Proceedings of the Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS-2026)
DVD Part Number: CFP261CP-DVD; ISBN: 979-8-3195-3129-2

An Interpretable Feature Selection–Driven Extra Trees Framework for COPD Severity Diagnosis using ABG and Biometric Data

Logeshwaran.G, Research Scholar
Department of Computer Science
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai, India
lwlogeshwaran@gmail.com

Dr.S.Kamalakkannan, Professor
Department of Computer Applications
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai, India.
kannan.scs@vistas.ac.in

Abstract- **One of the heterogeneous respiratory diseases is Chronic Obstructive Pulmonary Disease (COPD) that consists of several symptoms namely breathe shortness, persistent together with mucus production, chest tightness as well as wheezing,. Severity assessment in COPD has significant decision making to guide treatment process but the conventional methods are majorly highlight on binary diagnosis together with lagging interpretability to clinical usage. In spite of dataset availability that consists of patient Arterial Blood Gas along with biometrics, there are limited research concentrate on systematic integrated features with understandable Machine Learning (ML) methods in classifying the severity of the disease towards Global initiative for chronic Obstructive Lung Disease (GOLD) stages. To address the above gap, this current study presents a robust along with explainable architecture to classify the COPD severity with multi-classes. However, this current method influences the GOPD status with various available features using several Feature Selection (FS) techniques used for finding out most significant variables together minimize the dimensionality. Subsequently, predictive method used in this research is Extra Tree Classifier (ETC) with k-fold Cross Validation (CV) is considered for enhancing the generalizability along with minimizing overfitting. Hence, the feature transparency can be ensured through SHapley Additive exPlanations (SHAP) gets incorporated for interpreting feature has offered together with support of clinical understanding. Moreover, the experimental results has presented that subset of optimized features has effectively improved the performance of the classification while compared with traditional methods. Especially, the SHAP based FS integrate ETC with k-fold CV are accomplished with best accuracy in maintaining interpretability.**

Keywords: SHapley Additive explanations, Chronic Obstructive Pulmonary Disease, severity diagnosis, Global initiative for chronic Obstructive Lung Disease, Feature Selection, Extra Trees Classifier

1. Introduction

COPD is a respiratory disease that create long-term lung condition may progressively limits the airflow which make complex to patients for breathing along with frequent lead towards hospital visits as well as increases mortality. COPD may result to lung damage that minimizes breathing and the disease gets advanced. During this advanced stage of disease, exacerbations represent the respiratory condition with acute worsen that become more repetitive together with develop towards life-threatening complication. Eventually, this condition inclined to worsen the respiration condition. The major feature of its dwindle is about minimizing the maximal flow rate that can be measure at forceful death subsequently with full insight and this kind of measurement is said to be Peak Expiratory Flow (PEF) rate. The vulnerability of air pollution such as carbon monoxide, nitrogen dioxide, particulate matters and ozone have directed towards lungs inflammatory reaction that indicates the increase in lung's function risks along with symptomatic dwindle [1, 2]. The mortality rate of COPD has become the sixth leading mortality rate causing disease in US at 2020 and resulted to be most death caused associated with chronic low respiratory diseases [3]. In US, the COPD prevalence has continued to be stable at 6% from 2011 to 2021 [4].

Based on the data analysis across 46 states in the US country represented that New York city is considered to be high mortality rate with 10.3% approximately in COPD. In worldwide, the COPD global burden is high and the COPD is estimated to be third leading mortality rate in 2019 as per World Health Organization (WHO) records [5]. There are 213.39 million of COPD prevalent cases have estimated to be 2.5% of prevalence rate which is estimated to be 3.72 million of death because of COPD [6]. Moreover, COPD is considered to be large burden on healthcare system due to hospitalization for individuals with advance COPD as an usual along with recurrent with 90-days of readmission rates ranges from 16-48%. Each hospital admission length is get elevated with mean length of 9 days [7]. There are several COPD risk factors gets identified for adult includes foremost age, BMI less than 18.5 kg/m2, smoking and male sex and children who get admitted in hospital for severe respiration disease, obstructive lung disease as a family history, tuberculosis history, childhood asthma, exposure of biomass, ambient air pollution and occupational exposure to smoke or dust [8]. Thus, the urge intervention has become critical for the life quality of patients as an earlier intervention that slow down the disease progression along with permitting patients for managing symptoms with high efficient [9]. Moreover, the COPD diagnosing with gold standard has become spirometry testing [10].

Many researchers are focused on evaluating the traditional ML-based method performance along with quality as well as Deep Learning (DL)-based prognostic method to patients with COPD together with comparing this model performance with earlier predictive regression methods. To the evaluation purpose with clinical benefits with both DL and ML methods along with focusing the model performance with respect to selected features or method architecture. ML methods are assisted in identifying along with patient prediction who have COPD or possible suggestible for developing COPD towards future. There are several ML methods are utilized in predicting along with assessing a patient's potential in developing COPD. Most of these completely depend upon small dataset that may lead to less reliable methods. Moreover, there are certain ML methods have trained by spirometry along with imaging data that may get tested not available for all patients. Hence, the most ML methods have developed for predicting the exacerbation COPD amount subjected to experience the subsequent year [11]. There are certain methods to predict the exacerbations occurrence over broad forthcoming days with temporal window [12, 13]. Several factors are utilized as predictors includes individual patient-reported that Forced Expiratory Volume for one second (FEV1), level of C-Reactive Protein (CRP), measurement of potential spirometry, auscultation data of lung electronic, symptom of patients, lifestyle data, air pollutant exposure, influenza virus data and lifestyle data have measure frequently from monitoring fixed air quality [14, 15].

