

TWO-STAGE XGBOOST FRAMEWORK FOR SPIROMETRY-BASED CLASSIFICATION OF ASTHMA AND COPD AND GOLD SEVERITY PREDICTION

¹Tejaskumar V. Bhatt, ²Bhaveshkumar P. Patel

¹Associate Professor, Department of Biomedical Engineering, U. V. Patel College of Engineering, Faculty of Engineering & Technology, Ganpat University – 384012. Gujarat, India. Email: tejas.bhatt@ganpatuniversity.ac.in, ORCID ID:0000-0003-3016-5693

²Professor, Department of Mechanical Engineering, U. V. Patel College of Engineering, Faculty of Engineering & Technology, Ganpat University – 384012. Gujarat, India. Email: bpp01@ganpatuniversity.ac.in, ORCID ID:0000-0003-0907-4981

*Corresponding Author: Bhaveshkumar P. Patel, E-mail: bpp01@ganpatuniversity.ac.in

ABSTRACT

This study describes an explainable two-step XGBoost machine-learning pipeline for classifying spirometry reports as Normal, Asthma or Chronic Obstructive Pulmonary Disease followed by COPD staging according to the Global Initiative for Chronic Obstructive Lung Disease system. The study analysed 1600 spirometry reports, including 851 Normal, 320 Asthma and 429 COPD cases. Authors used demographic data, pre- and post-bronchodilator measurements and bronchodilator response features as inputs. After data quality checks and leakage audits, we split the dataset in an 80:20 ratio to train and test the model. Authors applied randomized hyperparameter optimization and 5-fold stratified cross-validation during the training stage. To compare the performance, we trained two models: one used a full-report set of 22 features, while the other employed a pre-bronchodilator feature set of 13 features. The full report model demonstrated an accuracy of 99.4% and a macro-F1 score of 99.1%, outperforming the conventional bronchodilator reversibility rule. The pre-bronchodilator model achieved an accuracy of 91.6%, a macro-averaged ROC-AUC of 0.988 and a recall of 98.8% for COPD, with no COPD cases incorrectly classified as Normal in the test set. Finally, the second-step model correctly classified all 86 COPD cases in the test set, while the entire pipeline correctly classified six diagnostic and severity categories with an accuracy of 99.4%. The study revealed that post-bronchodilator FEV1/FVC, bronchodilator response and smoking history were the most relevant features for diagnosing Asthma, COPD or Normal spirometry, while post-bronchodilator FEV1 percentage predicted was a critical determinant for COPD staging. In conclusion, the study proved the efficacy of the explainable machine learning model for classifying and staging COPD.

KEYWORDS: asthma, chronic obstructive pulmonary disease, spirometry, GOLD classification, XGBoost

1. INTRODUCTION

Chronic obstructive pulmonary disease is the third leading cause of death worldwide, causing 3.4 million deaths in 2023, with most deaths before the age of 70 occurring in low- and middle-income countries [1]. Chronic respiratory diseases have a huge burden in India that varies considerably between states [2]. Asthma is the other common obstructive disease and the two conditions share symptoms such as breathlessness, cough and wheeze. Their treatment differs, so an accurate distinction matters for patients [3 4].

Spirometry is the objective test at the centre of both diagnoses [5 6]. COPD is defined by a post-bronchodilator FEV1/FVC below 0.70 and its airflow limitation is graded GOLD 1 to 4 by post-bronchodilator FEV1 as a percentage of the predicted value [3]. Asthma is supported by variable airflow, classically an FEV1 increase of more than 12% and 200 mL after a bronchodilator [4]. These fixed thresholds are simple but imperfect. The 0.70 ratio misclassifies obstruction at the extremes of age [7], some people with asthma keep a low ratio after bronchodilation, and some people with COPD show a large bronchodilator response [3 4]. Interpretation also varies between clinicians: in a multicentre study, pulmonologists matched guideline-based interpretation of lung function tests in 74% of cases on average, against 100% for a machine-learning system [8].

Machine learning has therefore been proposed to support spirometry interpretation. Topalovic et al. trained a model on complete lung function tests that outperformed pulmonologists in both pattern recognition and diagnosis [8]. More recently, the Asthma/COPD Differentiation Classification tool, developed from routine-care records, reached 73% accuracy against an expert panel compared with 61% for pulmonologists and 50% for primary care physicians [9]. Among learning methods, gradient-boosted decision trees such as XGBoost [10] remain strong on tabular clinical data and often match or beat deep neural networks on medium-sized tables [11]. Tools based on Shapley values can explain their individual predictions [12 13].

