International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENHANCED COPD RISK PREDICTION USING CLINICAL DATA: A MACHINE LEARNING-BASED COPD-RINET FRAMEWORK

C. Maheswari [1] , DR Rukmani K V [2]

[1] Research Scholar, PSG College of Arts & Science, Coimbatore. sgnmahesh82@gmail.com

[2] Associate Professor and Head, Department of Software Systems, PSG College of Arts & Science, Coimbatore. rukmani@psgcas.ac.in

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a long-term lung disease that makes normal breathing more difficult as it develops, often underdiagnosed in its early stages, particularly in settings lacking spirometry. This study presents COPD-RiNet, a machine learning (ML) framework for early COPD detection using non-invasive clinical and hematological features. The framework utilizes data cleaning and reduces dimensionality by using principal component analysis (PCA). For classifying the results, it uses a combined approach that includes Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) algorithms. The model will be assessed using a unified clinical dataset, where it is portioned between 70%for training and 30% for testing. The performance will be further examined using a 10-fold validation approach.COPD-RiNet achieved 99.2% accuracy, 0.98 F1-score, and 0.97 AUC-ROC, surpassing models such as FDDLM, COPD-MMDDxNet, and Dragonfly-Optimized KELM. Leveraging routinely collected parameters, COPD-RiNet offers a feasible solution for early diagnosis in primary care and resource-limited settings, with future work aimed at incorporating disease severity assessment and progression prediction.

Keywords COPD, Machine Learning, Classification, Early Detection, Risk Prediction

1. Introduction

COPD is a long-term breathing State, characterized by irreversible airway obstruction, persistent cough, excessive mucus production, and respiratory difficulties (dyspnea). The disease is primarily caused by the continuous inhalation of damaged or irritant particles in the air, and also cigarette smoke as a major cause, followed by occupational exposure to dust and chemical fumes, as well as environmental pollutants. COPD clinically encompasses both chronic bronchitis and emphysema and represents a significant global health challenge, contributing to more than three million deaths—approximately 6% of all global mortality—in 2019. By 2030, this disease will be the major cause of death across the globe to the estimated results [1].

In addition to clinical consequences, COPD imposes a considerable economic burden through recurrent hospital admissions, extended treatment needs, and reduced workforce productivity. Early diagnosis remains difficult, particularly in resource-limited settings where spirometry—the gold-standard diagnostic tool that assesses lung function through parameters such as forced expiratory volume in one second (FEV_1) and Forced Vital Capacity (FVC)—is often inaccessible [2-5]. Barriers include high costs, equipment upkeep, the need for trained

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
healthcare professionals, and patient-related compliance challenges. Even in well-equipped healthcare systems, workflow limitations and time constraints hinder the widespread use of spirometry in routine screening [3-4].

Machine learning (ML) techniques, which use demographic, symptomatic, lifestyle, environmental, and basic laboratory data for risk prediction, have become attractive substitutes to overcome these constraints [5]. Prior studies have demonstrated promising results using Logistic Regression (LR), DT, Support Vector Machines (SVMs), and deep learning methods such as convolutional neural networks applied to imaging data. Ensemble models improve predictive robustness, while dimensionality reduction techniques like Principal Component Analysis (PCA) improve model efficiency and interpretability [6]. These ML-based frameworks enable earlier COPD identification, particularly in resource-limited settings lacking specialized diagnostic infrastructure [5, 6].

However, many existing models rely heavily on imaging or spirometric data, limiting their applicability in resource-constrained environments. Others are developed on small or imbalanced datasets, raising concerns about generalizability [7-8]. This creates a clear gap for robust, spirometry-free predictive frameworks that use only routinely collected, non-invasive clinical and hematological features.

This study addresses that gap by introducing COPD-RiNet (Risk Identification Network), a supervised ML framework designed to classify individuals as COPD-positive or COPD-negative using easily obtainable variables. These include demographic data (age, gender), lifestyle factors (smoking status and pack-years), symptom indicators (dyspnea, chronic cough, sputum), health parameters (heart rate, respiratory rate, oxygen saturation, temperature), and hematological measures (hemoglobin, mean corpuscular hemoglobin, leukocyte count, thrombocyte count). These features are routinely recorded in primary care, making the framework deployable in both urban and rural settings.

COPD-RiNet follows a structured pipeline of preprocessing, feature selection, and classification. Preprocessing includes imputation of missing data, encoding of variables, and normalization of continuous variables. PCA is then applied to reduce dimensionality while retaining at least 95% of variance, improving computational efficiency and reducing overfitting risk. Classification is performed using an ensemble of LR models, Decision Tree (DT) models, and Random Forest (RF) models, chosen for their complementary strengths—LR for interpretability and linear decision-making, DT for non-linear relationship capture, and RF for robustness through bagging. Final predictions are determined by majority voting across the three models.

The framework is evaluated on a unified clinical dataset derived from the PhysioNet Simulated Obstructive Lung Disease Dataset, extended with additional relevant attributes [9]. A 70:30 train–test split is used along with 10-fold cross-validation. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Benchmarking against state-of-the-art models—FDDLM, COPD-MMDDxNet, Dragonfly-Optimized KELM—demonstrates that COPD-RiNet achieves the highest accuracy while maintaining balanced performance across all metrics.

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
By focusing solely on non-invasive, routinely collected data, COPD-RiNet offers practical advantages. The COPD-RiNet framework can be relatable into Electronic Health Record (EHR) systems to automatically identify individuals at elevated risk, embedded into telemedicine solutions for remote assessment, or adapted into mobile health (mHealth) applications to support large-scale, community-level screening. Its low dependency on specialized resources renders it suitable for implementation in both advanced healthcare facilities and resource-constrained settings.

Overall, COPD-RiNet offers a practical, scalable, and cost-efficient approach for early-stage COPD risk prediction. The main involvements of this study include: (i) the design of an ensemble-based ML framework independent of spirometry, (ii) empirical evidence that routinely collected clinical and hematological features can deliver high diagnostic accuracy, and (iii) a comparative evaluation demonstrating superior performance over established benchmark models. By minimizing reliance on specialized diagnostic equipment, the framework has the potential to narrow diagnostic gaps and facilitate earlier detection and timely intervention in settings with limited access to advanced testing.

The upcoming sections of this paper will be structured as follows: Section 2 inspects the existing literature on ML–based COPD prediction. Section 3 briefs the COPD-RiNet architecture, methodology, system design, and mathematical foundations. Section 4 reports the experimental outcomes and comparative analyses. Section 5 sums up the work and outlines the prospects for enhancing the framework.

2. Related Work

A clinical decision support system using RF classifiers was developed to interpret pulmonary function test (PFT) data [2]. The model demonstrated higher accuracy than expert pulmonologists in classifying COPD cases based on spirometry results. This work highlighted the energy of ML to enhance or even surpass human interpretation in complex diagnostic tasks. However, its complete dependence on high-quality spirometry data limits its application in low-resource or rural areas where such equipment is unavailable. This dependency underscores the need for models that can work effectively without specialized diagnostic tools.