This limitation has minimized clinical relevance as treatment decision along with prognostic evaluation based upon accurate severity instead of mere detection. Moreover, traditional ML method often inadequacy in interpretability, make it complex for clinicians for adopting them in practice and few studies have systematic integration of physiological relevance variable towards an explainable ML methods. Hence, this research concentrate on improving prediction on COPD using FS concept that assist in minimizing the bias of the model sample and extracting the ML model with k-Fold CV to reduce the overfitting. Thus, the prediction on COPD severity get improved through better selection of feature with highly significant features weights together with k-fold CV may generate good predictive ML model.These experimental study further proceeds with four sessions are session 1 detailed with literature support about the significant studies that focus on predicting COPD severity and lung cancer using several feature selection and feature extraction methods. Session 2 discusses the proposed research methodology of different FS techniques with K-fold CV models for predicting GOLD stage of COPD severity. Session 3 determines the proposed SHAP-based ML with K-Fold CV classifiers model with other conventional methods for predicting COPDseverityalong with session 5 as finally presented the conclusion of the proposed SHAP-ETC with 5-fold CV classifiers model in predicting COPD severity precisely.

2. Literature Review

This literature study concentrated on various predictive ML methods are generated for influencing computational power together with clinical databases in improving the predictive system as well as COPD exacerbation characterizations. Emam et al. have focused on improving the ML method efficiency over medical diagnosis using various FS techniques and applied in the medical data along with identifying relevant features. It has no capability in capturing the interactions among features that gets mandatory to medical data [16]. Xia et al. have developed Support Vector Machine with Recursive Feature Elimination (SVM-RFE) in segregating COPD patients based on management of treatment guidelines from patients and the proposed SVM-RFE method has accomplishes AUC with 0.987 [17]. Tarek et al. have presented a FS technique using Snake Optimization Algorithm (SOA) in which the cardiovascular diseases are detected rapidly. This proposed technique is bio-inspired along with has a high classification rate that assist in minimizing the feature dimensionality. Thus, the potential Evolutionary Algorithm (EA) significance gets formulated with the solution of healthcare quality which gets interpretable [18]. Pramanik et al. have utilized the adaptable deep FS technique to identify the pneumonia by integrating Particle Swarm Optimization (PSO) with DL features obtained from the pre-trained methods [19]. Ladani and semnani have presented the novel RegNet together with XOR-based PSO model to select deep features over pneumonia diagnosis and produce high accuracy levels [20].

Ku et al. have employed the endured CV method in both gradient boosting as well as Gaussian process regression (GPR) methods correspondingly in developing two distinct prediction methods. In this research, the data is segregated towards 75% of data for training and remaining 25% of data for testing. Moreover, the developed prediction model to the test set is utilized to determine the estimated respiratory disease patient count. Hence, the performance of each methods get evaluated through coefficient of determination (R2) along with Root Mean Square Error (RMSE) among actual and predicted counts of respiratory disease patients. Thus, the SHAP technique is used to interpret the predicted findings of each ML method. SHAP analysis has demonstrated that average temperature deduction, average humidity, total amount of solar isolation, daylight period along with temperature differences has increases the respiratory disease patient amount. It increases the atmospheric pressure, particulate matter lesser than 2.5 µm in the aerodynamic diameter and carbon monoxide have increases the respiratory disease patients number [21].

Orangi-Fard et al. have highlighted the performance of top predictive methods performance in breathing pattern by leave-one-out CV (LOOCV) together with k-fold CV correspondingly. This study concentrated on both 5-fold along with 10-fold CV that performed for assessing model robustness. Moreover, these findings with LOOCV to data partition along with validation. Hence, the prediction of breathing pattern accuracy is 88% while the breathing rate estimation has involved another input features with no LOOCV accuracy is 85%. Similarly, the k-fold CV is analyzed with SVM method and Random Forest Classifier (RFC) and the best method performance is accomplished for both methods. The breathing pattern prediction accuracy is of 75% for RFC at 5-folds and in the case of 10-folds with SVM, the breathing pattern prediction accuracy is 76%. Thus, the proposed method irrespective of data partition along with validation method performance of RFC has performed better in both LOOCV as well as k-fold CV than other ML methods for the clinical tasks [22].

