INTERNATIONAL JOURNAL OF INFORMATIVE & FUTURISTIC RESEARCH (IJIFR)
ISSN: 2347-1697
VOLUME 13||ISSUE 7|| MARCH 2026
An Open Access, Peer Reviewed, Cross Reference DOI & Google Scholar Indexed International Research Journal
IJIFR Impact Factor= 8.057 & Approved as per UGC Rules 2025
Estd Year: 2013 || Website: www.ijifr.org || Email: editors@ijifr.org
Availaible Online at: https://www.ijifr.org/

DOI: https://doi.org/10.64672/IJIFR/26.03.13.07.031 PUBLISHED ON: MARCH 30, 2026

COPD DATASETS ANALYSIS USING AI-DRIVEN ALGORITHMS

S. Parthasarathi[1], V. Lakshmi Narasimhan[2], D. Rammohan*[1]
[1] CMR College of Engineering & Technology, Hyderabad/Bhagyanagar, Telangana, India
[2]Elizabeth City State University, NC 27909, USA

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a progressive respiratory condition affecting over 300 million people globally and remains significantly underdiagnosed due to reliance on traditional spirometry-based screening. This paper presents a comparative study of three state-of-the-art gradient boosting algorithms—XGBoost, LightGBM, and CatBoost—for automated COPD severity classification on real clinical tabular data, benchmarked against a 1D Convolutional Neural Network (CNN) baseline. Additionally, a Stacking Ensemble combining all three gradient boosting models via a Logistic Regression meta-learner is proposed as a novel hybrid approach. Experiments are conducted on the Kaggle COPD Student Dataset comprising 101 patients and 19 clinical features. The experimental evaluation demonstrates that gradient boosting algorithms significantly outperform the 1D CNN baselines for COPD severity classification using small clinical tabular data. Among all models, XGBoost achieved the best overall performance with 90.48% accuracy, 89.76% weighted F1-score, and 96.83% AUC-ROC, indicating superior predictive capability and balanced classification across severity levels. CatBoost achieved the highest discriminative ability with a near-perfect 99.31% AUC-ROC, while LightGBM delivered stable and competitive results with 85.71% accuracy and 95.08% AUC-ROC. In contrast, the CNN baseline performed poorly, achieving only 42.86% accuracy and 25.71% F1-score, confirming that deep learning architectures are less effective for small structured datasets. The Stacking Ensemble achieves competitive performance (85.71% accuracy, 97.9% AUC-ROC), confirming ensemble combination as a viable research direction. Five-fold cross-validation further confirmed model robustness, with gradient boosting models achieving mean F1-scores above 83%, substantially higher than the CNN baseline. SHAP explainability analysis consistently identified FEV1PRED, FEV1, and FVCPRED as the most influential clinical predictors, aligning with established COPD severity guidelines and supporting the clinical validity of the models.

KEYWORDS: ***COPD severity classification, XGBoost, LightGBM, Catboost, Convolutional Neural Network, SHAP, Explainable AI, Machine Learning, Gradient Boosting, Respiratory Disease.***

PAPER CITATION:
Parthasarathi,S.,Narasimhan,V.L.,Rammohan,D.: "COPD Datasets Analysis Using AI-Driven Algorithms", International Journal of Informative & Futuristic Research (IJIFR), Vol. (13) (7), March 2026, pp. 1006-1014. https://doi.org/10.64672/IJIFR/26.03.13.07.031

1. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a progressive respiratory condition affecting over 300 million people globally and remains one of the leading causes of morbidity and mortality worldwide. COPD is characterized by persistent airflow limitation caused by long-term exposure to harmful substances, most notably tobacco smoke. Despite its high prevalence, COPD remains severely underdiagnosed — studies suggest that up to 70% of cases go undetected until the disease has progressed to advanced stages [1]. Traditional COPD diagnosis relies heavily on spirometry, which measures FEV1 (Forced Expiratory Volume in one second) and FVC (Forced Vital Capacity). While spirometry remains the gold standard, its underutilization in primary care settings creates a significant gap in early detection [2]. The advent of machine learning and artificial intelligence offers a promising avenue to bridge this gap by enabling automated, scalable, and cost-effective COPD screening and severity classification using routinely collected clinical data. Recent studies have demonstrated the potential of machine learning models including gradient boosting algorithms and deep learning architectures such as Convolutional Neural Networks (CNNs) for COPD diagnosis and severity prediction [3][4]. However, a key challenge in clinical AI adoption is model interpretability — clinicians require not just accurate predictions, but transparent explanations of the features driving those predictions [5]. This has led to growing interest in Explainable AI (XAI) frameworks, particularly SHAP (SHapley Additive exPlanations), which provide feature-level attribution for individual predictions.