Three gaps motivate this study. First, most reports judge a model in isolation rather than against the guideline rules that clinicians actually apply. Second, data leakage, when an input reveals the answer, is rarely checked in spirometry datasets although it can inflate performance. Third, diagnosis and severity grading are usually studied separately, whereas clinical practice applies them in sequence.

This study proposes an interpretable two-step XGBoost diagnostic framework that accurately classifies Normal, Asthma, and COPD patients, followed by GOLD-standard COPD staging. To facilitate clinical data leakage, we perform data-quality audit and design a pre-bronchodilator-only model variant that eliminates post-test confounders. Our pipeline is rigorously benchmarked against three clinical guideline rules using bootstrapped confidence-intervals and exact McNemar tests for diagnostic utility.

2. MATERIALS AND METHODS

2.1 Study design and data

This study was performed using one thousand six hundred adult spirometry reports. For each report, the patient's age, sex, height, weight, body mass index and smoking status - never, ex-smoker or current were recorded, along with pre-bronchodilator forced vital capacity, FEV1 and FEV1/FVC with predicted values and percent predicted. For patients with abnormal results, post-bronchodilator FEV1, FEV1 percent predicted, absolute and percent change in FEV1 were also recorded. Each report was assigned a final diagnosis as Normal, Asthma or COPD and for COPD, a GOLD grade with severity label as shown in Figure 1. The study followed the TRIPOD+AI reporting guideline for prediction models [14].

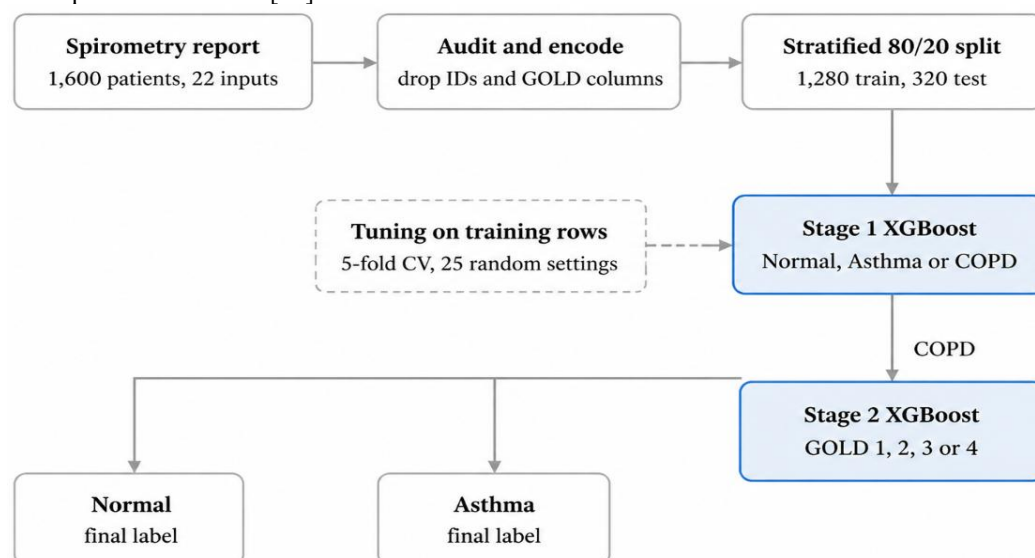


Fig. 1: Schematic overview of the eight-step study pipeline

2.2 Data-quality and leakage audit

Firstly, ensured that FEV1 did not exceed FVC; that the stored ratio, bronchodilator change, per cent predicted values and BMI matched their raw inputs; that every COPD patient met the GOLD criterion of a post-bronchodilator FEV1/FVC below 0.70 [3]; and that each GOLD grade matched the post-bronchodilator FEV1 % predicted cut-offs ($\geq 80\%$, $50\text{--}79\%$, $30\text{--}49\%$, $< 30\%$) [3]. Authors also checked for whether the pattern of missing values had a relationship with the outcome, a known source of leakage [15].

