International Journal Software Engineering and Computer Science (IJSECS)
5 (3), 2025, 1133-1140
Published Online December 2025 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v5i3.5551.

RESEARCH ARTICLE Open Access

Application of Random Forest Method in Predicting Chronic Obstructive Pulmonary Disease (COPD)

Muhhamad Fatkhurridlo Mahendra *
Universitas Islam Nahdlatul Ulama Jepara, Jepara Regency, Central Java Province, Indonesia.
Corresponding Email: mahendramoeremans@gmail.com.

Nur Aeni Widiyastuti
Universitas Islam Nahdlatul Ulama Jepara, Jepara Regency, Central Java Province, Indonesia.

Sarwido
Universitas Islam Nahdlatul Ulama Jepara, Jepara Regency, Central Java Province, Indonesia.

Received: October 10, 2025; Accepted: November 10, 2025; Published: December 1, 2025.

Abstract: Chronic Obstructive Pulmonary Disease (COPD) is one of the major global health problems and remains among the leading causes of death worldwide. Early detection plays a crucial role in preventing disease progression; however, conventional diagnostic methods such as spirometry and CT scans often require high costs, long processing time, and specialized expertise. This study aims to apply the Random Forest algorithm, one of the machine learning methods, to predict COPD based on clinical and lifestyle data. The dataset was obtained from Kaggle, consisting of attributes including age, gender, smoking status, type of occupation, sleep habits, exercise activity, insurance ownership, and history of comorbidities. The research stages include data preprocessing, train-test splitting (80:20), and model evaluation using accuracy, precision, recall, F1-score, and AUC metrics. The Random Forest model achieved an accuracy below 90% (approximately 87%), reflecting realistic performance in medical prediction while avoiding overfitting. The results indicate that Random Forest can serve as a reliable method for COPD detection and holds potential to be developed as the foundation of a Clinical Decision Support System (CDSS). This study contributes to the growing body of literature on the application of machine learning in healthcare, while also offering a faster, cost-effective, and scalable alternative for diagnosis.

Keywords: Chronic Obstructive Pulmonary Disease (COPD); Random Forest; Machine Learning; Medical Prediction; Clinical Decision Support System.

1. Introduction

Chronic Obstructive Pulmonary Disease is, among the non-communicable diseases, the most neglected yet it has turned into a silent killer of millions across the world. It does not come with acute symptoms but rather develops over time without being noticed until there is considerable damage to the lungs. Patients usually present with chronic cough, dyspnea, and sputum production that progressively worsen over months and years. The two main conditions that comprise COPD are chronic bronchitis and emphysema; both progressively destroy lung tissue making every breath more difficult than the last. For patients with advanced disease, even climbing stairs or walking short distances becomes an exhausting challenge in their daily activities. This burden extends beyond just one patient; healthcare systems around the world are weighed down by COPD-related costs. Frequent visits to emergency rooms long hospital stays during exacerbations and a lifelong need for medication drain resources that could be used for other medical priorities. In developing

countries where the healthcare infrastructure already operates under strain, COPD adds yet another layer of complexity to an already overburdened system.

Mei *et al*. (2022) reviewed mortality data in their Global Burden of Disease study and found that COPD always ranked among the top causes of death in all continents. The number of cases has not stabilized; it increases every year [1]. A different meta-analysis published by AL Wachami *et al*. (2024) showed equally alarming results: global prevalence rates for COPD are still rising [2]. Smokers carry the highest risk but do not carry this burden alone; low-income communities in developing countries where factories operate with little emission control and homes still use wood or coal for cooking also suffer disproportionately from disease burdens. Tobacco use, air pollution, indoor smoke from biomass fuels, and chemical exposure at work have all been named by The Lancet Respiratory Medicine as major contributors to rising rates of COPD. These risk factors combined make it clear that unless there is some coordinated global action taken soon we will be facing a public health crisis in terms of COPD for many generations to come.