3. Research Methodology

One of the global health burden caused disease is COPD that can be identified through progressive airflow limitation together with high morbidity. However, there are several diagnostic methods existing methods concentrated majorly on binary classification that can able to distinguish among healthy as well as diseased status with no adequate address about severity gradation in term of GOLD stages. Therefore, the limit restriction in their clinical relevance as decision making in treatment along with prognostic evaluation highly based on severity assessment accuracy instead of common detection. In spite of availability over dataset namely "Lung condition patient (COPD Severity)" consists of ABG and biometric parameters. Most of the research is basically integrating these relevant features towards interpretable ML architecture. Moreover, traditional ML model transparency is lagging that hinders the clinicians to understand feature contributions and it reduces the trust together with adopting over real-time setup. Hence, there is a vital requirement for robust along with explainable ML architecture about ability in severity classification of multi-class COPD. Figure 1 illustrate the architecture of proposed SHAP-based ML with K-Fold CV classification that focus on efficient integration of FS technique with k-fold CV ML method for identifying the most revealing features that uses predictable classifier namely ETC with k-fold CV to confirm generalizability together with explainable tool encompassing namely SHAP for improving the interpretability. When addressing these gaps will not only enhance predictive accuracy and even consolidate the clinical decision-making that enables personalized management together with earlier intervention to COPD patients.

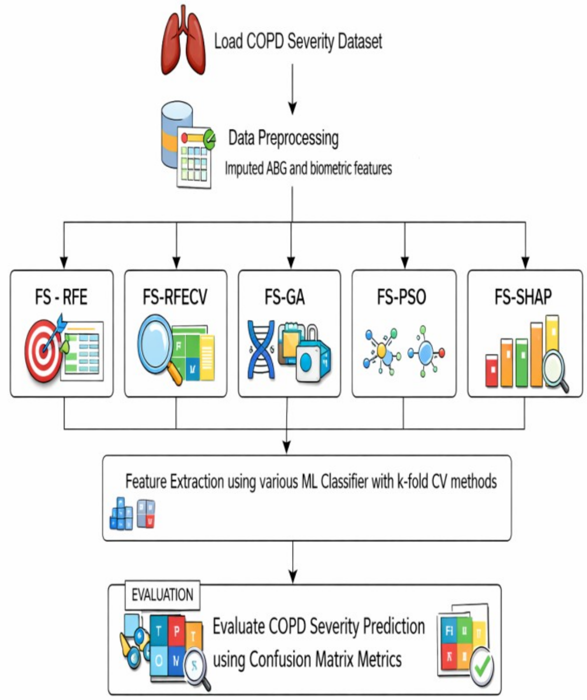

Figure 1. SHAP-based ML with K-Fold CV classifiers to detect GOLD stages of COPD

A. Data collection

The dataset used in this research work is an open source in which the comprehensive patient record collections has designed for predicting health outputs especially to the respiratory condition context like COPD. This integrates two major feature factors such as lifestyle factors and environmental factors that can be offered through holistic patient health persuasion. Lifestyle factors illustrate about patient behaviors that involves exercise frequency, smoking status together with further habits which prominently influences the overall patient's records. According to environmental features that involves feature namely air quality, temperature together with subsequent environment conditions which influence direct towards respiratory functions. Each records in the dataset relates to the COPD patients with labels represent their health status along with level of severity through GOLD stage towards four distinct classes. Moreover, the dataset contain 899 patients records as overall sample that get distributed towards four distinct severity stages namely mild (stage 1) as 309 records, moderate (stage 2) as 237 records, severe (stage 3) as 180 records and very severe (stage 4) as 173. Hence, this "Lung patient condition (COPD severity)" dataset can able to serve as beneficial resources for the healthcare professionals who seek for improving COPD severity assessment along with patient management recommendations.

B. Data Preprocessing

In data preprocessing, missing imputation is considered to be the initial process in which the feature space missed data are managed by median imputation process by simpleimputer. This assists to minimize the extreme value impact along with data consistency management. Once it is done, the imputed data get reconstructed towards clean data frames with proper aligned columns that is significance to compatibility with FS techniques and also with ML k-fold CV classifier models. Moreover, categorical labels are encoded towards numerical form by LabelEncoder that is considered as target variable and ensured the dependent variable is exactly represented to the respective classification target classes. As final, the target vector has reshaped towards one dimensional array to come across the input need of scikit learn classifier. Thus, the preprocessing workflow ensures the dataset is clean and get segregated as 80% for training and 20% for testing to be the data split.

C. Working of SHAP

The SHAP technique manage on the cooperative game theory foundation in which every feature to the predictive method has been considered to be a player contributor to the final outcome. The SHAP significance fabricate to fair distribution of model's prediction between any features with respect to their individual contributions. Basically, the model f(x) contain M features has focused for determining each feature "i" provide to the model's output deviations from an average prediction E[f(x)]. The SHAP score (ϕ_i) mentioned their contribution along with expression in equation 1.

$$\phi_i = \sum_{S\subseteq F/\{i\}} \frac{|S|!\,.(M-|S|-1)!}{M!}[f(S\cup\{i\}) - f(S)] \quad (1)$$

Where,

S = Any probable subsets of a feature exclude “i”

$f(S\cup\{i\}) - f(S)$ = Feature marginal contribution for “i” during subset addition.