2.3 Preprocessing and feature sets

Identifiers and both GOLD columns were removed from the Stage 1 inputs. Sex was coded female = 0 and male = 1 and smoking status as an ordinal value as never = 0, ex-smoker = 1 and current = 2. Missing post-bronchodilator values were left as missing, since XGBoost learns a default branch for them at each split [10]. No scaling was applied. As the post-bronchodilator test was missing for every Normal patient, two Stage 1 feature sets were defined in advance:

- Set A, full report (22 inputs): demographics, pre- and post-bronchodilator values and bronchodilator response.
- Set B, pre-bronchodilator (13 inputs): demographics and pre-bronchodilator values only, thus removing the missing-test signal and mimicking a baseline only test.

Stage 2 used Set A, because GOLD grading requires post-bronchodilator FEV1 [3].

2.4 Model development

Both stages of our approach used XGBoost, which sequentially adds shallow regression trees, each fitted to the gradient of the remaining loss [10]. The model minimises a regularised objective:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

where l is the multi-class log-loss, T the number of leaves and w the leaf weights [10]. A softmax objective gave class probabilities. Training rows were weighted inversely to class frequency. Stage 1 predicted Normal, Asthma or COPD; Stage 2 predicted GOLD 1–4 and was trained on COPD patients only.

2.5 Validation and tuning

Patients were split once into a training set (80%, $n = 1280$), and a test set (20%, $n = 320$), stratified by diagnosis. Stage 2 inherited this split: COPD patients in the training set ($n = 343$) were used for training, and those in the test set ($n = 86$) for testing, so that no patient appeared in both. Within the training set, a randomised search over 25 combinations of eight hyper-parameters selected the setting with the best macro-F1 in 5-fold stratified cross-validation [16–17]. The selected models were then evaluated once on the test set, which avoids the optimistic bias of reporting the selection score [18]. Finally, Stage 1 predictions were passed to Stage 2 to evaluate the full pipeline over six labels - Normal, Asthma, GOLD 1 to 4.

2.6 Rule baselines

Three guideline-like rules were applied to the same sets of test patients [3,4]:

- Rule A: FEV1 increase $\geq 12\%$ and ≥ 0.20 L $\rightarrow$ asthma; otherwise, pre-bronchodilator FEV1/FVC $< 70\% \rightarrow$ COPD; else normal
- Rule B: similar to Rule A, but with a 10% cut-off and no volume criterion
- Rule C (GOLD definition): no post-bronchodilator $\rightarrow$ normal; post-bronchodilator FEV1/FVC $< 70\% \rightarrow$ COPD; else asthma

For Stage 2, the GOLD cut-offs were directly applied to % predicted FEV1 post-bronchodilator as the baseline.

2.7 Statistical analysis

For the cohort demographics, medians or counts are reported, compared with Kruskal–Wallis test for continuous and chi-square test for categorical variables. The test-set results were summarised by accuracy, balanced accuracy, macro-F1, Cohen’s κ and one-versus-rest macro-ROC–AUC. Accuracy, macro-F1 and ROC–AUC were accompanied by 95% percentile bootstrap confidence intervals from 2000 resamples [19]. Pairwise comparisons between models and rules were made using the exact McNemar test on discordant predictions [20–21]. For stability assessment, the tuned rules were re-evaluated using a 10 $\times$ 5-fold repeated stratified cross-validation and reported as mean $\pm$ standard deviation over all eligible patients. A two-sided P value < 0.05 was considered significant.

2.8 Explainability

SHAP values were calculated for every prediction by decomposing it into additive feature contributions using Shapley values [12]. These were calculated exactly using TreeExplainer [13] and the mean absolute SHAP value on the test set was used to rank the features for every class.

3. RESULTS

3.1 Data quality and cohort

The data passed all audit tests. FEV1 was never greater than FVC in any report, all derived data corresponded to their raw inputs, all COPD patients had a post-bronchodilator FEV1/FVC below 0.70, and all GOLD grades corresponded to their post-bronchodilator cut-offs. Pre-bronchodilator FEV1 % predicted would have correctly graded only 78.1% of COPD patients. Table 1 represents the patient characteristics across diagnostic group. There was one discrepancy found in the audit: 851 out of all 1600 Normal patients did not have a post-bronchodilator test, whereas all the other 749 patients did have one.