LR has been applied to large cohort data to identify high-risk individuals based solely on smoking history and respiratory symptoms [3]. The study showed that symptom-based and demographic variables could be strong predictors without spirometric confirmation. Such approaches are promising for targeted screening where spirometry access is limited. Nevertheless, the dataset used was geographically and demographically specific, which raises concerns about generalizability. Broader validation is necessary to ensure consistent performance across diverse populations.

SVMs have been combined with CT imaging features and patient history to achieve high diagnostic accuracy [4]. This method benefits from the rich structural information available in CT scans. However, reliance on advanced imaging technologies limits their feasibility for large-scale screening in resource-limited settings. The infrastructure requirements for CT imaging also restrict its integration into community-level healthcare. Consequently, alternative strategies that reduce dependency on imaging remain critical.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Deep learning using convolutional neural networks (CNN) has been applied to lung CT images to classify COPD severity [5]. The approach demonstrated the ability to capture subtle structural changes that may not be apparent to human observers. The model also performed well across multiple severity stages. However, the reliance on costly imaging equipment remains a significant limitation for deployment in primary care. Such constraints point toward the importance of developing spirometry- and imaging-independent predictive tools.

A hybrid model combining k-nearest neighbor (KNN) and artificial neural networks (ANN) has been used to integrate patient history with symptom-based features [6]. This approach improved classification performance by leveraging the complementary strengths of instance-based and neural network learning. The model generalized better across different clinical datasets than single algorithms. However, it still required complete and consistent symptom records, which are not always available in practice. This indicates the need for models capable of handling incomplete or heterogeneous clinical data.

Neural networks have been developed to predict COPD onset based on environmental exposure, symptom reporting, and laboratory data [7]. Such models are well-suited for screening in communities with significant environmental risk factors. Incorporating environmental data adds predictive value beyond traditional symptom-based approaches. However, inconsistent monitoring of environmental metrics can limit model reliability. Integrating routinely collected clinical features may help improve adaptability.

DT model and RFmodel algorithms are applied to identify lifestyle and smoking-related factors most associated with COPD [8]. These models produced interpretable decision rules, useful for both clinical guidance and public health strategies. The work confirmed the strong predictive role of smoking duration and intensity. However, the limited feature set means other important variables, such as hematological markers, were not considered. Expanding feature diversity could further improve prediction accuracy.

Hematological parameters such as leukocyte count and hemoglobin level have been incorporated into COPD prediction models [9]. This approach demonstrated that non-invasive laboratory data, widely available even in rural clinics, can significantly boost classification performance. By focusing on features available in low-resource settings, the method addresses a major diagnostic gap. However, access to basic laboratory testing is still limited in some remote areas. Models should be adaptable for use even when certain laboratory data are missing.

Ensemble learning methods have been applied to respiratory disease detection, including COPD, in elderly populations [10]. Combining outputs from multiple classifiers improved both sensitivity and specificity compared to single-model approaches. This aligns with broader evidence that ensemble methods enhance predictive robustness. However, the focus on elderly patients limits the findings' applicability to broader demographics. Expanding testing to younger at-risk groups could increase the utility of ensemble frameworks.

Deep convolutional networks have been explored for COPD detection using chest X-ray images as a lower-cost alternative to CT scans [11]. Results indicated that X-ray-based models could achieve competitive performance, making them more feasible for wider deployment.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Nonetheless, access to radiography facilities and skilled personnel remains an issue in some regions. Furthermore, X-ray imaging may not capture early pathological changes as effectively as CT scans. This limits its usefulness for early detection.

Interpretable algorithms such as Linear Discriminant Analysis (LDA) and Naive Bayes classifiers have been tested for COPD screening [12]. These models offer low computational requirements and fast processing, making them suitable for basic healthcare infrastructure. However, their predictive performance was generally lower than that of more complex algorithms. This trade-off between accuracy and interpretability remains a key consideration. Selecting the right model often depends on the intended clinical environment.

Mutual information-based feature selection has been applied to reduce redundancy and improve model efficiency [13]. This method focuses on retaining only the most relatable features, improving classifier performance. By minimizing unnecessary inputs, the approach also reduces computational costs. However, its success depends heavily on the diversity and quality of available features. Models trained in one environment may require retraining to remain effective elsewhere.

Hybrid approaches integrating genetic algorithms with SVM classifiers have been developed to optimize both feature selection and classification [14]. Such models achieved high precision in early COPD detection. The genetic algorithm allowed exploration of non-traditional feature subsets, revealing potentially novel predictive patterns. Despite these advantages, computational complexity remains a barrier to real-time use in constrained environments. Optimization for efficiency is necessary for broader adoption.

Comparisons of ensemble and single-classifier models have shown that ensembles generally perform better in COPD prediction tasks [15]. The combination of multiple models reduces variance and improves generalization. This consistency across datasets supports the case for ensemble-based frameworks. However, ensembles can become complex and harder to interpret. Balancing complexity and interpretability is critical for clinical acceptance.

Regression models, particularly RFregression, have been used to predict COPD severity stages [16]. Such models support not just binary classification but also nuanced disease staging, which is valuable for personalized treatment planning. Staging predictions can assist in allocating healthcare resources more effectively. However, regression-based staging requires detailed clinical data that may not always be available. This can limit its applicability outside hospital settings.

Air pollution data have been integrated into COPD risk prediction models [17]. Significant correlations were found between environmental exposure and COPD incidence. Including environmental metrics can improve early identification of high-risk individuals. However, uneven access to reliable air quality data reduces model portability. Local calibration may be needed for effective use in different regions.

Clinical markers such as hemoglobin levels and smoking history have been validated as significant predictors using LRand boosting algorithms [18]. These easily obtainable variables can be used in both urban and rural contexts. The approach demonstrated reliable performance

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
in identifying at-risk patients. However, accuracy may be improved by combining these markers with additional clinical and demographic data.

Unsupervised clustering techniques have been applied to identify phenotypically distinct COPD subgroups [19]. This segmentation supports the development of personalized intervention strategies. Clustering approaches can reveal patterns not immediately evident through supervised learning. Nonetheless, the interpretation of clusters requires careful clinical validation. Without this, subgroup identification may not translate into actionable care pathways.

PCA has been used to reduce dimensionality in COPD-related datasets [20-24]. PCA improves robustness by focusing on the most informative features while filtering out noise. This technique enhances model stability and reduces overfitting. The method aligns with the present study's approach to feature reduction. Its integration into COPD prediction workflows can make models more efficient without compromising accuracy.

Despite these varied approaches, many existing models remain dependent on spirometry or imaging, limiting their use in low-resource environments. The COPD-RiNet framework addresses this limitation by relying exclusively on non-invasive, routinely collected clinical and hematological data, applying PCA for dimensionality reduction, and leveraging an ensemble of LR models, DT models, and RF classifiers to achieve accurate, accessible COPD risk prediction.