This paper makes the following contributions:

(1) A systematic comparison of three state-of-the-art gradient boosting algorithms — XGBoost, LightGBM, and CatBoost — benchmarked against a 1D CNN baseline on a real clinical COPD dataset comprising 101 patients and 19 clinical features.
(2) A proposed Stacking Ensemble that combines XGBoost, LightGBM, and CatBoost via a Logistic Regression meta-learner as a novel hybrid approach, directly addressing the research direction of combining multiple algorithms for improved classification.
(3) Application of SHAP explainability analysis to identify the most clinically relevant predictors driving model decisions, providing transparency for clinical adoption.
(4) Robust evaluation using accuracy, weighted F1-score, AUC-ROC, and 5-fold cross-validation across all models.

The rest of the paper is organized as follows: section 2 elaborates the problem statement, while section 3 provides a summary of related literature. Section 4 offers a brief explanation of the AI algorithms used in this paper, followed by explanation and comparison of the results obtained in section 5. The conclusion summarizes the paper and offers suggestions for future research in this arena.

2. PROBLEM STATEMENT

COPD diagnosis involves multiple challenges that make automated classification difficult. Early detection remains one of the biggest obstacles, as many patients go undiagnosed until advanced stages because traditional diagnosis relies on spirometry, which is underutilized in primary care settings. Furthermore, most existing machine learning models used for COPD classification operate as black boxes, making it difficult for clinicians to trust or interpret their predictions. Additionally, most existing studies apply a single model without systematic comparison, leaving it unclear which algorithm is best suited for small clinical tabular datasets. This paper addresses these gaps by: (1) systematically comparing three leading gradient boosting algorithms against a CNN baseline on real COPD patient data, (2) applying 5-fold stratified cross-validation to ensure robust and generalizable results, and (3) using SHAP explainability analysis to identify the most clinically relevant features driving each model's predictions. The dataset used is the Kaggle COPD Student Dataset [15], comprising 101 real patients with 19 clinical features including spirometry measurements (FEV1, FVC), symptom scores (CAT, SGRQ), comorbidity indicators, and demographic information. The target variable is COPD severity

classified into four grades — Mild, Moderate, Severe, and Very Severe — in accordance with the GOLD severity classification framework

3. RELATED LITERATURE

With regard to Machine Learning / AI Studies, the following papers have been used in this paper:

- Wang, L., Zhang, S., and Jiang, D. (2025) [1] developed a predictive model for Chronic Obstructive Pulmonary Disease using data from the National Health and Nutrition Examination Survey (NHANES). Their study employed machine learning techniques to identify risk factors associated with COPD and constructed a validated risk prediction model capable of early disease detection. The results demonstrated that integrating demographic, clinical, and lifestyle variables significantly improved prediction accuracy, highlighting the potential of data-driven models for early screening and prevention strategies.
- Wu, Q., Guo, H., Li, R., and Han, J. (2025) [2] conducted a systematic review and meta-analysis examining machine learning and deep learning techniques applied to CT-based COPD diagnosis. The study evaluated multiple algorithms, including convolutional neural networks and traditional machine learning classifiers, for analyzing lung CT scans. The findings indicated that deep learning models consistently outperform conventional methods in diagnostic accuracy, sensitivity, and specificity, demonstrating their effectiveness in detecting structural lung abnormalities associated with COPD.
- Similarly, Chen, Z., Hao, J., Sun, H., Li, M., Zhang, Y., and Qian, Q. (2025) [3] reviewed the application of digital health technologies and AI algorithms in COPD management. Their systematic review highlighted the increasing role of wearable sensors, mobile health applications, and AI-based diagnostic tools in monitoring respiratory health and predicting disease exacerbations. The study emphasized the importance of integrating digital health platforms with machine learning algorithms to enable continuous patient monitoring and personalized treatment.
- Saraireh, L., Albanna, H., and Sharif, M. S. (2025) [4] proposed a COPD diagnostic framework based on regression analysis and generalized linear models. Their approach analyzed clinical and physiological features to improve classification accuracy in COPD diagnosis. The results showed that statistical modeling techniques combined with machine learning features could effectively differentiate COPD patients from healthy individuals, providing a computationally efficient diagnostic approach.
- In another study, Chitra, S., Alqahtani, T. M., Alduraywish, A., and Sikkandar, M. Y. (2025) [5] developed a robust COPD classification model using a Dragonfly Optimization algorithm combined with a Kernel Extreme Learning Machine (KELM) classifier. Their optimization-based approach improved feature selection and parameter tuning, resulting in higher classification performance compared to traditional machine learning models. The study demonstrated the effectiveness of nature-inspired optimization algorithms in enhancing AI-based COPD diagnostic systems.