Table 1: Cohort characteristics across diagnostic groups

Characteristic	Normal (n = 851)	Asthma (n = 320)	COPD (n = 429)	p
Age, years	62.0 (51.0–73.0)	57.0 (49.0–68.0)	67.0 (57.0–75.0)	< 0.001
Male sex	480 (56.4)	136 (42.5)	286 (66.7)	< 0.001
BMI, kg/m ²	25.9 (22.6–29.1)	27.0 (23.3–30.4)	23.9 (20.4–27.4)	< 0.001
Never smoker	567 (66.6)	221 (69.1)	0 (0.0)	< 0.001
Ex-smoker	221 (26.0)	60 (18.8)	188 (43.8)	< 0.001
Current smoker	63 (7.4)	39 (12.2)	241 (56.2)	< 0.001
Pre-BD FEV1, % predicted	94.5 (87.0–102.1)	75.8 (68.7–82.0)	49.2 (34.6–58.5)	< 0.001
Pre-BD FVC, % predicted	97.9 (92.7–103.5)	88.9 (82.5–95.1)	78.8 (69.1–86.2)	< 0.001
Pre-BD FEV1/FVC, %	81.9 (77.2–86.8)	69.6 (66.2–73.4)	50.0 (36.6–65.8)	< 0.001
Post-BD FEV1, % predicted	not tested	88.5 (80.4–95.9)	52.7 (36.9–71.0)	< 0.001

Post-BD FEV1/FVC, %	not tested	77.1 (73.0–81.8)	60.5 (52.1–67.0)	< 0.001
BD change in FEV1, %	not tested	16.8 (13.1–20.9)	5.6 (3.8–7.3)	< 0.001

Of the patients included with COPD, 64 (14.9%) were GOLD 1, 166 (38.7%) GOLD 2, 138 (32.2%) GOLD 3 and 61 (14.2%) GOLD 4. There were two clinically significant subgroups that crossed traditional thresholds as given in Figure 2. Twenty-six COPD patients (6.1%) had a bronchodilator response of 10% or more, to a maximum of 76.6%. Forty-three asthma patients (13.4%) had post-bronchodilator FEV1/FVC below 0.70.

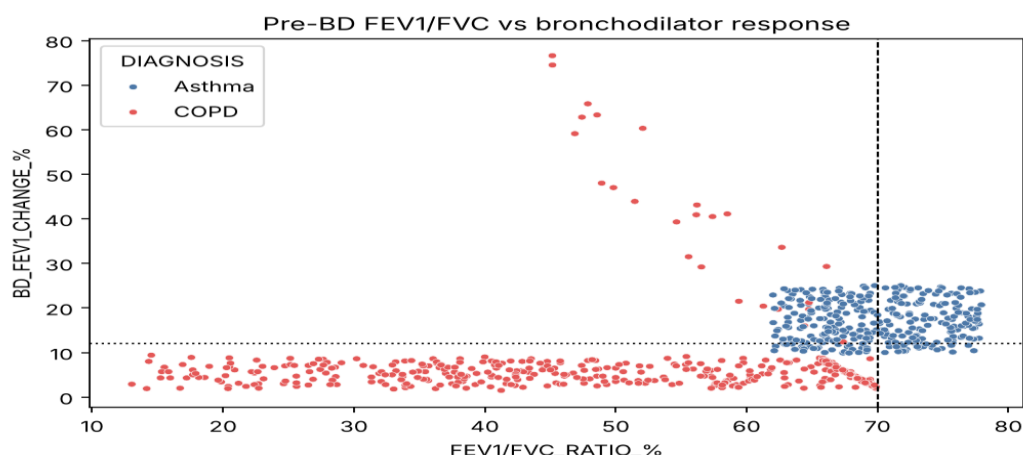


Fig. 2: Pre-bronchodilator FEV1/FVC versus bronchodilator response

3.2 Stage 1: diagnosis

The XGBoost model misclassified 2 out of 320 patients. Both were COPD patients with large bronchodilator responses (16% and 47%) who were labelled Asthma. XGBoost was significantly more accurate than Rule A, correctly classifying instances where Rule A failed in 17 patients and failing where Rule A succeeded in only 1 ($p < 0.001$). Its superiority over Rule C (8 versus 2 discordant patients, $p = 0.11$) and Rule B (2 versus 0, $p = 0.50$) was not statistically significant. Rule C failed predominantly on asthma with persistent obstruction, and Rule B on COPD with high reversibility.