This factorial term has made certain legitimate weighted over all combinations that maintain correspondence together with efficiency. The commence of workflow with model's baseline prediction computations as E[f(x)] accompanied by evaluation of each feature that get modified the prediction while involved. These marginal impacts have accumulated in generating the SHAP score that measures the each feature significance. The model prediction is expressed in equation 2.

$$f(x) = E[f(x)] + \sum_{i=1}^{M}\phi_i \quad (2)$$

At last, the SHAP score with mean absolute over all samples have been employed for ranking features through its influence is expressed in equation 3.

$$Importance(i) = \frac{1}{N}\sum_{n=1}^{N}|\phi_{i,n}| \quad (3)$$

The top ranked SHAP features are Partial pressure of oxygen (PaO_2), Partial pressure of carbon Dioxide (($PaCO_2$), Blood acidity (pH), oxygen saturation ((SpO_2), smoking status, Body Mass Index (BMI) and age. The significance of SHAP derived features that align COPD clinical trajectories. Features related to mild or moderate stage as an early stage are decline in PaO_2, SpO_2 with respect to smoking and age is considered to be risk factors. Features associated with severe or very severe stage as an advanced stage are elevated $PaCO_2$, imbalance pH and with low BMI. However, the SHAP integration ensures the feature significance in both mathematical rigorous together with clinical relevance. By measuring the ABG contribution along with biometric parameter, this SHAP based FS assist in bridging the gap among computational prediction and decision making in COPD management. This kind of ranking has assist FS technique to retain the most significant features for model training. This research concentrate on GOLD status as COPD severity classification assist in finding out which parameter has most efficient in predicting severity and ensures both interpretability along with clinical pertinence.

When considering FS process before classification training, each FS method is analyzed and compared with complete original dataset with 899 records and 22 available features. The number of retained features is justified to be significant features to determine the better result with classifier model. There are different FS involved in this research are Recursive Feature Elimination (RFE), Genetic Algorithm (GA), Recursive Feature Elimination with Cross-Validation (RFECV), SHAP and PSO. RFE technique consists of selected 14 features from the 22 features focused on ABG parameters along with BMI and smoking status. RFECV technique get selected 12 features from 22 features with balanced predictive accuracy with generalization through CV. GA technique get selected 15 features from 22 features with optimized subset by evolutionary search includes both ABG and environmental factors. PSO technique get selected 13 features from 22 features emphasize respiratory biomarkers and lifestyle indicators. SHAP technique get selected 10 most influential features consistently ranking as top contribution to COPD severity progression. This SHAP-based FS is computationally efficient along with clinical meaningful and minimize the redundancy while preserving diagnostic insight.

D. ETC with k-fold CV working principle

ETC with k-fold CV working principle initiated by data split D towards k-fold is represented as D={D1,D2,…,DK}. For each fold, the model get trained on for all fold other than Kth as well as validated by Dk. This ensures that each data subset has been utilized to both training along with validation. In ETC framework, the introduction of tree construction randomness at every node has a random feature subset along with the threshold of random split have produced to those features. Each split quality gets evaluated through impurity measurement namely Gini Index and it is expressed in equation 4.

$$G = 1 - \sum_{c=1}^{C} p_c^2 \quad (4)$$

Where,

pc = Sample proportion to class c at the node

With no pruning and capture of complex relationship over data is done while trees are grown to their maximum depth. This kind of process gets iterative over all folds along with the metrics performance are recorded to each validation set. At last, the outcome get averaged over folds and provide a robust classifier estimation generalization capability during reduction og variance and bias. Hence, this randomized tree construction integration together with systematic CV involved in ETC has high efficiency to difficult set of classification tasks namely COPD severity prediction. During the ETC with k-fold CV with ensemble prediction stage, each fold generates its individual accumulation of classification outcome using prediction combination of all trees in the folds. Particularly in k-fold, the prediction can be determined through major voting over T trees is expressed in equation 5.

$$\hat{y}_k = argmax_{c \epsilon C} \sum_{t=1}^{T} I(T_{k,t}(x) = C) \quad (5)$$

Where,

I(.) = Indicator function

$T_{k,t}(x)$ = Tree prediction “t” in k-fold

C = Potential classes as target

However, it make sure that the final prediction to each fold exhibit the ensemble unified decision instead of a single tree. Hence, the prediction is accomplished to all folds and the final performance gets estimated by measuring the average accuracy of k-fold is expressed in equation 6.

$$Accuracy = \frac{1}{K} \sum_{k=1}^{K} Accuracy(D_k) \quad (6)$$

Thus, the provision of averaging process with robust along with unbiased classifier’s measure of the generalization capability and minimize the overfitting risk and ensure that the model’s performance has consistent across distinct dataset partitions. In addition, the algorithm for SHAP- ETC with k-fold CV is explained below

E. Pseudocode for SHAP based ETC with k-fold CV

- Input: Dataset D, K folds, T trees
- Output: Feature importance ranking, performance metrics

```
Split D into K folds

Initialize metrics_list = []

Initialize shap_values = []

For k in 1..K:

D_train = D \ D_k

D_val = D_k

  # Train Extra Trees Classifier

model = ExtraTreesClassifier(T)

model.fit(D_train)

  # Compute SHAP values

shap_k = SHAP(model, D_train)

shap_values.append(shap_k)

  # Rank features

importance = mean(|shap_k|)

  # Validate

predictions = model.predict(D_val)

metrics = evaluate(predictions, D_val.labels)

metrics_list.append(metrics)

# Aggregate results

final_importance = aggregate(shap_values)

final_metrics = average(metrics_list)

Return final_importance, final_metrics
```