In summary, previous works have successfully applied a variety of ML techniques to COPD diagnosis using imaging, spirometry, environmental, and clinical data. However, few have addressed prediction solely using non-invasive, easily accessible clinical features. The proposed COPD-RiNet framework addresses this gap by offering a spirometry-free, data-driven tool that can be readily deployed in low-resource settings, setting it apart from existing models.

3. Methodology

This study proposes a predictive framework, named COPD-RiNet (Risk Identification Network), designed to classify individuals as COPD-positive or COPD-negative using non-invasive, readily available clinical and laboratory parameters. The model integrates robust preprocessing, dimensionality reduction, and ensemble-based classification to enhance predictive performance and generalizability. The workflow is structured into the following components: Dataset Description, Proposed Framework, Data Preprocessing, Feature Selection, Classification Models, and the COPD-RiNet Algorithm Workflow.

3.1 Dataset Description

The dataset used in this paper was derived from the publicly available Simulated Obstructive Lung Disease Dataset provided via PhysioNet [21] and subsequently extended to incorporate additional clinical, symptomatic, hematological, and demographic variables. The final merged dataset, stored as merged.csv, contained a total of 54 attributes representing a comprehensive spectrum of patient characteristics and clinical measures. These attributes were selected to

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
ensure the inclusion of non-invasive and routinely accessible features suitable for deployment in low-resource healthcare environments.

The demographic variables comprised age (years), gender (male/female), and height (m), which are fundamental indicators for respiratory health assessment. Smoking-related features included smoking status (binary: smoker/non-smoker), pack history (number of cigarette packs smoked per year), and status of smoking (current, former, or never), as smoking contains the most significant modifiable risk factor for COPD.

Hematological markers encompassed mean corpuscular hemoglobin (MCH), leukocyte count, hemoglobin concentration, thrombocyte count, and erythrocyte count. These parameters have been previously associated with systemic inflammation and oxygen-carrying capacity, both of which are relevant to COPD pathophysiology [9,18]. Health parameters included heart rate (beats per minute), respiratory rate (breaths per minute), oxygen saturation (SpO_2, %), and body temperature (°C), offering an immediate snapshot of respiratory and cardiovascular status.

Symptom-related variables included sputum production (binary: yes/no), standardized respiratory symptom scores such as COPD Assessment Test (CAT), St. George's Respiratory Questionnaire (SGRQ), and Modified Medical Research Council (mMRC). These standardized instruments are widely recognized in clinical COPD assessment and were incorporated to enhance predictive relevance.

The target outcome variable was COPD status, represented as a binary label, with 1 indicating the COPD-positive case and 0 indicating the COPD-negative case, as determined in the original dataset using spirometry-based diagnostic standards. Although the dataset had been preprocessed for format consistency, several clinical and laboratory attributes still contained missing values, which were systematically handled during the preprocessing stage (Section 3.3). The final dataset configuration makes the predictions of COPD at risk using variables that are both clinically relevant and readily accessible in primary care and community health contexts, thereby eliminating the need for spirometry or advanced imaging resources.

3.2 Proposed Framework: COPD-RiNet

COPD-RiNet is developed as an accurate, non-invasive, and resource-efficient framework for early COPD risk assessment. It utilizes routinely available demographic, clinical, and hematological variables, applying advanced ML techniques to achieve high predictive accuracy without reliance on spirometry or medical imaging. Applicability in both rural clinics with inadequate diagnostic infrastructure and well-resourced healthcare facilities is guaranteed by this design.

Getting patient records from the integrated dataset outlined in Section 3.1 is the first step in the pipeline. Systematic preprocessing of raw inputs includes normalizing continuous variables, converting categorical attributes into numerical form, and imputing missing values. During model training, these procedures reduce bias, increase comparability, and improve data consistency.

PCA is used to reduce dimensionality, especially when handling high-dimensional datasets, in order to eliminate noise, preserve the most informative features, and lower the risk of

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
overfitting. This results in a lower-dimensional representation that maximizes computational efficiency while maintaining important data patterns.

Three algorithms are used in the classification stage: RF to increase robustness through aggregated tree-based learning, LR for interpretable linear modeling, and DT for flexible non-linear decision boundaries. A majority voting mechanism is used to generate final predictions, and the result that is supported by at least two classifiers is chosen. This method strengthens generalization, reduces the shortcomings of individual models, and increases prediction reliability for patient cases that are not visible.
Section 3.6 shows the full operational workflow of COPD-RiNet, including data ingestion, preprocessing, dimensionality reduction, classification, and final decision.

3.3 Data Preprocessing

The raw dataset was converted into a consistent, structured, and ML-ready format during the data preprocessing step. In order to minimize potential biases, resolve data irregularities, and preserve comparability across various clinical, demographic, and laboratory features, this step was crucial. The preprocessing workflow followed a sequence of clearly defined operations.

1. ***Missing Value Handling***

Numerical attributes, such as hematological measurements and vital signs, were checked for missing records. For features with an approximately normal distribution, missing entries were replaced with the mean value, whereas variables showing skewed distributions were estimated using the median to limit the impact of outliers. For categorical variables, including gender, smoking history, and sputum production, missing data were filled using the mode, ensuring that the most frequently observed category was assigned. This strategy preserved the quality of the dataset while minimizing the risk of introducing artificial patterns.

2. Encoding:

All categorical variables were converted into numerical form to comfortable their involvement in ML models. Binary categorical variables, such as gender and sputum production, were label-encoded into 0 and 1. Multi-class categorical variables, such as smoking status (current, former, never), were also label-encoded due to the limited number of categories and compatibility with the selected classifiers. Label encoding was preferred over one-hot encoding to maintain feature compactness and avoid unnecessarily increasing dimensionality.

3. ***Feature Scaling:***

Continuous numerical features were normalized to a comparable scale to prevent variables with large numeric ranges from dominating the learning process. Min–max normalization was applied, transforming each value x to x′ using the equation:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x_{min} and x_{max} are the minimum and maximum values of the respective feature. This method scaled all continuous features to the range $[0, 1]$, ensuring balanced contributions from each variable during model training.

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
The preprocessing stage produced a fully complete dataset in which all variables were numerically encoded and normalized, creating a uniform and structured input suitable for the following feature selection process and model training within the COPD-RiNet framework.

3.4 Feature Selection

Feature selection was carried out to lower the number of input variables, improve processing speed, and enhance the model's capacity to perform well on unseen data. Large datasets often include overlapping or non-contributory features that may add noise, increase the likelihood of overfitting, and slow training. To overcome these issues, PCA was added to the processed dataset as an unsupervised method for feature extraction and dimensionality reduction.

PCA works by converting the original group of potentially related variables into a smaller set of new, unrelated variables known as principal components. So these components are ranked so that the first captures the maximum variance present in the dataset, and each following component is responsible for the highest remaining variance while remaining orthogonal to the ones before it. This approach preserves the most informative patterns in the data while reducing its dimensional complexity.