Table 2. Stage 1 test set performance

Method	Inputs	Accuracy	Balanced accuracy	Macro-F1	Cohen's kappa	Macro ROC-AUC	McNemar p vs Set A
XGBoost, Set A (full report)	22	0.994 (0.984–1.000)	0.992	0.991 (0.977–1.000)	0.990	1.000 (1.000–1.000)	–
XGBoost, Set B (pre-BD only)	13	0.916 (0.884–0.947)	0.913	0.903 (0.868–0.936)	0.863	0.988 (0.981–0.994)	–
Rule A (12% and 200 mL)	3	0.944 (0.916–0.969)	0.909	0.919 (0.881–0.953)	–	–	< 0.001
Rule B (10%)	2	0.988 (0.975–0.997)	0.983	0.982 (0.962–0.996)	–	–	0.50
Rule C (GOLD definition)	2	0.975 (0.956–0.991)	0.958	0.963 (0.935–0.986)	–	–	0.11

Using pre-bronchodilator values only, accuracy dropped to 0.916 but remained high for discrimination (ROC-AUC 0.988). Table 2 represents the Stage 1 performance on the held-out test set. COPD was identified with a recall of 0.988 and an F1 of 0.977, with no COPD patient labelled Normal. Most errors occurred between Normal and Asthma: 16 Normal patients were labelled Asthma and 7 Asthma patients Normal as shown in Figure 3. Asthma therefore had the lowest F1 (0.800; precision 0.761, recall 0.844).

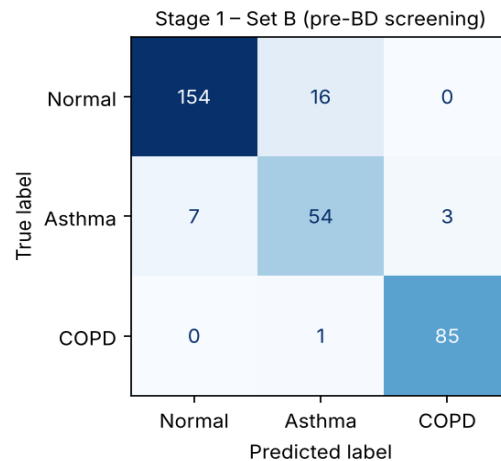


Fig. 3: Confusion matrix for the Set B – pre bronchodilator test model

3.3 Stage 2 and the full pipeline

The GOLD model correctly classified all 86 test COPD patients (accuracy and macro-F1 1.000; GOLD 1, n = 10; GOLD 2, n = 33; GOLD 3, n = 28; GOLD 4, n = 15). Applying the GOLD cut-offs directly to the post-bronchodilator FEV1 % predicted was also accurate. Run in sequence, the two stages achieved an accuracy of 0.994 and a macro-F1 of 0.986 across six labels. The only two errors were the Stage 1 misclassifications described above, one GOLD 1 and one GOLD 2 patient, who never reached Stage 2 as shown in Figure 4.

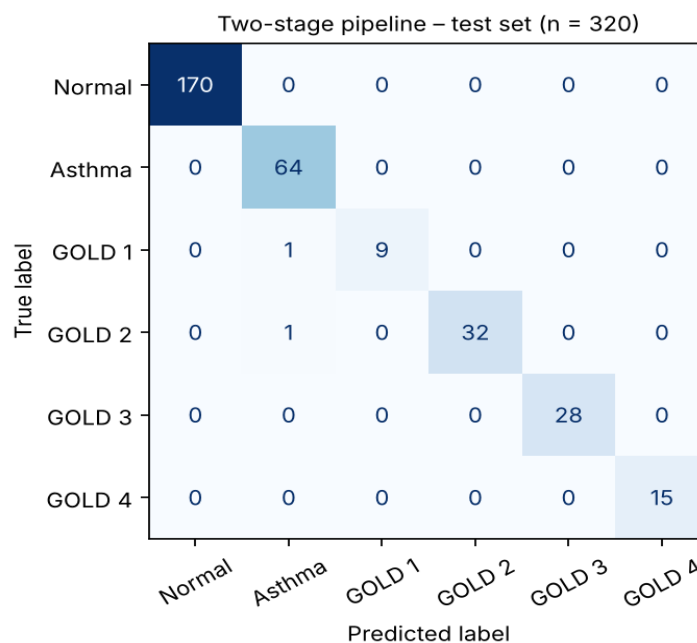


Fig. 4: Confusion matrix of the full two-stage pipeline on the test set

3.4 Stability

Repeated cross-validation produced results close to the single test set with a small variance between folds. 10 × 5-fold repeated stratified cross validation results are given in Table 3. Also, since the settings were optimized on the training portion of the same set, these results might be slightly overestimated.