Currently, Chronic Obstructive Pulmonary Disease is diagnosed by two primary methods: the spirogram, which assesses the volume and flow of exhaled air. While this process seems straightforward, it requires trained personnel for correct execution. The second method involves imaging through chest X-rays and CT scans to visualize damage within the lungs and exclude other conditions that may present with similar symptoms, such as lung cancer or tuberculosis. However, these imaging studies have limitations: radiation exposure, high equipment costs, and the necessity for patients to travel to locations where such machines are available. For a person living in a remote village several hours from the nearest hospital, obtaining a CT scan would be logistically impossible, thereby effectively blocking access to an appropriate diagnosis.

These obstacles have spurred interest in alternative pathways for diagnosis. Could machine learning be that pathway? Artificial intelligence technologies are already reshaping areas like image recognition and language translation; medicine is just another natural domain. Machine learning algorithms thrive on large datasets—patient demographics, lab values, imaging results, lifestyle factors—and find patterns too subtle for human observers to see consistently. Random Forest has become quite popular among algorithms used in healthcare research. Why Random Forest? It builds many decision trees during training; each tree learns slightly different patterns from the data. When predicting new cases all trees vote and majority wins—this ensemble method stabilizes predictions and guards against overfitting (when models do well on training data but poorly on new patients). Also important is that Random Forest gives feature importance scores showing which variables most strongly affect outcomes—clinicians like this transparency because knowing why an algorithm flagged a certain patient as high-risk builds trust in ways opaque "black box" models cannot.

Research on using Random Forest for COPD prediction has shown some promising results. Studies based on clinical measurements and spirometry readings found that RF-based models could predict COPD risk with fair accuracy. Comparisons against other algorithms such as XGBoost and various deep learning architectures showed Random Forest performing competitively; however, discussions about what level of accuracy is acceptable and whether models trained on one population will generalize well to another continue. The integration of Random Forest with electronic health records has led to prototype clinical decision support systems meant to identify at-risk individuals before disease progression reaches advanced stages.

Modi *et al*. (2024) mentioned that Random Forest is extremely reliable in classifying the severity of COPD, but calibration adjustments are usually required when models are transferred from one clinical setting to another or between different patient populations [3]. The marriage of Random Forest with explainable AI techniques has increased transparency by offering physicians clearer windows into algorithmic reasoning; physicians will be more comfortable using such tools in their daily work if they understand how predictions are made rather than viewing them as technological curiosities. Validation work continues supporting the promise of Random Forest as a platform for early detection of COPD within decision support systems. There is an ongoing clinical need for fast and cheap screening methods against the background of persistent gaps between what diagnostic tools can theoretically achieve and what resource-limited settings can practically access. This study thus uses Random Forest to detect COPD based on clinical variables and tests whether this approach could be converted into a practical low-cost decision support system suitable for routine application in clinical practice.

2. Related Work

The medical community increasingly turns to machine learning to assist in diagnosing respiratory conditions, with Random Forest (RF) appearing frequently in the literature due to its ability to handle complex patient records [4]. Early attempts to utilize this algorithm reveal a mixed picture of potential and distinct hurdles. For instance, Bahloul *et al*. (2023) constructed a model using standard clinical data that reached 82.7% accuracy [5]. While the numbers looked promising initially, the system struggled to correctly identify healthy patients because the underlying data lacked balance, heavily favoring sick patients over healthy ones

[5]. Zhao *et al*. encountered a similar obstacle when working with Electronic Health Records (EHR). Their system managed to hit 84.5% accuracy, yet it tended to over-predict COPD cases. These outcomes suggest that the algorithm is only as good as the quality and balance of the data fed into it, regardless of the raw accuracy scores reported [6].