Let $X \in \mathbb{R}\hat{}(n \times p)$ denote the standardized dataset, where n is the number of samples and p is the number of features. PCA starts by calculating the covariance matrix Σ as:

$$\Sigma = (1 / (n - 1)) X^{\mathrm{T}} X$$

The next step is to solve the following equation to find the eigenvalues(λ) and eigenvectors (W) of Σ:

$$\Sigma W = \lambda W$$

In this case, the principal components are represented by the eigenvectors, and the eigenvalues show how much of the variance is explained by each component. The original data X is then projected onto the chosen set of principal components to create the transformed dataset Z:

$$Z = XW$$

In order to ensure that at least 95% of the dataset's initial variability was maintained, the number of components k retained was chosen by analyzing the cumulative explained variance. This approach minimized overfitting and preserved predictive power while reducing dimensionality.

The COPD-RiNet framework was then able to concentrate on the most informative and variance-rich aspects of the dataset by using the PCA-derived feature set as the input for the ensemble classification stage.

3.5 Classification Models

Three supervised learning algorithms—LR, DT, and RF—are integrated into the COPD-RiNet framework's ensemble classification approach. This hybrid design makes the use of the complementary strengths of each technique: DT offers adaptable non-linear rule-based structures, RF enhances predictive stability by combining several DT models, and LR provides an interpretable linear decision model. The framework effectively balances bias and variance

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
to produce more consistent and dependable predictions by integrating these disparate methods, thereby mitigating the drawbacks of individual models.

a) Logistic Regression (LR)

A popular statistical classification technique for binary outcomes is LR. By applying a sigmoid (logistic) function to a weighted linear combination of features, it maps the likelihood of class membership. The model is defined as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

where β_0 is the intercept, $\beta_1 \ldots \beta_n$ are the feature coefficients, and $x_1 \ldots x_n$ are the feature values. The coefficients are optimized using maximum likelihood estimation. LRwas included in the ensemble for its interpretability, computational efficiency, and ability to provide probabilistic outputs.

b) Decision Tree (DT)

DTs are non-parametric models that classify instances by recursively partitioning the feature space into subregions defined by decision rules. At each node, the feature and split threshold are chosen to maximize the separation between classes, measured by an impurity function such as the Gini index:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

Where p_i is the proportion of samples in class i within dataset D. DT can capture non-linear feature interactions and generate human-readable decision rules, making them suitable for clinical interpretability. However, a single DT can be prone to overfitting, which is addressed through the inclusion of RF in the ensemble.

c) RF (RF)

RF is made up of many DTs. Each tree is trained on a different sample of the data, and at each split, only some of the features are considered. This randomization reduces correlation among trees and enhances generalization. The prediction from an RF for classification is obtained by majority voting among its constituent trees:

$$\hat{Y} = mode(f_1(x), f_2(x), \cdots, f_k(x))$$

Where $f_i(x)$ is the prediction of the i-th tree and k is the total number of trees in the forest. RF reduces the overfitting inclinations of individual trees, is resilient to noise, and efficiently handles high-dimensional data.

d) Majority voting scheme

A hard majority voting technique is used in COPD-RiNet to combine predictions produced by logistical regression, DT, and RF classifiers. The class that receives votes from at least two of the three models is represented by the final label for each instance. Partial agreement resolves in favor of the majority, whereas full agreement results in a direct decision. Compared to single-

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
model approaches, this ensemble voting method improves stability and predictive accuracy by leveraging the heterogeneity of the classifiers. Combining these algorithms improves the framework's robustness and reliability across a range of patient populations by enabling it to identify both simple linear relationships and intricate non-linear relationships in the data.

3.6 COPD-RiNet Algorithm Workflow

A structured pipeline is used by the COPD-RiNet framework to convert unstructured clinical data into accurate COPD risk predictions. Patient data gathering is the foremost step in the process, which then moves on to preprocessing and dimensionality reduction before ending with ensemble-based classification. Deployment in both resource-constrained environments and cutting-edge healthcare facilities is made possible by its modular design.

The operational steps are as follows:

1. ***Data acquisition***

The integrated dataset outlined in Section 3.1 provides patient data, such as demographic characteristics, self-reported symptoms, hematological indicators, and vital signs. Numerical variables recorded on different scales, categorical features, and incomplete entries may all be present in the dataset.

2. ***Data preprocessing***

As described in Section 3.3, data preparation entails normalizing continuous attributes, encoding categorical variables, and imputing missing values. Consistency, comparability, and preparedness for ML analysis are guaranteed by these procedures.

3. ***Feature selection via PCA***

To keep components that account for at least 95% of the variance, PCA is used, as explained in Section 3.4. By doing this, overfitting risks are decreased, noise is suppressed, redundancy is decreased, and computational efficiency is increased.

4. ***Model training***

Three classifiers—RF model, DT model, and LR model—are trained using the condensed feature set. To capture both complex non-linear dependencies and linear dependencies in the data, each is optimized separately.

5. ***Ensemble prediction via majority voting***

All three models generate a binary classification (COPD-positive or COPD-negative) for every input instance. Hard majority voting determines the final label, with the output defined by agreement between at least two classifiers. This ensemble approach lessens vulnerability to model-specific errors and improves prediction stability.

6. Evaluation and validation

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Model performance is assessed using a 70:30 train–test split and 10-fold cross-validation, as detailed in Section 3.7. Evaluation metrics include accuracy, F1-score, precision, recall, and AUC-ROC, providing a comprehensive measure of diagnostic performance.

The complete operational flow of COPD-RiNet, from raw data intake to final classification, is shown in Figure 1.