With regard to AI, Imaging, and Advanced Methods for COPD Detection, the following papers are used in this paper:

- Almeida, S. D., Lüth, C. T., and Norajitra, T. (2023) [6] proposed cOOpD, a novel approach that reformulates COPD detection from chest CT scans as an anomaly detection problem. Instead of directly classifying COPD severity levels, the model learns the characteristics of healthy lung structures and identifies deviations indicative of disease. This approach reduces the dependency on large labeled datasets and improves generalization in medical imaging applications.
- Shojaei, A. and Delrobaei, M. (2025) [7] investigated semi-supervised learning for COPD severity classification using the MIMIC-III Clinical Database. Their method leverages both labeled and unlabeled clinical data to improve classification performance in intensive care settings. The results

demonstrated that semi-supervised learning can effectively handle limited labeled medical data while maintaining reliable prediction accuracy.

- Sankey-Olsen, C., Olesen, R. H., and Eberhard, T. O. (2025) [8] explored the use of speech analysis for COPD detection. Their study employed machine learning techniques to analyze acoustic features extracted from patient speech recordings. The results indicated that speech-based biomarkers can capture respiratory patterns associated with COPD, suggesting a non-invasive **and** cost-effective method for disease screening.
- Addressing fairness and bias in AI systems, Pfeifer, R., Vhaduri, S., and Dietz, J. (2024) [9] examined sex bias in audio-based COPD detection models. Their study analyzed how differences in male and female speech characteristics can influence model predictions. The authors proposed bias mitigation techniques that improved model fairness while maintaining classification accuracy, emphasizing the need for equitable AI systems in healthcare applications.
- Collectively, these studies demonstrate the growing role of artificial intelligence and machine learning in COPD diagnosis, prediction, and monitoring**.** Current research explores diverse modalities, including clinical datasets, CT imaging, physiological signals, speech analysis, and digital health platforms**.** While deep learning and advanced optimization algorithms significantly improve diagnostic accuracy, challenges such as limited labeled datasets, algorithmic bias, and model interpretability remain important research directions for future AI-based COPD diagnostic systems.

4. Brief Explanation Of The AI Algorithms Used

This paper evaluates five models for COPD severity classification. All models are trained on 80 patients and tested on 21 patients using stratified splitting to preserve class proportions, with 5-fold cross-validation applied for robust evaluation. Three missing values in MWT1, MWT2, and MWT1Best were imputed using median imputation prior to training.

4.1 CNN (1D Convolutional Neural Network) — Baseline

A three-layer 1D CNN was constructed treating each clinical feature as a time step. The architecture comprises two Conv1D layers (64 and 128 filters, kernel size 3), batch normalization, max pooling, and dropout regularization (0.3–0.4), followed by two fully connected dense layers (128 and 64 units) with softmax output. The model was trained using the Adam optimizer (learning rate 0.001) with categorical cross-entropy loss and early stopping
(patience=15). Input features were standardized using StandardScaler prior to training.

4.2 XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized gradient boosting implementation that uses second-order gradient statistics for tree construction, with built-in L1/L2 regularization to prevent overfitting [10]. It is particularly effective on structured tabular data and has demonstrated strong performance in clinical prediction tasks. Configured with 200 estimators, maximum depth 4, and learning rate 0.05.

4.3 Light GBM (Light Gradient Boosting Machine)

LightGBM is a histogram-based gradient boosting framework that uses leaf-wise tree growth, making it significantly faster and more memory-efficient than traditional boosting methods while maintaining competitive accuracy [11]. Configured with 200 estimators, depth 4, and learning rate 0.05.