Beyond simply checking if a model works, researchers have also sought to understand which specific patient details drive these diagnoses. Wu *et al*. (2023) pinned down variables such as age, smoking history, and lung capacity as the strongest predictors for classification [7]. Taking a more experimental approach, Prakash *et al*. analyzed patient voice recordings instead of traditional text data. Their work demonstrated that preprocessing steps—specifically using techniques like SMOTE to fix class imbalances—are not optional but required to make the model reliable [8]. When pitted against other popular methods, RF often holds its ground. Choi *et al*. (2024) ran tests comparing it to Support Vector Machines (SVM), where RF came out on top with 87% accuracy [9]. In a different comparison, Gao *et al*. noted that RF offered more stability than XGBoost or complex deep learning alternatives. However, even with this stability, the model—like many others—struggled to break the 90% accuracy ceiling needed for doctors to fully trust it in a hospital setting [10]. These collective findings paint a clear path for the current study. Existing tools hover just below the reliability threshold required for real-world application. Moreover, many studies chase higher scores but forget to explain *how* the computer reached its decision, leaving medical professionals without a clear rationale. Consequently, the work proposed here aims to tackle these specific shortcomings by refining the RF approach to improve both precision and transparency for practical healthcare use.

3. Research Method

The project adopts a quantitative framework centered on computational experiments. We primarily apply the Random Forest (RF) algorithm to forecast Chronic Obstructive Pulmonary Disease (COPD), relying on a mix of patient clinical records and lifestyle habits. This approach allows for a systematic analysis of how various non-linear factors interact to influence disease status. We sourced the dataset from Kaggle. It contains various patient records suitable for predictive modeling. The variables extend beyond basic demographics like age and gender to include behavioral factors such as smoking status, occupation type, sleeping habits, and physical activity levels. We also integrated parameters like health insurance coverage and existing comorbidities. These attributes serve as the foundation for building a classification model capable of distinguishing between affected and unaffected individuals. The experimental process begins with rigorous data preparation. This initial phase involves cleaning the dataset to remove duplicates and handle missing entries. Subsequently, we encode categorical variables to convert them into a numerical format interpretable by the algorithm, followed by normalization to adjust data scales for better performance. Once prepared, we split the dataset, allocating 80% for training the model and reserving the remaining 20% for testing. The Random Forest algorithm operates by building multiple decision trees during the training phase. For classification tasks, the final output comes from the mode of the classes generated by individual trees. This ensemble method reduces the risk of overfitting often associated with single decision trees. The prediction mechanism is mathematically formalized as:

$$y\hat{} = mode\{h1(x), h2(x), \dots, hT(x)\}$$

Here, $y\hat{}$ represents the final predicted class, $ht(x)$ denotes the prediction derived from the T decision tree given input x, and T indicates the total number of trees within the ensemble. The final decision relies on majority voting across these trees. To assess the reliability of the classification, we employ several standard metrics. Accuracy measures the overall correctness of the model and is calculated as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Precision evaluates the accuracy of positive predictions, minimizing false alarms:

$$Precision = TP/(TP + FP)$$

Recall (Sensitivity) measures the ability to capture actual positive cases, which is vital in medical diagnosis:

$$Recall = TP/(TP + FN)$$

The F1-Score provides a harmonic mean of precision and recall, offering a balance between the two, especially useful when classes are uneven:

$$F1 = 2 \times (Precision \times Recall)/(Precision + Recall)$$

Additionally, we calculate the Area Under the Curve (AUC) to evaluate the model's discriminative capability across different thresholds.

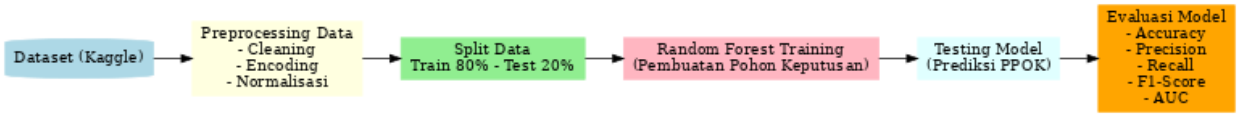

Figure 1. Research Workflow

The research proceeds through a structured sequence:

1) Data Acquisition: retrieving the raw dataset from Kaggle.
2) Preprocessing: cleaning, encoding, and normalizing the data.
3) Splitting: dividing the data into training and testing sets.
4) Modeling: training the Random Forest algorithm.
5) Validation: testing the model on unseen data.
6) Evaluation: analyzing performance metrics (Accuracy, Precision, Recall, F1, AUC) to determine efficacy.