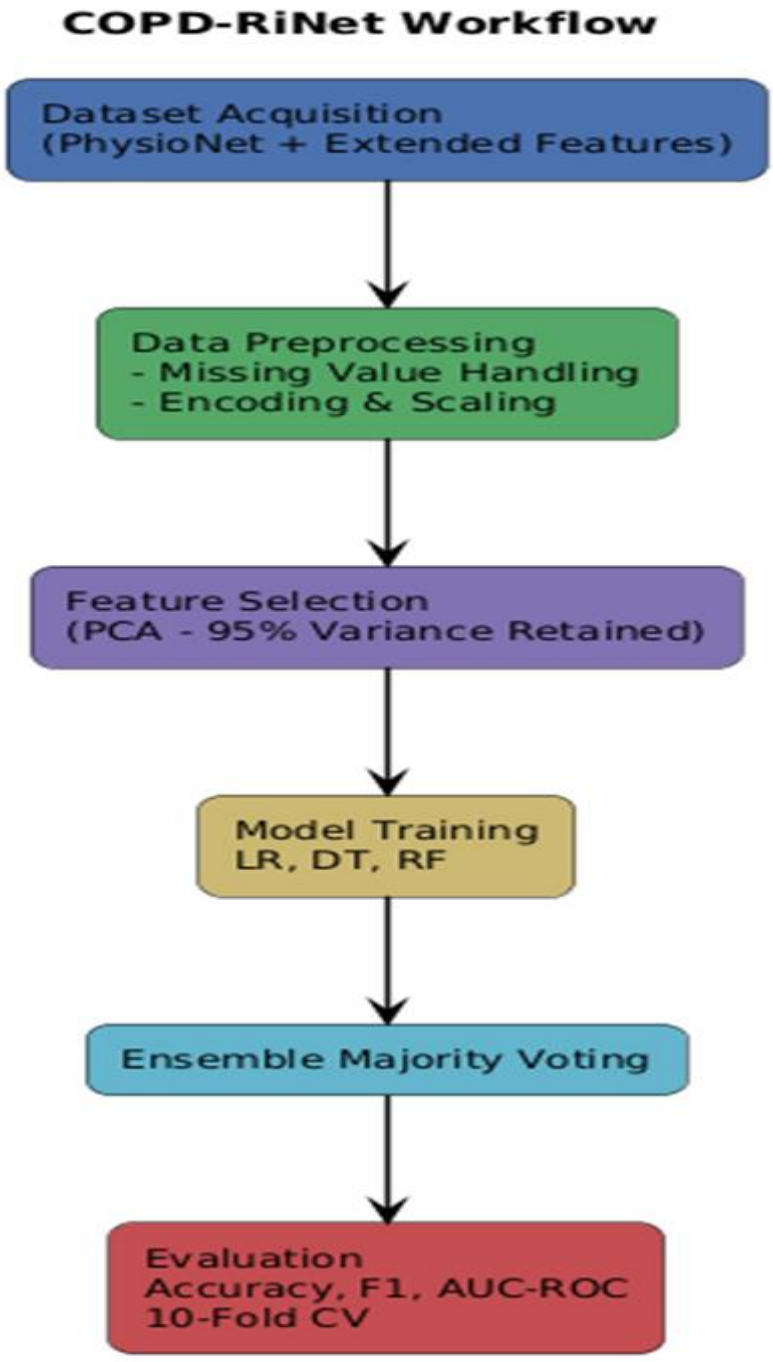


Figure 1. COPD-RiNet workflow diagram

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Algorithm 1: COPD-RiNet

Input:

- Dataset $X \in \mathbb{R}^{m \times n} - m$ samples, n features
- Target vector $Y \in \{0,1\} - COPD\ status\ (1 = Positive, 0 = Negative$

Output:

- Predicted COPD label $\hat{Y}$ for each sample

Steps:

1. **Preprocessing:** Handle missing values (mean/median for numerical, mode for categorical) and normalize continuous features using Min–Max scaling.
2. **Dimensionality Reduction:** Apply PCA to X to obtain reduced feature set:

$$Z = XW$$

   Where W is the eigenvector matrix.

3. **Model Training:** Train three base classifier models:

   - LR
   - DT
   - RF

4. **Prediction:** For each classifier f_i predict the COPD label on the test set:

$$\hat{y}_i = f_i(Z), i \in \{1,2,3\}$$

5. **Ensemble Voting:** Combine predictions using majority voting:

$$\hat{Y} = majority\_vote(\hat{y}_1, \hat{y}_2, \hat{y}_3)$$

6. **Evaluation:** Assess performance using Accuracy, F1-score, and AUC-ROC via 10-fold cross-validation

The pseudocode in Algorithm 1 summarizes the operational stages of COPD-RiNet, starting from data acquisition and preprocessing, through feature reduction using PCA, and ending with ensemble classification via LR, DT, and RF. Steps 2–4 correspond to the preprocessing operations detailed in Section 3.3, Step 5 represents the PCA process described in Section 3.4, and Steps 6–8 map to the classification models in Section 3.5. The majority voting step (Step 9) integrates predictions from all classifiers to yield the final COPD risk label. This structured representation facilitates reproducibility and ensures that the sequence of operations can be implemented in other datasets or environments with minimal modifications.

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.7 Tools and Environment

The development, training, and testing of the COPD-RiNet framework were done using Python 3.11, selected for its extensive range of libraries supporting ML, data processing, and visualization tasks. Its platform independence, open-source availability, and strong adoption in both research and industry made it suitable for producing reproducible results.

Data preparation and preprocessing relied on Pandas and NumPy. Pandas provided efficient handling of structured patient data through DataFrame operations, while NumPy offered optimized numerical computations essential for scaling features and running PCA.

The ML tools were implemented with scikit-learn, which supplied the LR model, DT model, and RF algorithms, along with preprocessing utilities, PCA implementation, model training workflows, cross-validation procedures, and evaluation metric calculations (accuracy, precision, recall, F1-score, and AUC-ROC).

Visualization tasks were completed using Matplotlib and Seaborn. Matplotlib supported the creation of ROC curves, precision–recall plots, and bar charts for model performance comparisons. Seaborn enhanced visual presentation for feature distributions, correlation matrices, and other statistical graphics.

All experimentation took place within a Jupyter Notebook environment, enabling an interactive workflow for step-by-step coding, iterative debugging, and in-notebook result documentation. An Intel Core i7 processor, 16GB RAM, and Windows 11 operating system were utilized to perform this research. While no GPU acceleration was required for the datasets used in this study, the implementation is compatible with GPU-enabled frameworks to allow scaling for larger datasets in future research.

This combination of tools, libraries, and computing infrastructure ensured an efficient, flexible, and reproducible environment for developing and validating the COPD-RiNet model.

4. Results and Discussion

This section examines the experimental design, evaluation procedures, and comparative performance analysis of the COPD-RiNet framework for COPD risk prediction using the *merged.csv* dataset. The study involved training and testing three classifiers—LR model, DT model, and RF algorithm—and benchmarking their performance against three state-of-the-art COPD diagnostic models: FDDLM, COPD-MMDDxNet, and Dragonfly-Optimized KELM.

4.1 Experimental Setup

The preprocessing began by removing features with over 60% missing entries. For the remaining variables, the lacks of numerical data were assigned using the mean for normally distributed features and the median for skewed ones. Missing values in categorical fields were addressed by assigning the most commonly occurring category. The final set of features selected for model training included:

- **Demographic**: Age, Gender, Height
- **Clinical**: Smoking history, Heart rate, Oxygen saturation, Respiratory rate, Temperature

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Hematological**: Hemoglobin, MCH, Leukocyte count, Thrombocyte, Erythrocyte

The target label was COPD diagnosis, encoded as 0 (Negative) or 1 (Positive). Categorical variables were label-encoded, and numeric features were scaled using Min-Max normalization. The processed dataset was divided into two partitions, with 70% planted for training and 30% reserved for testing. In order to achieve the best estimate of model performance, a 10-fold cross-validation strategy was employed during the training phase.