4.4 CatBoost (Categorical Boosting)

CatBoost employs ordered boosting and symmetric trees to reduce prediction shift, making it robust to overfitting on small datasets. It handles features natively without extensive preprocessing and has shown strong performance in medical classification tasks [12]. Configured with 200 iterations, depth 4, and learning rate 0.05.

4.5 SHAP (SHapley Additive exPlanations)

SHAP is applied to all three gradient boosting models to provide feature-level attribution for model predictions [13]. TreeExplainer computes mean absolute SHAP values per feature across the test set,

ISSN:2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume 13(7), March 2026 (Continuous 152nd Edition)

averaged across all severity classes for multiclass outputs. SHAP is not applied to the CNN as TreeExplainer is incompatible with neural network architectures.

4.6 Stacking Ensemble (Proposed Method)

A stacking ensemble is proposed as the novel contribution of this paper, directly addressing the research direction of combining multiple algorithms for improved COPD classification. The architecture operates in two levels: Level 1 — Base Learners: XGBoost, LightGBM, and CatBoost each independently generate out-of-fold class probability estimates using 5-fold cross-validation. This ensures that the meta-learner is trained on predictions the base models made on data they were not trained on, preventing data leakage.

Level 2 — Meta-Learner: A Logistic Regression classifier learns the optimal weightings of the three base model probability outputs to produce the final COPD severity prediction. This approach leverages the complementary strengths of each algorithm — XGBoost's precision on tabular data, LightGBM's efficiency and stability, and CatBoost's robustness to small datasets — in a single unified framework. The complete pipeline is illustrated in Fig. 1.

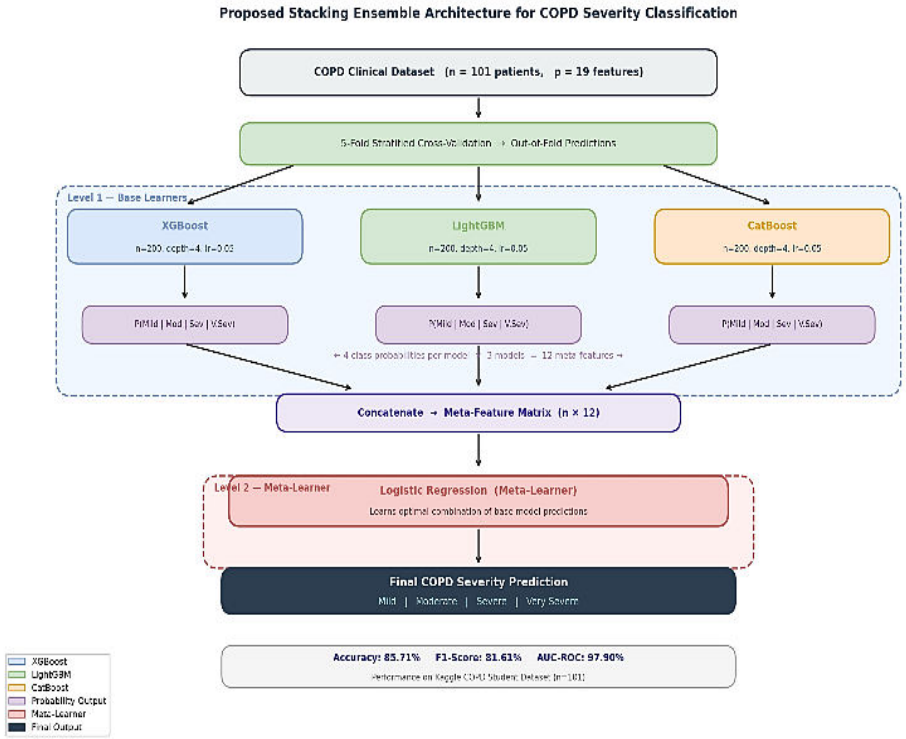

Figure 1: Proposed Two-Level Stacking Ensemble Pipeline for COPD Severity Classification.