4. Result and Discussion

This experimental research work has promote a detailed comparison about proposed SHAP-ETC with k-fold CV enable detection with high accurate together with interpretable method to predict multi-class COPD severity. This experiment has utilized the jupyter Notebook environment for python on the Google Colab platform that influences libraries like scikit-learn, SHAP, Pandas, LightGBM and NumPy. The computational environment involves high performance VM with intel Xeon processors, 16GB optional GPU, 12 GB RAM to ensure effective training along with several ML methods evaluation. The four COPD severity stages are defined with GOLD criteria that mention the progressive disease classification with respect to lungs function together with symptom burden. Table 1 illustrate the dataset contain 899 patients records as overall sample that get distributed towards four distinct severity stages namely mild (stage 1) as 309 records, moderate (stage 2) as 237 records, severe (stage 3) as 180 records and very severe (stage 4) as 173. Hence, this "Lung patient condition (COPD severity)" dataset can able to serve as beneficial resources for the healthcare professionals who seek for improving COPD severity assessment along with patient management recommendations. When preparing the data for method development along with evaluation, the overall sample get segregated towards 80% of sample dataset for training as 719 records and 20% of sample dataset for testing as 180 records. This provides an unbiased model performance evaluation. This stratified split maintains the dataset integrity while requiring the balance to the robust training along with predictable validations.

Table 1 amount of sample with severity level of COPD status

COPD Status	**Severity Level**	**Records Available**
1 – Stage1	Mild	309
2 – Stage2	Moderate	237
3 – Stage3	Severe	180
4 – Stage4	Very Severe	173

This experimental setting is designed to compare the various FS techniques impacts involves Recursive Feature Elimination (RFE), Genetic Algorithm (GA), Recursive Feature Elimination with Cross-Validation (RFECV), SHAP and PSO. These techniques are employed to identify the most significant features from several parameters involved in the dataset. The subset of selected feature are then evaluated by several tree-based ML classifier with k-fold CV such as DT, RFC, ETC and Light Gradient Boosting Machine (LGBM). From the above ML classifier with k-fold CV, ETC with k-fold CV and RF with k-fold CV is ensemble-based method for providing robust performance because of its capability in managing feature interactions together with minimized variance. In the case of LGBM, it can provide efficient gradient boosting with both rapid training along with improved accuracy. The SHAP technique integration has not only improve the FS process and also assist in interpretability using quantified FS as utilize the method suitability to support clinical decision.

Table 2 Performance comparison of different FS with various tree-based ML classifications

Feature Selection	**Classifier method with 5-fold CV**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
RFE	DT	86.11	86.14	88.78	87.44
	RF	86.67	86.27	89.80	88.00
	LGBM	87.22	85.71	91.84	88.67
	ETC	**88.33**	**87.38**	**91.84**	**89.55**
RFECV	DT	86.67	86.27	89.80	88.00
	RF	88.33	87.38	91.84	89.55
	LGBM	88.89	88.46	92.00	90.20
	ETC	**89.44**	**88.57**	**93.00**	**90.73**
GA	DT	89.44	89.42	92.08	90.73
	RF	88.33	87.38	91.84	89.55
	LGBM	**90.56**	**90.48**	**93.14**	**91.79**
	ETC	89.44	88.57	93.00	90.73
PSO	DT	90.56	90.48	93.14	91.79
	RF	91.11	91.35	93.14	92.23
	LGBM	91.67	91.43	94.12	92.75
	ETC	**92.22**	**92.38**	**94.17**	**93.27**
SHAP	DT	91.15	91.38	93.17	92.29
	RF	91.69	91.46	94.15	92.79
	LGBM	92.22	92.38	94.17	93.27
	ETC	**92.78**	**93.33**	**94.23**	**93.78**

The performance comparison in Table 2 demonstrates how different FS methods influence the effectiveness of tree-based classifiers under 5-fold CV. There are five different FS with tree-based classifiers using 5-fold CV namely DT, RF, LGBM and ETC. The performance metrics considered are accuracy, precision, recall and F1-score. From the five FS technique SHAP technique has performed better outcome with accuracy of DT, RF, LGBM and ETC are 91.11%, 91.67%, 92.22% and 92.78%. Similarly, F1-score is the major performance metrics that can predict COPD status better for tree-based methods are DT, RF, LGBM and ETC are 92.29%, 92.79%, 93.27% and 93.78%. This clearly determines that SHAP based ML methods with 5-fold CV have better result while compared to other FS methods of respective ML methods metrics.