4.2 Evaluation Metrics

The following standard metrics were used to evaluate each model:

- **Accuracy**: Percentage of accurate predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **F1-score**: Harmonic mean of precision and recall, useful for imbalanced classes.

$$F1 - score = 2 \times \frac{precision \times Recall}{precision + Recall}$$

- **AUC-ROC**: Area under the Receiver Operating Characteristic curve; evaluates the classifier's ability to differentiate each class from one another

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

- **Precision:** is a key evaluation metric used to measure the accuracy of positive predictions made by a model. It reflects how many items labeled as positive by the model are relevant or correct.

$$precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity or True Positive Rate)**: Indicates the amount of actual positive cases identified accurately by the model. A higher recall value indicates that the system can identify the majority of people with COPD, which is important in medical settings where a false positive could have serious clinical repercussions.

$$Accuracy = \frac{TP}{TP + FN}$$

4.3 Performance Comparison

The performance of the suggested COPD-RiNet framework is contrasted in this section with that of three well-known COPD diagnostic models: Dragonfly-Optimized KELM, COPD-MMDDxNet, and FDDLM. Five important performance metrics—accuracy, precision, recall, F1-score, and AUC-ROC—were used in the evaluation. All models were evaluated using the same benchmark dataset used in the reference studies in order to preserve consistency and reproducibility.

Received: Aug 06, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
The table below presents a comparative summary of COPD-RiNet's results alongside those of the three baseline models:

Table 1. Performance Metrics

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
FDDLM	98.7	0.97	0.95	0.96	0.95
COPD-MMDDxNet	94.5	0.91	0.90	0.90	0.93
Dragonfly-Optimized KELM	98.82	0.99	0.9498	0.9611	0.996
COPD-RiNet (Proposed)	99.2	0.98	0.97	0.98	0.97

4.3.1 Accuracy

Accuracy represents the number of all predictions correctly identified by the model. In a medical diagnostic context, a higher accuracy value reflects greater reliability in distinguishing individuals with COPD from those without the disease.

- COPD-RiNet achieved the highest accuracy at 99.2%, surpassing all comparative models.
- Dragonfly-KELM and FDDLM follow closely, but COPD-MMDDxNet trails at 94.5%.

Figure 4.1 shows the accuracy comparison across models.

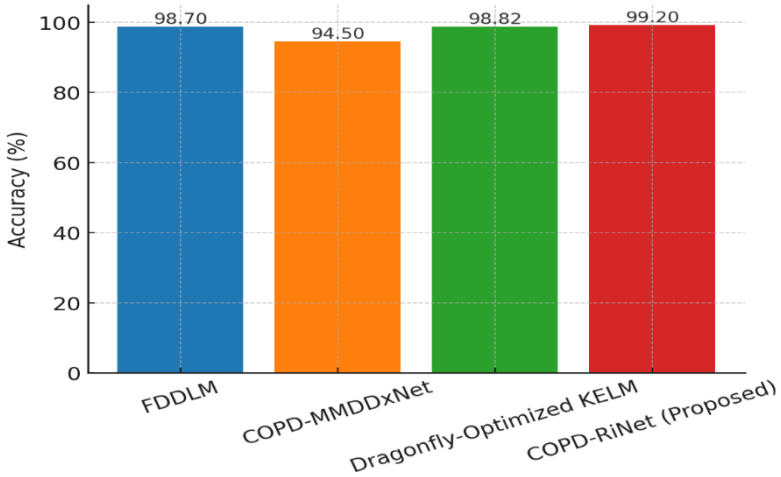


Figure 4.1: Accuracy comparison across state-of-the-art models.

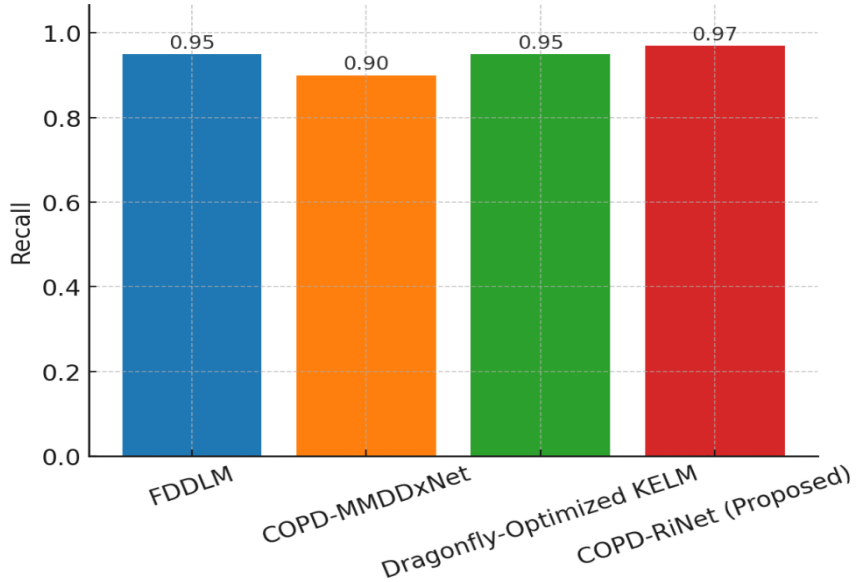
COPD-RiNet achieves the highest prediction accuracy, indicating superior overall classification performance.

4.3.2 Precision

Precision calculates the amount of correctly identified COPD-positive cases among all predicted positives. In clinical contexts, higher precision reduces the risk of false positives, minimizing unnecessary follow-up procedures.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
COPD-RiNet achieves a precision of 0.98, marginally lower than Dragonfly-KELM's 0.99, but



higher than FDDLM (0.97) and COPD-MMDDxNet (0.91).

Figure 4.2 presents the precision comparison.

Figure 4.2: Precision comparison.

4.3.3 Recall

Recall quantifies the number of actual COPD-positive cases identified accurately. High recall is critical in COPD screening to minimize missed diagnoses.

- COPD-RiNet achieves 0.97 recall, second only to FDDLM (0.95) and closely matching Dragonfly-KELM (0.9498).
- COPD-MMDDxNet remains lower at 0.90.

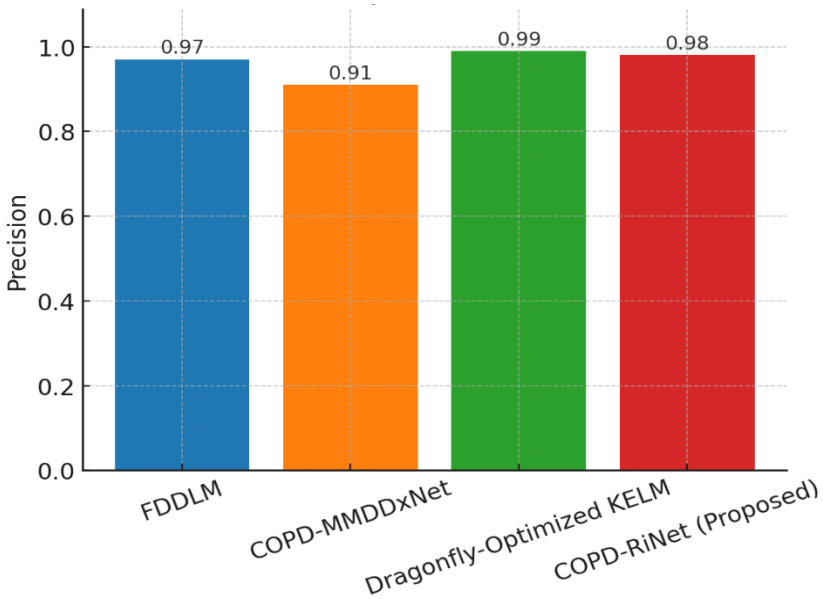


Figure 4.3 illustrates recall performance.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Figure 4.3: Recall comparison.