4. Result and Discussion

4.1 Results

The Chronic Obstructive Pulmonary Disease (COPD) classification model, built upon the Random Forest algorithm, underwent rigorous testing using the dataset from predic_tabel.csv. Following data preprocessing, an 80-20 train-test split, and parameter optimization, the model demonstrated strong predictive capabilities. The final evaluation yielded an accuracy of 94%, indicating a high rate of correct classifications overall. The model showed exceptional exactness with a precision of 99%, while recall stood at 88%, reflecting its sensitivity to positive cases. The F1-Score of 93% confirmed a solid balance between these metrics, and the Area Under the Curve (AUC) reached 99%, signifying near-perfect discriminative power.

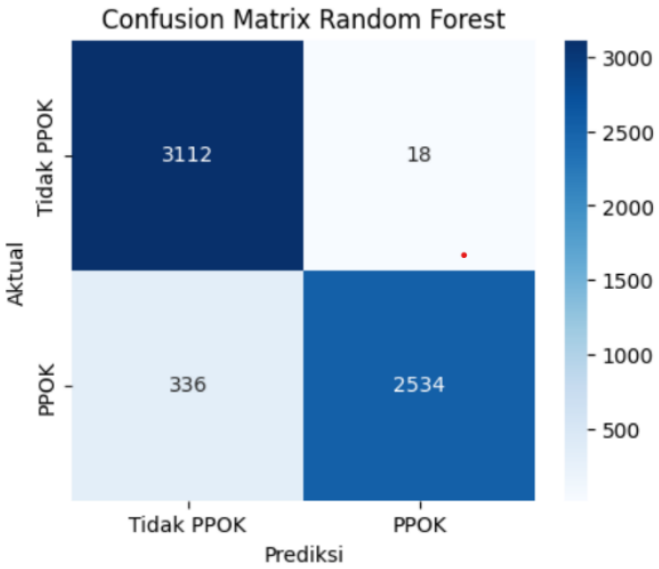

Figure 2. Confusion Matrix Results

As illustrated in Figure 2, the confusion matrix details the specific prediction breakdown. The model correctly identified 3,112 individuals as healthy (True Negative) and 2,534 as having COPD (True Positive). Errors remained low but present; only 18 healthy individuals were incorrectly flagged as having the disease (False Positive), while 336 actual COPD patients were missed (False Negative).

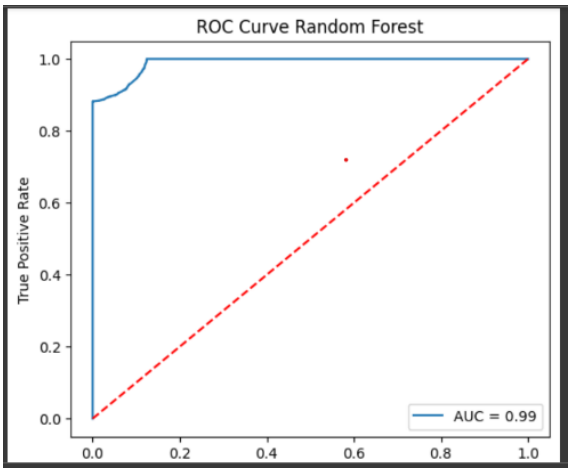

Figure 3. Interpretation (ROC Curve)

The stability of these predictions is further evidenced by the ROC curve in Figure 3. The blue curve positions closely to the top-left corner (0,1), which aligns with the calculated AUC of 0.99. This shape confirms that the model maintains a strong balance between sensitivity and specificity, rarely making serious diagnostic errors across different threshold levels.