Table 3 Performance comparison for different FS with best ML classifier methods

FS with Classifier method with 5-fold CV	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
RFE-ETC	88.33	87.38	91.84	89.55
RFECV-ETC	89.44	88.57	93.00	90.73
GA-LGBM	90.56	90.48	93.14	91.79
PSO-ETC	92.22	92.38	94.17	93.27
SHAP-ETC	92.78	93.33	94.23	93.78

Table 3 demonstrate the best ML with 5-fold CV methods from each FS technique in which SHAP-ETC is 92.78% as best accuracy while compared to other best model of FS methods are RFE-ETC, RFECV-ETC, GA-LGBM and PSO-ETC are 88.33%, 89.44%, 90.56% and 92.22% respectively. Moreover, the This show the better section of FS with 5-fold

CV of ML method has better accuracy in predicting GOLD status of COPD severity. Moreover, F1-score determines the better performance metrics that can predict COPD status precisely better than existing methods. The F1-Score of SHAP-ETC is 93.78% as best while compared to other best model of FS methods are RFE-ETC, RFECV-ETC, GA-LGBM and PSO-ETC are 89.55%, 90.73%, 91.79% and 93.27% respectively.

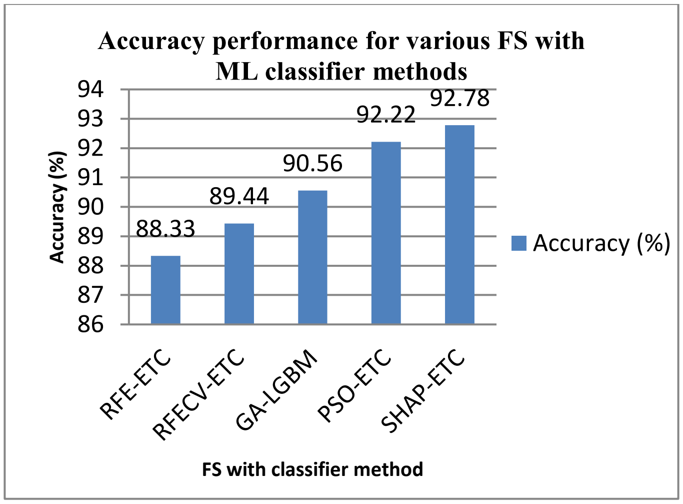

Figure 2 Comparison of accuracy performance for various FS with ML classifier in COPD severity detection

Figure 2 illustrate the best ML with 5-fold CV methods from each FS technique in which SHAP-ETC is 92.78% as best accuracy while compared to other best model of FS methods are RFE-ETC, RFECV-ETC, GA-LGBM and PSO-ETC are 88.33%, 89.44%, 90.56% and 92.22% respectively. This show the better section of FS with 5-fold CV of ML method has better accuracy in predicting GOLD status of COPD severity.

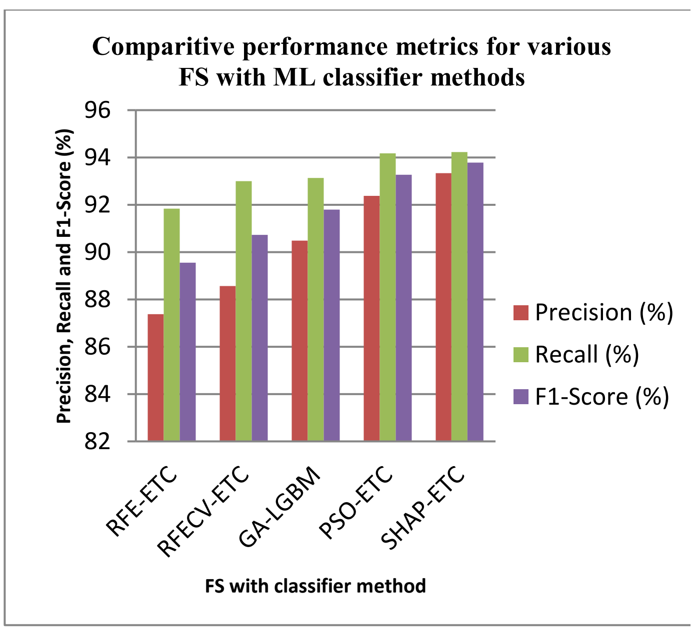

Figure 3 Comparison of performance metrics for various FS with ML classifier in COPD severity detection

Figure 3 illustrate the comparison performance about F1-score, recall (True positive rate) and precision in which SHAP-ETC with 5-fold CV has accomplish high performance metrics namely F1-Score, recall and precision are 93.78%, 94.23% and 93.33% respectively. Followed by PSO-ETC with 5-fold CV and the least method is RFE-ETC with F1-Score, recall and precision are 89.55%, 91.84% and 87.38% respectively. This determines the better section of FS with 5-fold CV of ML method with SHAP-ETC with 5-fold CV has better accuracy in predicting GOLD status of COPD severity.

The robustness of proposed SHAP-ETC model is evaluated through two different clinical scenarios namely missing ABG value and noisy ABG measurement. In the case of missing ABG value, the values are imputd by median statistics that minimize the extreme outlier impacts. It involves 10% to 20% of ABG features were masked randomly. In the scenario of noisy ABG data, Gaussian noise ($\sigma = 0.05$) is considered for features like pH, $PaCO_2$ and PaO_2 for mimic measurement variability. Table 4 illustrate the robustness summary of the proposed SHAP-ETC with complete dataset, 20% missing ABG dataset and noisy ABG with $\sigma = 0.05$.