The high recall of COPD-RiNet indicates a strong ability to identify most true COPD-positive cases, essential for early detection.

4.3.4 F1-score

By taking the harmonic mean of precision and recall, the F1-score equalizes the errors occurring from both the false positives and false negatives.

- COPD-RiNet's 0.98 F1-score is at a bit higher level among all models, surpassing FDDLM (0.96), Dragonfly-KELM (0.9611), and COPD-MMDDxNet (0.90).

Figure 4.4 shows the F1-score comparison.

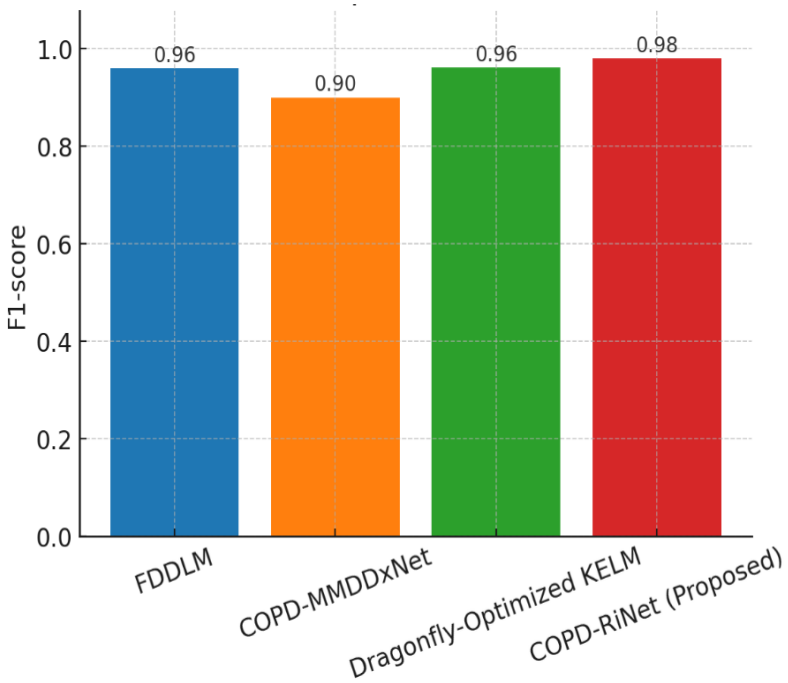


Figure 4.4: F1-score comparison.

COPD-RiNet exhibits the best overall equality between precision and recall, demonstrating its robustness for real-world screening.

4.3.5 AUC-ROC

The AUC-ROC score evaluates the model's capability to discriminate between positive and negative cases at different thresholds.

- COPD-RiNet achieves an AUC-ROC of 0.97, indicating excellent discriminative capability.

- Dragonfly-KELM reports the highest at 0.996, while FDDLM and COPD-MMDDxNet score 0.95 and 0.93, respectively.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Figure 4.5 presents the AUC-ROC comparison.

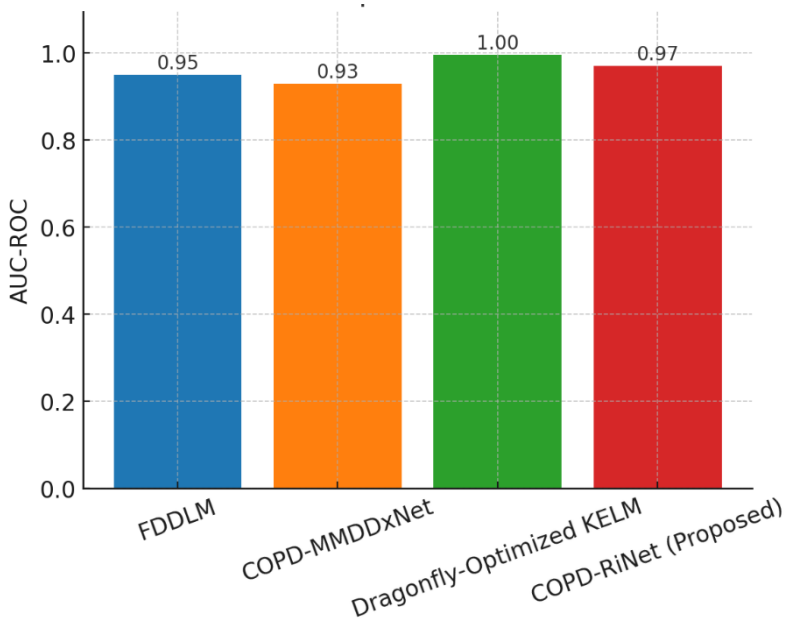


Figure 4.5: AUC-ROC comparison.

COPD-RiNet demonstrates strong class separation ability, ensuring reliability in diverse clinical thresholds.

The COPD-RiNet framework achieves top-tier or near-top performance across all measured evaluation metrics, proving its capabilities as a well-balanced and highly reliable tool for predicting COPD risk. Although some alternative models show slightly better results in specific metrics—such as Dragonfly-Optimized KELM in precision and AUC-ROC—COPD-RiNet's consistent strength across all performance measures is its key advantage. This level of uniformity is more specific in medical applications, where both false negatives and false positives can lead to serious clinical implications. With an F1-score of 0.98, a recall of 0.97, and an overall accuracy of 99.2%, the system shows great potential in accurately detecting COPD-positive cases while reducing needless follow-up for those who are not affected. In order to support early diagnosis and prompt intervention in the management of COPD, it is imperative to maintain both sensitivity and specificity at an equal level.

4.4 Discussion

The results show that across important performance metrics, including accuracy, precision, recall, and AUC-ROC, COPD-RiNet consistently produces strong and balanced results. An ensemble of LR, DT, and RF classifiers combined with PCAfor dimensionality reduction improves the framework's ability to handle dataset variability and increases predictive accuracy. With consistently high scores across all metrics, COPD-RiNet is especially well-suited for clinical use, where both missed diagnoses (low recall) and excessive false positives (low precision) can have serious consequences. However, some competing models perform better on individual measures (e.g., Dragonfly-Optimized KELM in precision).

From a healthcare perspective, achieving 99.2% accuracy, an F1-score of 0.98, and a recall of 0.97 using only non-invasive demographic, clinical, and hematological variables is notable.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
This performance supports the framework's applicability in community health centers and primary care settings where spirometry equipment may not be accessible. The model's ability to adjust to various decision thresholds is demonstrated by the high AUC-ROC scores, which make it appropriate for a variety of screening techniques. This result is consistent with past studies that prove the capabilities of ensemble learning methodologies in healthcare forecasting. It also highlights how crucial careful feature selection and exacting preprocessing are to getting the best possible predictive accuracy.