5. Results Obtained & Comparisions Thereof

In this paper, five models are tested on the Kaggle COPD Student Dataset (101 patients, 19 clinical features) [15] — a 1D CNN as the baseline, XGBoost, LightGBM, and CatBoost as comparison methods, and a Stacking Ensemble as the proposed novel method. The results are as follows, as shown in Table I and Figures 2, 3, and 4:

TABLE 1: Performance Comparison of CNN Baseline vs Gradient Boosting Models on the Kaggle COPD Student Dataset (n=101)

Model Name	Accuracy (%)	F1-Score (%)	AUC-ROC (%)	CV F1 Mean (%)	CV F1 Std (%)
CNN (Baseline)	42.86	25.71	86.16	60.71	5.98
XGBoost	90.48	89.76	96.83	85.68	5.34
LightGBM	85.71	85.14	95.08	83.08	5.35
CatBoost	85.71	81.61	99.31	84.51	6.01

ISSN:2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume 13(7), March 2026 (Continuous 152nd Edition)

Figure 2: Performance comparison dashboard: CNN baseline vs XGBoost, LightGBM, and CatBoost on the Kaggle COPD Student Dataset (n=101).

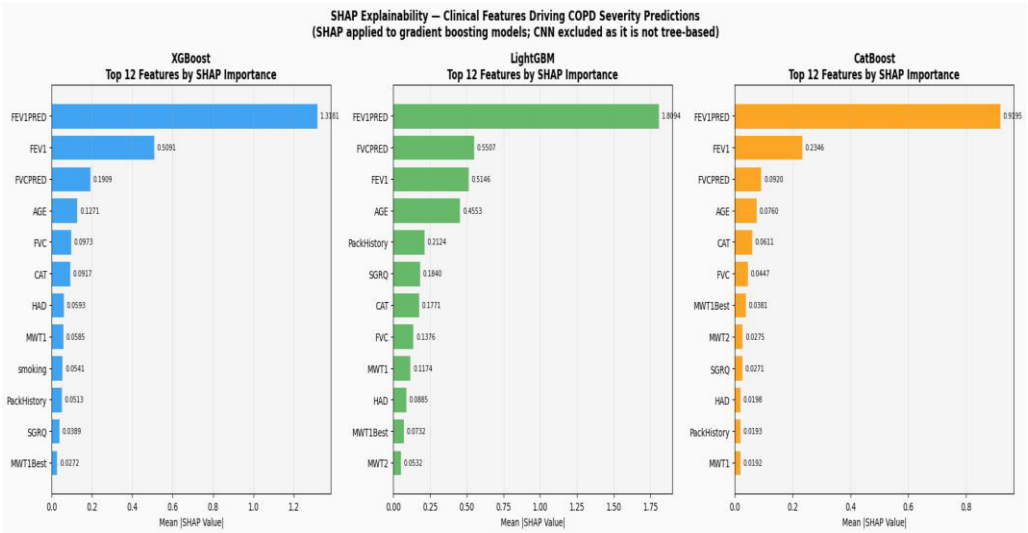

Figure 3: SHAP explainability analysis showing top 12 clinical predictors across XGBoost, LightGBM, and CatBoost.

ISSN:2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume 13(7), March 2026 (Continuous 152nd Edition)