Table 4 Robustness summary of various ABG dataset

Dataset Scenarios	**F1-Score (%)**	**Accuracy (%)**	**Stability of Model Interpretability**
Complete ABG Dataset	93.78	92.78	High
20% of Missing ABG	92.1	91.2	High
Noisy ABG with $\sigma = 0.05$	91.5	90.9	High

Table 4 illustrates the robustness summary of proposed SHAP-ETC model with three different scenarios whereas FS with complete dataset with accuracy, F1-score and stable interpretability are 92.7%, 93.78% and high respectively. In the missing ABG with 20% maximum, the accuracy and F1-score are 91.2% and 92.1% which comparatively lesser than complete dataset that demonstrate efficient imputation is justified with high stable interpretability. In the scenario of noisy ABG dataset for the features of PaO_2, $PaCO_2$, and pH, the accuracy and F1-score are 90.9% and 91.5% that get minimized while compared to complete dataset and 20% missing ABG dataset. Overall, the proposed SHAP-ETC model has reliability to classify the COPD severity in real-time environment.

Table 5 Execution time comparison for different FS with best ML classifier methods

FS with Classifier method with 5-fold CV	Execution Time (Sec)
RFE-ETC	75
RFECV-ETC	92
GA-LGBM	118
PSO-ETC	135
SHAP-ETC	168

Table 5 illustrates the execution time comparison in computational demand over FS methods. SHAP-ETC has recorded the longest runtime at 168 Sec and computes marginal contribution across all features and trees but the modest increase in justifying its high accuracy along with interpretability. RFE-ETC is the least runtime at 75 Sec followed by RFECV-ETC requires 92 Sec, GA-LGBM has considered 118 Sec reflect the iterative nature of genetic search and PSO-ETC has involved 135 Sec due to swarm-based optimization cycles.

However, complexity of SHAP require to evaluate the marginal contribution of each features over multiple subsets and it measured by $0(M.N.T)$.

Where,

M = Feature amount

N = Sample number

T = Tree number ensemble

Hence, this complexity is higher than simple method such as RFE and GA but it ensures the interpretability by quantifying features significant rigorously.

Limitation

In spite of likely results in SHAP-ETC model, there are certain limitations are identified and initially the study depends upon open source dataset with 899 patient records with 22 ABG along with biometric features. This dataset get balanced distribution over GOLD stages and its size and scope is limited. The execution time is basically consumed more due to tree based ensemble method as ETC is integrated with SHAP-based FS consume more execution time while compared to simple FS technique like RFE, RFECV, GA ad PSO techniques. Moreover, the complexity of model is high due to ETC as ensemble method and it integrate with SHAP based FS method because of involvement of various features, sample and trees.

5. Conclusion

This research work proposed consist of explicable ML architecture for classifying multi-class COPD severity using integrated advanced FS technique with tree-based ML classifier by 5-fold CV. The findings highlight the optimal FS has significantly influences the performance of the methods. There are traditional and optimized based FS technique has limited amount of capturing complex feature interaction along with limited notable performance gain. From the above method, SHAP-based ETC with 5-fold CV has most efficient and it improves the predictive accuracy if ensure interpretability. In contrast, feature extraction method minimizes dimensionality at the clinical relevance task. Hence, the SHAP has capability in retaining the original features along with ranking them with respect to actual contribution for predictions. However, the characteristic has specific valuable in healthcare application in which transparency is significant. Hence, the SHAP-ETC with 5-fold CV can accomplished best performance metrics as 92.78% in accuracy, 93.78% in F1-score, 94.23% in recall along with 93.33% in precision. Moreover, the robustness in stability of model interpretability is high for complete ABG dataset, 20% missing ABG dataset and noisy ABG dataset. Thus, the proposed SHAP-ETC 5-fold CV has better robustness in predicting GOLD stage of COPD severity with consistent results. Future work of this research concentrate on parallelized SHAP FS technique for minimizing the execution time with no sacrificing interpretability along with DL methods integrated using explainable AI for improving the accuracy of the model in diagnosing COPD severity. In addition, the limited dataset can be expand by considering multi-center clinical records along with longitudinal patient monitor data to improve external validity.

Reference

[1] Evangelopoulos, D. et al. Personal exposure to air pollution and respiratory health of COPD patients in London. Eur. Respir. J.[SPACE] https://doi.org/10.1183/13993003.03432-2020 (2021).

[2] Boehm, A. et al. COPD exacerbations are related to poor air quality in Innsbruck: A retrospective pilot study. Heart Lung 50, 499–503. https://doi.org/10.1016/j.hrtlng.2021.02.012 (2021).