Table 3: Tuned model performance across 10 × 5-fold repeated stratified cross-validation

Model	Patients	Accuracy	Macro-F1
Stage 1, Set A	1,600	0.997 ± 0.003	0.996 ± 0.005
Stage 1, Set B	1,600	0.933 ± 0.012	0.915 ± 0.017
Stage 2, GOLD	429	0.992 ± 0.011	0.991 ± 0.013

3.5 Explainability

For Set A, a post-bronchodilator column (POST_FVC_L) predominantly carried Normal decision through its missing value that confirmed the leak detected in the audit is given in Figure 5. Post bronchodilator FEV1/FVC dictated the COPD decision, whereas the bronchodilator response dominated the Asthma decision, and the smoking status added significance towards COPD. For Set B, the pre-bronchodilator FEV1/FVC dominated all

three classes; FEV1 % predicted and smoking status were primarily relevant for COPD, whereas the age, height and FVC % predicted contributed slightly to Asthma is shown in Figure 6. In Stage 2, the post-bronchodilator FEV1 % predicted was accountable for nearly all of the decision with a minor role for pre-bronchodilator FEV1 % predicted.

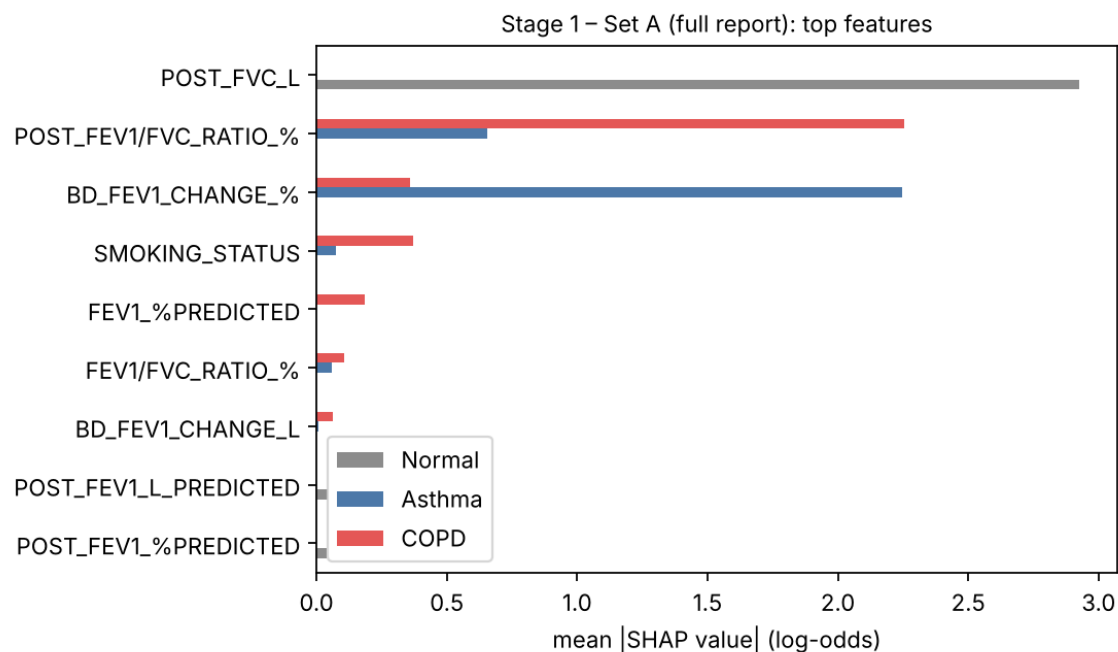


Fig. 5: Mean absolute SHAP values by class - Set A

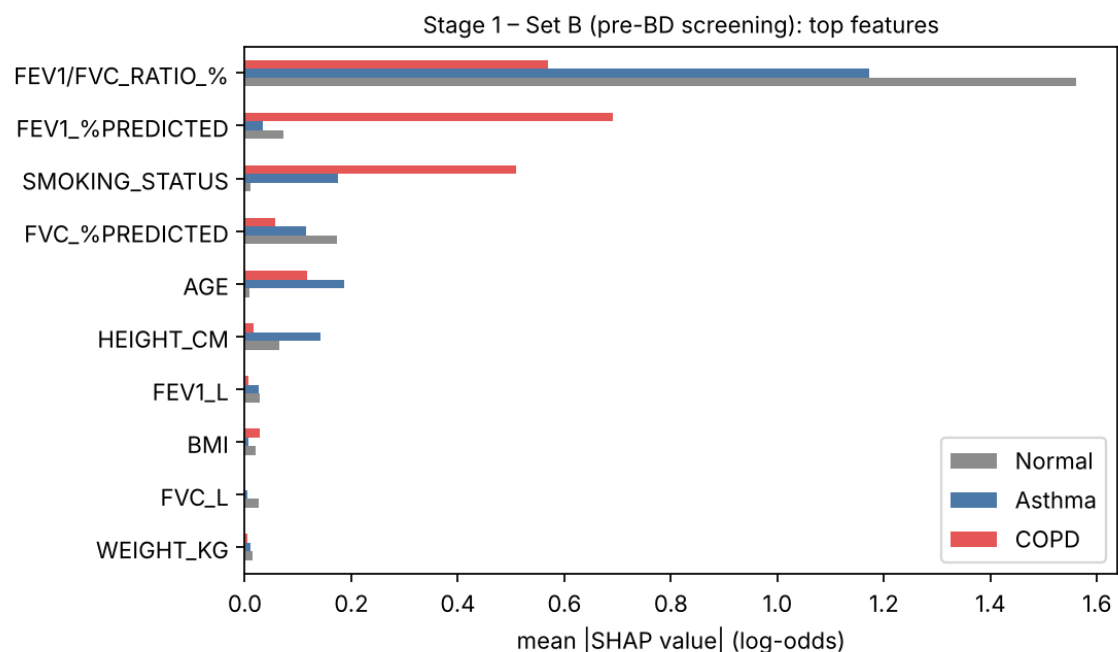


Fig. 6: Mean absolute SHAP values by class - Set B

4. DISCUSSION

This study shows that an explainable two-stage XGBoost pipeline accurately reproduces spirometry-based diagnoses and GOLD severity grading (0.994 accuracy, $n = 320$), substantially outweighing standard reversibility rules (12% and 200 mL) and halving the errors of the best simple guideline rule. With pre-bronchodilator spirometry alone, XGBoost maintains utility for diagnosing COPD (0.988 recall and ROC-AUC) and thus can serve as a triage tool in primary care to identify patients requiring further workup without mistakenly labelling COPD as Normal [1 2]. SHAP feature