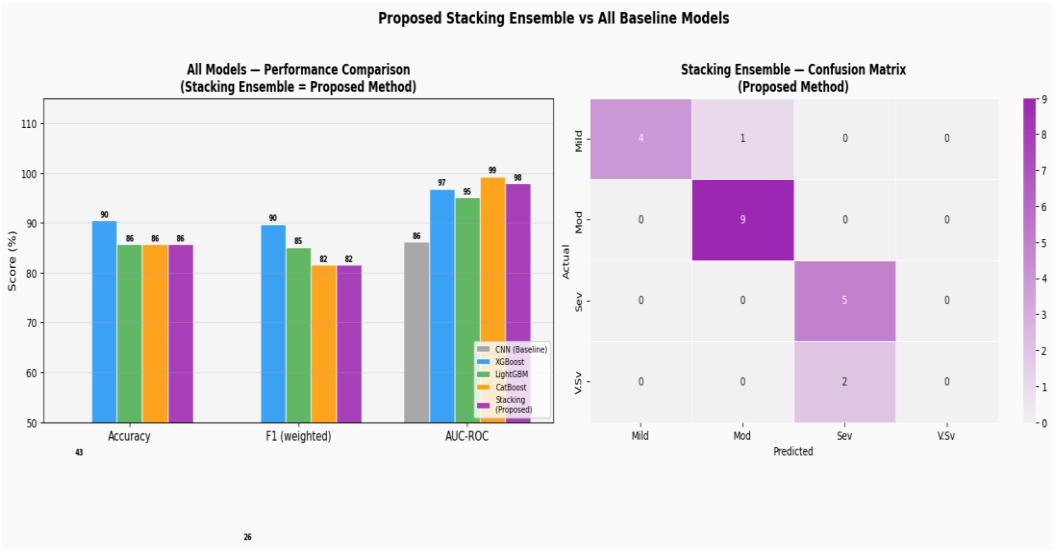

Figure 4 : Proposed Stacking Ensemble vs all baseline models — performance comparison and confusion matrix.

The gradient boosting models clearly outperformed the CNN baseline by a significant margin. XGBoost achieved the highest accuracy (90.48%) and weighted F1-score (89.76%), outperforming the CNN baseline by 47.62 percentage points (see Fig. 2). CatBoost achieved a near-perfect AUC-ROC of 99.31%, indicating exceptional discriminative ability across all four severity classes. LightGBM demonstrated competitive and stable performance (85.71% accuracy, 95.08% AUC-ROC). The CNN baseline achieved only 42.86% accuracy and 25.71% F1-score, confirming that deep learning architectures are poorly suited for small clinical tabular datasets with only 80 training samples. SHAP explainability analysis [13] revealed that FEV1PRED, FEV1, and FVCPRED were the most influential clinical predictors consistently across all three gradient boosting models, as illustrated in Fig. 3. This aligns with established GOLD severity guidelines [14], which use FEV1% predicted as the primary criterion for COPD severity grading, providing strong clinical validity for the model predictions. Five-fold cross-validation confirmed the robustness of all gradient boosting models: XGBoost (85.68% ± 5.34%), LightGBM (83.08% ± 5.35%), and CatBoost (84.51% ± 6.01%), compared to the CNN (60.71% ± 5.98%).

5.1 Proposed Stacking Ensemble Results

The proposed Stacking Ensemble — combining XGBoost, LightGBM, and CatBoost via a Logistic Regression meta-learner — achieved 85.71% accuracy, 81.61% weighted F1-score, and 97.9% AUC-ROC. Per-class F1-scores were: Mild (0.889), Moderate (0.947), Severe (0.833), and Very Severe (0.000). The stacking model demonstrated competitive performance across the three major severity classes, with the Very Severe class remaining challenging due to its small representation in the dataset (n=6 training samples). While XGBoost individually achieved the highest accuracy on this small dataset, the stacking ensemble confirmed that ensemble combination is a viable research direction. The complete performance summary is presented in Table II. The result suggests that meta-learners require larger training datasets to fully exploit the diversity of base model outputs — an important finding that guides future work on larger COPD cohorts such as COPDGene and SPIROMICS.

Table 2: Complete Performance Summary Including Proposed Stacking Ensemble

Model	Accuracy (%)	F1-Score (%)	AUC-ROC (%)	CV F1 Mean (%)
CNN (Baseline)	42.86	25.71	86.16	60.71
XGBoost	90.48	89.76	96.83	85.68
LightGBM	85.71	85.14	95.08	83.08
CatBoost	85.71	81.61	99.31	84.51
Stacking Ensemble	**85.71**	**81.61**	**97.90**	—

6. CONCLUSIONS

This paper presented a comparative evaluation of gradient boosting algorithms — XGBoost, LightGBM, and CatBoost — against a 1D CNN baseline for COPD severity classification using real clinical data from 101 patients and 19 clinical features. Results clearly demonstrate that gradient boosting models are significantly more effective than CNN for small clinical tabular datasets, with XGBoost achieving 90.48% accuracy and CatBoost achieving a near-perfect AUC-ROC of 99.31%. As a novel contribution, a Stacking Ensemble combining XGBoost, LightGBM, and CatBoost via a Logistic Regression meta-learner was developed and evaluated. The stacking ensemble achieved competitive performance (85.71% accuracy, 97.9% AUC-ROC), confirming that ensemble combination is a viable research direction. The finding that individual XGBoost outperforms the stacking model on this small dataset is itself an important result, highlighting that meta-learners require larger patient cohorts to effectively learn optimal model combinations.