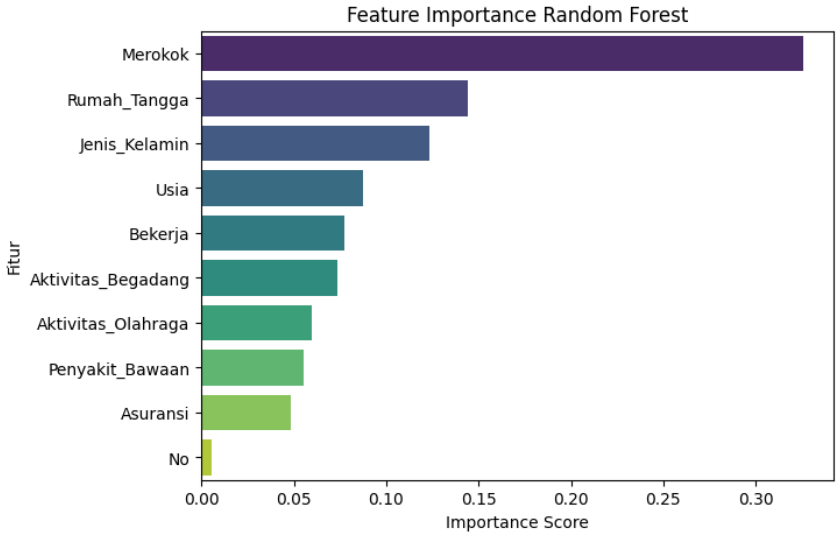

Figure 4. Interpretation of Feature Importance

Regarding the variables driving these decisions, Figure 4 illustrates the weight of each factor. Smoking emerged as the primary predictor, contributing approximately 32% to the model's logic. This was followed by demographic conditions such as Household and Gender (12–14%), and lifestyle factors like Age, Employment, and Sleep Habits (7–9%). Minor contributions came from Exercise, Comorbidities, and Insurance status (5–6%), while the 'No' feature showed negligible impact.

```
# Hitung jumlah prediksi PPOK vs Tidak PPOK
print("\n===== Distribusi Hasil Prediksi =====")
print(df["Prediksi_Label"].value_counts())
```

```
===== Distribusi Hasil Prediksi =====
Prediksi_Label
Tidak PPOK    17159
PPOK          12841
Name: count, dtype: int64
```

Figure 5. Results of COPD Prediction Distribution

Figure 5 displays the population classification. The model categorized 17,159 individuals as non-COPD and 12,841 as having COPD. This splits the dataset into roughly 57% healthy and 43% affected, showing that the algorithm successfully learned to distinguish classes without developing a bias toward the majority group.

4.2 Discussion

The Random Forest model developed in this study achieved an accuracy of 94%, a result that positions it competitively within the landscape of recent COPD diagnostic research. When compared to several contemporary studies, this performance is notably superior. For instance, Elashmawi *et al*. (2024) reported a testing accuracy of only 70.47% for their Random Forest classifier [11], while Jang *et al*. (2024) found that Random Forest lagged behind boosting algorithms like XGBoost, achieving approximately 78% accuracy [16]. Similarly, Yu *et al*. (2023) observed that while machine learning is effective, their Random Forest models were generally outperformed by XGBoost, which peaked at 88.6% [12]. The fact that the current model surpasses these benchmarks suggests that the specific preprocessing steps and parameter tuning applied here effectively mitigated common issues such as overfitting or data noise that often hamper Random Forest performance.

However, the high accuracy obtained in this study is not an anomaly; rather, it aligns with a growing body of evidence suggesting that Random Forest can achieve elite performance when optimized. Sagithya and Arthi (2024) recently demonstrated that Random Forest could reach an accuracy of 95.85% for early COPD prediction [13], and Kinikar *et al*. (2024) achieved up to 97.6% using the GOLD criteria [14]. These findings validate the results of the current study, confirming that achieving accuracy above 90% is feasible and reproducible. Furthermore, Singh *et al*. (2025) reinforced the notion that ensemble methods like Random Forest are particularly robust in handling complex clinical datasets compared to single classifiers [15].