attributions track clinical intuition across post-BD ratios, reversibility, smoking history, and post-BD FEV%; in combination, these factors enable the XGBoost model to distinguish among difficult cases falling across static single-threshold rules, such as longstanding asthma with fixed airflow limitation [4] or COPD with marked bronchodilator response [3]. Though prior work reported 73%-82% accuracy against expert panels [8 9], our approach achieves higher accuracy by learning standardized spirometric criteria

[3]. Our strength lies in the pre-specified data audit to prevent feature leakage, repeated cross-validation, and rigorous paired benchmarking with bootstrap confidence intervals. Clinical implementation should consider that ground-truth labels were derived from spirometric definitions rather than expert consensus, normal controls lacked post-BD testing [15], diagnoses were based on a fixed 0.70 ratio rather than lower-limit-of-normal z-scores [6 7 22], and the evaluation lacks external validation and other clinical variables such as symptoms, eosinophil counts, and imaging [3 4].

5. CONCLUSION

An explainable two-stage XGBoost pipeline achieved an accuracy of 99.4% when classifying spirometry reports into Normal, Asthma or COPD and correctly graded COPD by GOLD on held out patients. It outperformed the standard reversibility rule and was able to correctly classify atypical patients who have COPD with high reversibility and asthma with persistent obstruction better than a single threshold rule. A single model using only baseline spirometry correctly identified patients with COPD and could be used to triage patients where bronchodilator testing is not available. The pipeline requires external validation against diagnostic categories and age predicted reference values before clinical deployment.