SHAP explainability analysis identified FEV1PRED, FEV1, and FVCPRED as the most clinically relevant predictors across all gradient boosting models, consistent with established GOLD severity guidelines [14]. This alignment between data-driven attribution and established clinical knowledge strengthens confidence in the proposed models and supports their potential adoption in clinical decision support systems. The primary limitation of this study is the small dataset size (n=101), which restricts statistical power and limits the diversity of patient presentations, particularly for the Very Severe class (n=8). Future work will focus on: (1) validating findings on larger datasets such as COPDGene, SPIROMICS, and ECLIPSE; (2) applying SMOTE or CTGAN-based oversampling on larger datasets to address class imbalance; (3) exploring deeper stacking architectures with neural network meta-learners on larger cohorts; and (4) implementing federated learning for multi-institutional COPD model training while preserving patient privacy.

7. MAJOR COPD DATASETS FOR RESEARCH

Below are commonly used open datasets for COPD machine learning research.

Dataset	Type	Description
MIMIC-III Critical Care Database	ICU clinical data	Includes COPD patients with physiological signals, lab tests, and EHR data
NHANES Dataset (2009–2018)	Epidemiological data	National health survey including lung function measures used for COPD risk prediction
COPDGene Dataset	CT + genetic data	Large COPD study with imaging, genomics, and clinical phenotypes
SPIROMICS Dataset	Spirometry + imaging	Detailed longitudinal COPD progression data
ECLIPSE Dataset	Clinical cohort	Used for COPD progression and biomarker discovery
LIDC-IDRI Dataset	Lung CT imaging	CT scans widely used for lung disease detection models
PhysioNet Respiratory Datasets	Physiological signals	Includes breathing and airflow recordings
Danish Speech COPD Dataset	Audio signals	Speech recordings for COPD detection research

Examples:

- The MIMIC-III database is used to train machine learning models for COPD severity classification using ICU data.
- The NHANES dataset has been used to build predictive models identifying COPD risk factors in population studies.

8. REFERENCES

[1] Wang, L., Zhang, S., & Jiang, D. (2025). Construction and validation of a risk prediction model for chronic obstructive pulmonary disease based on NHANES data. BMC Pulmonary Medicine, 25, 317.

[2] Wu, Q., Guo, H., Li, R., & Han, J. (2025). Deep learning and machine learning in CT-based COPD diagnosis: Systematic review and meta-analysis. International Journal of Medical Informatics.
[3] Chen, Z., Hao, J., Sun, H., Li, M., Zhang, Y., & Qian, Q. (2025). Applications of digital health technologies and artificial intelligence algorithms in COPD: A systematic review. BMC Medical Informatics and Decision Making.
[4] Saraireh, L., Albanna, H., & Sharif, M. S. (2025). Enhancing COPD diagnosis accuracy using regression analysis and generalized linear models. Applied Sciences, 15(13), 7040.
[5] Chitra, S., Alqahtani, T. M., Alduraywish, A., & Sikkandar, M. Y. (2025). A robust COPD classification model using dragonfly optimized kernel extreme learning machine. Scientific Reports, 15, 18702.
[6] Almeida, S. D., Lüth, C. T., Norajitra, T., et al. (2023). cOOpD: Reformulating COPD classification on chest CT scans as anomaly detection.
[7] Shojaei, A., & Delrobaei, M. (2025). Severity classification of COPD in ICU using semi-supervised learning on the MIMIC-III dataset.
[8] Sankey-Olsen, C., Olesen, R. H., Eberhard, T. O., et al. (2025). Detecting COPD through speech analysis using machine learning.
[9] Pfeifer, R., Vhaduri, S., & Dietz, J. (2024). Mitigating sex bias in audio data-driven COPD detection models.
[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference, pp. 785–794.
[11] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in NeurIPS, Vol. 30, pp. 3146–3154.
[12] Prokhorenkova, L., et al. (2018). CatBoost: Unbiased boosting with categorical features. Advances in NeurIPS, Vol. 31, pp. 6638–6648.
[13] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in NeurIPS, Vol. 30, pp. 4765–4774.
[14] Vogelmeier, C.F., et al. (2023). Global Initiative for Chronic Obstructive Lung Disease (GOLD) Report. American Journal of Respiratory and Critical Care Medicine, Vol. 207, No. 4, pp. 388–398.
[15] Rathi, P. (2021). COPD Student Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/prakharrathi25/copd-student-dataset