonary disease. *Respiration*, *84*(1), 17–26. https://doi.org/10.1159/000335467

Matsumura, K., & Ito, S. (2020). Novel biomarker genes which distinguish between smokers and chronic obstructive pulmonary disease patients with machine learning approach. *BMC Pulmonary Medicine*, *20*(1), 29. https://doi.org/10.1186/s12890-020-1062-9

Min, X., Yu, B., & Wang, F. (2019). Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD. *Scientific Reports*, *9*(1), 2362. https://doi.org/10.1038/s41598-019-39071-y

Moreno-Barea, F. J., Jerez, J. M., & Franco, L. (2020). Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications*, *161*, 113696. https://doi.org/10.1016/j.eswa.2020.113696

Piquet, J., Chavaillon, J. M., David, P., Martin, F., Blanchon, F., & Roche, N. (2013). High-risk patients following hospitalisation for an acute exacerbation of COPD. *European Respiratory Journal*, *42*(4), 946–955. https://doi.org/10.1183/09031936.00180312

Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., . . . , & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, *74*, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

Safiri, S., Carson-Chahhoud, K., Noori, M., Nejadghaderi, S. A., Sullman, M. J., Heris, J. A., . . . , & Kaufman, J. S. (2022). Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: Results from the global burden of disease study 2019. *BMJ*, *378*, e069679. https://doi.org/10.1136/bmj-2021-069679

Sun, Y., Li, Z., Li, X., & Zhang, J. (2021). Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. *Applied Artificial Intelligence*, *35*(4), 290–303. https://doi.org/10.1080/08839514.2021.1877481

Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., . . . , & Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLOS ONE*, *12*(11), e0188532. https://doi.org/10.1371/journal.pone.0188532

Topalovic, M., Das, N., Burgel, P. R., Daenen, M., Derom, E., Haenebalcke, C., . . . , & Janssens, W. (2019). Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *European Respiratory Journal*, *53*(4), 1801660. https://doi.org/10.1183/13993003.01660-2018

Wang, J. M., Labaki, W. W., Murray, S., Martinez, F. J., Curtis, J. L., Hoffman, E. A., . . . , & Hatt, C. (2023). Machine learning for screening of at-risk, mild and moderate COPD patients at risk of FEV1 decline: Results from COPDGene and SPIROMICS. *Frontiers in Physiology*, *14*, 1144192. https://doi.org/10.3389/fphys.2023.1144192

Wang, Y., Stavem, K., Dahl, F. A., Humerfelt, S., & Haugen, T. (2014). Factors associated with a prolonged length of stay after acute exacerbation of chronic obstructive pulmonary disease (AECOPD). *International Journal of Chronic Obstructive Pulmonary Disease*, *9*(1), 99–105. https://doi.org/10.2147/COPD.S51467

Wu, W., Wang, J., Huang, Y., Zhao, H., & Wang, X. (2021). A novel way to determine transient heat flux based on GBDT machine learning algorithm. *International Journal of Heat and Mass Transfer*, *179*, 121746. https://doi.org/10.1016/j.ijheatmasstransfer.2021.121746

How to Cite: Kang, M., Zhao, J., & Farid, F. (2024). Implications of Classification Models for Patients with Chronic Obstructive Pulmonary Disease. *Artificial Intelligence and Applications*, *2*(2), 97–106. https://doi.org/10.47852/bonviewAIA32021406