REFERENCES

1. World Health Organization. (2026, June 10). Chronic obstructive pulmonary disease (COPD) [Fact sheet]. [https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-\(copd\)](https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd))
2. India State-Level Disease Burden Initiative CRD Collaborators. (2018). The burden of chronic respiratory diseases and their heterogeneity across the states of India: The Global Burden of Disease Study 1990–2016. *The Lancet Global Health*, 6(12), e1363–e1374. [https://doi.org/10.1016/S2214-109X\(18\)30409-1](https://doi.org/10.1016/S2214-109X(18)30409-1)
3. Global Initiative for Chronic Obstructive Lung Disease. (2025). Global strategy for the prevention, diagnosis and management of COPD: 2026 report. <https://goldcopd.org/2026-gold-report-and-pocket-guide/>
4. Global Initiative for Asthma. (2026). Global strategy for asthma management and prevention (2026 update). <https://ginasthma.org/2026-gina-strategy-report/>
5. Graham, B. L., Steenbruggen, I., Miller, M. R., Barjaktarevic, I. Z., Cooper, B. G., Hall, G. L., Hallstrand, T. S., Kaminsky, D. A., McCarthy, K., McCormack, M. C., Oropez, C. E., Rosenfeld, M., Stanojevic, S., Swanney, M. P., & Thompson, B. R. (2019). Standardization of spirometry 2019 update: An official American Thoracic Society and European Respiratory Society technical statement. *American Journal of Respiratory and Critical Care Medicine*, 200(8), e70–e88. <https://doi.org/10.1164/rccm.201908-1590ST>
6. Stanojevic, S., Kaminsky, D. A., Miller, M. R., Thompson, B., Aliverti, A., Barjaktarevic, I., Cooper, B. G., Culver, B., Derom, E., Hall, G. L., Hallstrand, T. S., Leuppi, J. D., MacIntyre, N., McCormack, M., Rosenfeld, M., & Swenson, E. R. (2022). ERS/ATS technical standard on interpretive strategies for routine lung function tests. *European Respiratory Journal*, 60(1), Article 2101499. <https://doi.org/10.1183/13993003.01499-2021>
7. Swanney, M. P., Ruppel, G., Enright, P. L., Pedersen, O. F., Crapo, R. O., Miller, M. R., Jensen, R. L., Falaschetti, E., Schouten, J. P., Hankinson, J. L., Stocks, J., & Quanjer, P. H. (2008). Using the lower limit of normal for the FEV1/FVC ratio reduces the misclassification of airway obstruction. *Thorax*, 63(12), 1046–1051. <https://doi.org/10.1136/thx.2008.098483>
8. Topalovic, M., Das, N., Burgel, P.-R., Daenen, M., Derom, E., Haenebalcke, C., Janssen, R., Kerstjens, H. A. M., Liistro, G., Louis, R., Ninane, V., Pison, C., Schlessers, M., Vercauter, P., Vogelmeier, C. F., Wouters, E., Wynants, J., & Janssens, W. (2019). Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *European Respiratory Journal*, 53(4), Article 1801660. <https://doi.org/10.1183/13993003.01660-2018>
9. Kocks, J. W. H., Cao, H., Holzhauer, B., Kaplan, A., FitzGerald, J. M., Kostikas, K., Price, D., Reddel, H. K., Tsiligianni, I., Vogelmeier, C. F., Bostel, S., & Mastoridis, P. (2023). Diagnostic performance of a machine learning algorithm (Asthma/Chronic Obstructive Pulmonary Disease [COPD] Differentiation Classification) tool versus primary care physicians and pulmonologists in asthma, COPD, and asthma/COPD overlap. *The Journal of Allergy and Clinical Immunology: In Practice*, 11(5), 1463–1474. <https://doi.org/10.1016/j.jaip.2023.01.017>
10. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
11. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 507–520). Curran Associates.
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.
13. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
14. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated

- guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, Article e078378. <https://doi.org/10.1136/bmj-2023-078378>
15. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. <https://doi.org/10.1145/2382577.2382579>
 16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
 17. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.
 18. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
 19. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
 20. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>
 21. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923. <https://doi.org/10.1162/089976698300017197>
 22. Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., Enright, P. L., Hankinson, J. L., Ip, M. S. M., Zheng, J., & Stocks, J. (2012). Multi-ethnic reference values for spirometry for the 3–95-yr age range: The global lung function 2012 equations. *European Respiratory Journal*, 40(6), 1324–1343. <https://doi.org/10.1183/09031936.00080312>