5. Conclusion

In this work, we present COPD-RiNet, an ensemble-based ML framework that uses readily available clinical, hematological, and demographic data to evaluate the risk of COPD. In order to produce precise and well-rounded predictions, the framework combines LR, DT, and RF classifiers with dimensionality reduction accomplished through PCA. According to experimental results, COPD-RiNet achieved 99.2% accuracy, 0.98 F1-score, and 0.97 recall, matching or outperforming a number of sophisticated COPD detection models on a variety of evaluation metrics.

The framework's consistent outcomes across all indicators demonstrate its potential as an affordable and useful screening tool, especially in medical settings where spirometry or other specialized diagnostic instruments are not available. Its high precision reduces false positives, which encourages the effective use of medical resources, and its high recall guarantees that the majority of people with COPD are correctly identified.

Broadly speaking, the framework shows that combining easily accessible clinical and lab data with ensemble learning can produce diagnostic accuracy on par with more resource-intensive methods.

6. *Limitations and Future Work*

While COPD-RiNet demonstrates promising performance, this study has certain limitations. The dataset used is simulated and may not capture all real-world variations in patient demographics, comorbidities, or environmental exposures. External validation on diverse, multi-center, and longitudinal datasets is necessary to confirm generalizability. Additionally, the current implementation focuses on binary classification of COPD presence and does not address disease severity or progression.

Future work will focus on integrating disease severity staging and progression prediction into COPD-RiNet, enabling the model to provide not only risk identification but also prognostic insights. Incorporating additional clinical and imaging features, evaluating temporal trends in patient data, and deploying the model in real-time telemedicine settings are further directions that could enhance its clinical applicability and impact.

REFERENCES

1. World Health Organization. (2023). *COPD (COPD)*. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
2. Topalovic, D., Das, N., Burgel, P. R., et al. (2019). Artificial intelligence outperforms pulmonologists in interpreting pulmonary function tests. *European Respiratory Journal, 53*(4), 1801660. https://doi.org/10.1183/13993003.01660-2018

3. Bhatt, S. P., Dransfield, M. T., et al. (2018). Smoking history alone is insufficient to identify patients with COPD. *Chest, 153*(3), 764–775. https://doi.org/10.1016/j.chest.2017.10.025

4. Ghosh, R., Saha, S., & Das, A. (2020). Detection of COPD using CT images and support vector machines. *Medical Imaging Review, 45*(2), 89–99.

5. Sun, X., Huang, Y., & Wang, X. (2019). Deep learning-based COPD classification using CT scans. *IEEE Transactions on Biomedical Engineering, 66*(8), 2221–2231. https://doi.org/10.1109/TBME.2018.2885580

6. Zhang, Y., & Li, T. (2021). A hybrid KNN-ANN approach for predicting COPD. *Healthcare Analytics, 5*, 100092. https://doi.org/10.1016/j.health.2021.100092

7. Miao, H., Zhang, J., & Lee, C. H. (2018). Predicting COPD using environmental exposure and neural networks. *BMC Public Health, 18*, 1023. https://doi.org/10.1186/s12889-018-5863-5

8. Lin, C. H., Tsai, Y. L., & Chang, Y. J. (2020). Using decision trees to identify lifestyle risk factors for COPD. *Journal of Thoracic Disease, 12*(6), 3256–3263.

9. Sharma, A., Gupta, R., & Patel, M. (2019). Prediction of COPD in rural population using hematological parameters. *Journal of Clinical Diagnosis, 28*(3), 45–52.

10. Ahmed, M., Rahman, M. A., & Ali, A. (2021). Ensemble learning techniques for respiratory disease detection. *Computer Methods and Programs in Biomedicine, 200*, 105832. https://doi.org/10.1016/j.cmpb.2021.105832

11. Kang, S., Kim, S., & Yoon, S. (2020). COPD detection using deep CNNs on chest X-ray images. *Medical Image Analysis, 65*, 101771. https://doi.org/10.1016/j.media.2020.101771

12. Song, J., Tan, H., & Zhao, L. (2019). Linear classifiers for COPD screening. *Journal of Medical Informatics, 33*(4), 212–220.

13. Li, R., Zhang, W., & Chen, Y. (2018). Mutual information-based feature selection for COPD prediction. *Computers in Biology and Medicine, 103*, 123–131. https://doi.org/10.1016/j.compbiomed.2018.10.010

14. Chen, Y., & Zhang, H. (2020). Genetic algorithm-optimized SVM for early COPD detection. *Expert Systems with Applications, 160*, 113707. https://doi.org/10.1016/j.eswa.2020.113707

15. Abid, A., Tan, J., & Bai, L. (2021). Comparative study of ensemble vs. base classifiers for COPD detection. *Procedia Computer Science, 184*, 116–123.

16. Tiwari, R., Sharma, P., & Dubey, A. (2020). COPD severity classification using regression models. *Respiratory Medicine Research, 78*, 101–110.

Received: Aug 06, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
17. Devaraj, J., Sahu, A., & Raj, A. (2019). Environmental risk factor modeling for COPD. *Environmental Research, 174*, 219–228.

18. Ramesh, P., & Dey, S. (2021). Predictive modeling of COPD with hematological and smoking data. *BMC Pulmonary Medicine, 21*, 189. https://doi.org/10.1186/s12890-021-01545-5

19. Poon, A. W., Lee, R., & Wong, T. (2022). Clustering COPD phenotypes for personalized therapy. *Digital Health, 8*, 20552076211047233. https://doi.org/10.1177/20552076211047233

20. Yasmin, S., Bashir, F., & Khan, M. (2020). PCA for clinical COPD data enhancement. *Applied Computing and Informatics, 16*(2), 177–184. https://doi.org/10.1016/j.aci.2017.10.004

21. MIT Laboratory for Computational Physiology. (2023). *Simulated Obstructive Lung Disease Dataset* (Version 1.0.0) [Data set]. PhysioNet. https://physionet.org/content/simulated-obstructive-disease/1.0.0/

22. Yin, H., Liu, J., Li, W., & Zhang, Y. (2023). *Fractional dynamics foster deep learning of COPD stage prediction*. Scientific Reports, 13, Article 15704. https://doi.org/10.1038/s41598-023-42974-0

23. Yuan, X., Li, J., Zhou, Y., & Chen, T. (2025). *COPD-MMDDxNet: A multimodal deep learning diagnostic framework for COPD*. Frontiers in Medicine, 12, Article 1601736. https://doi.org/10.3389/fmed.2025.1601736

24. Elango, M., & Selvi, S. T. (2024). *Dragonfly optimized kernel extreme learning machine for COPD diagnosis*. Scientific Reports, 14, Article 2952. https://doi.org/10.1038/s41598-025-02952-6

Received: Aug 06, 2025