[3] Syamlal, G.; Kurth, L.M.; Dodd, K.E.; Blackley, D.J.; Hall, N.B.; Mazurek, J.M. Chronic Obstructive Pulmonary Disease Mortality by Industry and Occupation—United States, 2020. MMWR Morb. Mortal. Wkly. Rep. 2022, 71, 1550–1554. [CrossRef]

[4] 3. Liu, Y.; Carlson, S.A.; Watson, K.B.; Xu, F.; Greenlund, K.J. Trends in the Prevalence of Chronic Obstructive Pulmonary Disease Among Adults Aged≥18 Years—United States, 2011–2021. MMWR Morb. Mortal. Wkly. Rep. 2023, 72, 1250–1256. [CrossRef]

[5] 4. Lindberg, A.; Lindberg, L.; Sawalha, S.; Nilsson, U.; Stridsman, C.; Lundback, B.; Backman, H. Large underreporting of COPD as cause of death-results from a population-based cohort study. Respir. Med. 2021, 186, 106518. [CrossRef]

[6] 5. Wang, Z.; Lin, J.; Liang, L.; Huang, F.; Yao, X.; Peng, K.; Gao, Y.; Zheng, J. Global, regional, and national burden of chronic obstructive pulmonary disease and its attributable risk factors from 1990 to 2021: An analysis for the Global Burden of Disease Study 2021. Respir. Res. 2025, 26, 2. [CrossRef]

[7] 6. Bakthavatsalu, B.; Walshe, C.; Simpson, J. The experience of hospitalization in people with advanced chronic obstructive

pulmonary disease: A qualitative, phenomenological study. Chronic Illn. 2023, 19, 339–353. [CrossRef]

[8] 7. Adeloye, D.; Song, P.; Zhu, Y.; Campbell, H.; Sheikh, A.; Rudan, I.; NIHR RESPIRE Global Respiratory Health Unit. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: A systematic review and modelling analysis. Lancet Respir. Med. 2022, 10, 447–458. [CrossRef] [PubMed]

[9] K. Kahnert, R. A. Jörres, J. Behr, and T. Welte, "The Diagnosis and Treatment of COPD and Its Comorbidities," DtschArzteblInt, vol. 120, no. 25, p. 434, Jun. 2023, doi: 10.3238/ARZTEBL.M2023.027.

[10] Agusti et al., "GOLD COPD DOCUMENT 2023: a brief update for practicing cardiologists," Clinical Research in Cardiology, vol. 113, no. 2, pp. 195–204, Feb. 2024, doi: 10.1007/S00392-023-02217-0/FIGURES/3.

[11] Zeng, S., Arjomandi, M., Tong, Y., Liao, Z. &Luo, G. Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: Retrospective cohort study. J. Med. Internet Res. 24, e28953. https://doi.org/10.2196/28953 (2022).

[12] Patel, N., Kinmond, K., Jones, P., Birks, P. &Spiteri, M. Validation of copdpredictTM: Unique combination of remote monitoring and exacerbation prediction to support preventative management of copd exacerbations. Int. J. Chron. Obstruct. Pulmon. Dis. 16, 1887–1899. https://doi.org/10.2147/COPD.S309372 (2021).

[13] Wu, C. et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: Development and cohort study. JMIR MhealthUhealth 9, e22591. https://doi.org/10.2196/22591 (2021).

[14] Yin, H. et al. A machine learning model for predicting acute exacerbation of in-home chronic obstructive pulmonary disease patients. Comput. Methods Programs Biomed. 246, 108005. https://doi.org/10.1016/j.cmpb.2023.108005 (2024).

[15] Jo, Y., Han, S. & Lee, D. Development of a daily predictive model for the exacerbation of chronic obstructive pulmonary disease. Sci. Rep. 13, 18669. https://doi.org/10.1038/s41598-023-45835-4 (2023).

[16] Emam, S. S., Arafa, M. M., El-Attar, N. E. &Elshishtawy, T. Feature Selection in Medical Data as Coping Review from 2017 to 2022 155–171 https://doi.org/10.1007/978-981-99-4764-5_10 (2023).

[17] Xia, J. et al. <p>A Model using support vector machines recursive feature elimination (SVM-RFE) algorithm to classify whether COPD patients have been continuously managed according to GOLD guidelines</p>. Int. J. Chron. Obstruct. Pulmon. Dis. 15, 2779–2786. https://doi.org/10.2147/COPD.S271237 (2020).

[18] Tarek, Z., Alhussan, A. A., Khafaga, D. S., El-Kenawy, E. S. M. &Elshewey, A. M. A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biomed. Signal Process. Control 102, 107417 (2025).

[19] Pramanik, R., Sarkar, S. & Sarkar, R. An adaptive and altruistic PSO-based deep feature selection method for Pneumonia detection from Chest X-rays. Appl. Soft Comput. 128, 109464. https://doi.org/10.1016/j.asoc.2022.109464 (2022).

[20] Ladani, F. G. &Semnani, S. H. Optimized Deep Feature Selection for Pneumonia Detection: A Novel RegNet and XOR-Based PSO Approach (2023).

[21] Ku Y, Kwon SB, Yoon JH, Mun SK, Chang M. Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors. ClinExpOtorhinolaryngol. 2022 May;15(2):168-176. doi: 10.21053/ceo.2021.01536.

[22] N. Orangi-Fard, A. Bogdan, and H. Sagreiya, "Machine learning-based algorithms for at-home respiratory disease monitoring and respiratory assessment," arXiv preprint arXiv:2409.03180, Sep. 2024