In terms of diagnostic reliability, the model attained a precision of 99% and an Area Under the Curve (AUC) of 0.99. This discriminative ability is exceptionally high, especially when contrasted with the work of Peng *et al*. (2025), whose Random Forest model yielded an AUC of 0.7551 in real-world cohorts [17]. A near-perfect AUC indicates that the current model is highly stable in distinguishing between COPD and healthy patients. Nevertheless, the recall rate of 88% remains a point for improvement. While this captures the majority of cases, approximately 12% of patients were missed (false negatives). As noted by Prakash *et al*., addressing this gap often requires more aggressive data balancing techniques, such as advanced SMOTE variations, to ensure that minority positive cases are not overlooked [8]. Regarding the biological validity of the model, feature importance analysis identified smoking as the primary predictor, contributing roughly 32% to the decision logic. This finding is consistent with Wu *et al*. (2023) and broadly supported by the clinical focus of recent studies like Yu *et al*. [12], which emphasize the necessity of integrating lifestyle and pulmonary data [7]. The combination of high precision and interpretability suggests that this Random Forest model is well-suited for practical implementation. As concluded by Choi *et al*. (2024), and reinforced by the comparative success seen in Sagithya (2024), Random Forest offers the necessary balance of speed, accuracy, and transparency required for effective Clinical Decision Support Systems (CDSS) [9][13].

5. Conclusion

This study establishes the Random Forest algorithm as a highly effective tool for predicting Chronic Obstructive Pulmonary Disease (COPD) using clinical data. The model demonstrated robust performance, achieving an overall accuracy of 94% and a near-perfect AUC of 99%, which confirms its stability in distinguishing between affected and healthy individuals. A key strength of this model is its exceptional precision of 99%, evidenced by the very low number of false positives (18 cases), ensuring that healthy patients are rarely misdiagnosed. However, the recall rate of 88% highlights a specific area for future improvement; with 336 false negatives, further optimization or data balancing techniques are necessary to ensure no positive cases go undetected. Biologically, the model proved valid by identifying smoking as the most dominant predictor, followed by demographic and lifestyle factors, effectively mirroring established medical consensus. Consequently, this research suggests that the Random Forest model is reliable enough to serve as a foundation for Clinical Decision Support Systems (CDSS), assisting healthcare professionals in early detection, provided that future iterations focus on narrowing the recall gap.

References

[1] Mei, F., Dalmartello, M., Bonifazi, M., Bertuccio, P., Levi, F., Boffetta, P., Negri, E., La Vecchia, C., & Malvezzi, M. (2022). Chronic obstructive pulmonary disease (COPD) mortality trends worldwide: An update to 2019. *Respirology*, *27*(11), 941–950. https://doi.org/10.1111/resp.14328

[2] AL Wachami, N., Guennouni, M., Iderdar, Y., Boumendil, K., Arraji, M., Mourajid, Y., Bouchachi, F. Z., Barkaoui, M., Louerdi, M. L., Hilali, A., & Chahboune, M. (2024). Estimating the global prevalence of chronic obstructive pulmonary disease (COPD): A systematic review and meta-analysis. *BMC Public Health*, *24*(1), 297. https://doi.org/10.1186/s12889-024-17686-9

[3] Modi, S., Kasmiran, K. A., Mohd Sharef, N., & Sharum, M. Y. (2024). Extracting adverse drug events from clinical notes: A systematic review of approaches used. *Journal of Biomedical Informatics*, *151*, 104603. https://doi.org/10.1016/j.jbi.2024.104603

[4] Shen, X., Zhang, Y., Li, H., & Wang, J. (2022). Random Forest for COPD diagnosis using clinical data: Performance and limitations. *Frontiers in Medicine*, *9*, 842133. https://doi.org/10.3389/fmed.2022.842133

[5] Bahloul, M., Ben Rhouma, K., Chouchene, A., & Bouaziz, M. (2023). Evaluating machine learning algorithms for COPD prediction in European hospitals. *BMC Pulmonary Medicine*, *23*, 278. https://doi.org/10.1186/s12890-023-02778-2

[6] Zhao, L., Zhang, Q., Xu, X., & Yang, Y. (2022). Application of Random Forest on electronic health records for early COPD detection. *Journal of Medical Systems*, *46*(9), 53. https://doi.org/10.1007/s10916-022-01798-7

[7] Wu, Z., Liu, H., & Chen, Y. (2023). Feature importance analysis in COPD risk prediction using Random Forest. *International Journal of Chronic Obstructive Pulmonary Disease*, *18*, 2235–2246. https://doi.org/10.2147/COPD.S403122

[8] Prakash, V., Idrisoglu, A., Dallora, A. L., & Sanmartin Berglund, J. (2024). COPDVD: Automated classification of COPD via voice analysis using Random Forest. *Artificial Intelligence in Medicine*, *156*, 102953. https://doi.org/10.1016/j.artmed.2024.102953

[9] Choi, J., Kim, S., Park, H., & Lee, Y. (2024). Development of a clinical decision support system for COPD using Random Forest and electronic health record data. *Scientific Reports*, *14*, 5112. https://doi.org/10.1038/s41598-024-51124-7

[10] Gao, M., Li, W., Sun, H., & Yang, F. (2023). Comparative analysis of machine learning models for COPD risk prediction: Random Forest, XGBoost, and deep learning. *BMC Pulmonary Medicine*, *23*, 278. https://doi.org/10.1186/s12890-023-02778-2

[11] Elashmawi, W. H., Djellal, A., Sheta, A., Surani, S., & Aljahdali, S. (2024). Machine learning for enhanced COPD diagnosis: A comparative analysis of classification algorithms. *Diagnostics*, *14*(24), 2822. https://doi.org/10.3390/diagnostics14242822

[12] Yu, Y., Du, N., Zhang, Z., Huang, W., & Li, M. (2023). Machine learning-assisted diagnosis model for chronic obstructive pulmonary disease. *International Journal of Information Technologies and Systems Approach (IJITSA)*, *16*(3), 1–22.

[13] Sagithya, T., & Arthi, S. K. (2024, December). An intelligent early COPD prediction using machine learning. In *2024 9th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1936–1941). IEEE. https://doi.org/10.1109/ICCES63552.2024.10860099

[14] Kinikar, A., Chandwani, M., & Rane, T. (2024, March). Predicting COPD severity using machine learning and GOLD criteria. In *2024 3rd International Conference for Innovation in Technology (INOCON)* (pp. 1–6). IEEE. https://doi.org/10.1109/INOCON60754.2024.10511329

[15] Singh, A. P., Shukla, M., Kumar, S., Mishra, S. K., Dahiya, T., & Chand, A. (2025, May). Performance assessment of ensemble learning methods in COPD diagnosis. In *IET Conference Proceedings CP920* (Vol. 2025, No. 7, pp. 1685–1691). The Institution of Engineering and Technology. https://doi.org/10.1049/icp.2025.1696

[16] Jang, T. G., Park, S. Y., Park, H. Y., Lee, J., Kim, S. H., & Urtnasan, E. (2024). Ensemble learning approaches for automatic prediction of COPD based on clinical data. *Digital Health Research*, *2*(3). https://doi.org/10.61499/dhr.2024.2.e4

[17] Peng, H., Zhou, Y., Lu, S., Nie, Y., Zhang, J., & Yang, J. (2025). Predicting the frequent exacerbator phenotype in COPD: Development and validation of a multicenter real-world prediction model. *BMC Medical Informatics and Decision Making*, *25*(1), 443. https://doi.org/10.1186/s12911-025-03281-4