EVALUATING THE PREDICTION OF COPD USING DATA ANALYSIS AND ENSEMBLE MACHINE LEARNING TECHNIQUES

Arpita Nath Boruah[1], Mrinal Goswami[1], Elchin Rzayev[2]

•

[1]Faculty of Computer Technology, Assam down town University, Assam, India
[2]Azerbaijan Technical University, Azerbaijan, Baku
arpita.b@adtu.in, mrinal.g@adtu.in, elchin_rzayev@aztu.edu.az

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a progressive and debilitating respiratory condition characterized by persistent airflow limitation, typically associated with chronic bronchitis and emphysema. COPD represents a significant global health burden, affecting millions of individuals worldwide, with increasing prevalence and mortality rates. The primary risk factor for COPD is tobacco smoking, although other factors such as occupational exposure to pollutants, genetic predisposition, and respiratory infections also contribute to its development. Chronic inflammation, oxidative stress, and protease-antiprotease imbalance play pivotal roles in the pathogenesis of COPD, leading to structural changes in the airways and alveoli, progressive airflow limitation, and impaired gas exchange. In recent years, there has been growing interest in applying Machine Learning (ML) techniques to various aspects of COPD management, including diagnosis, prognosis, treatment optimization, and exacerbation prediction. So also data analysis plays an important part in the performance the ML techniques. This work investigates the performance of different machine learning classifiers used in COPD prediction, especially in single and ensemble classification. A detailed performance comparison among all the classifiers is also done, considering accuracy, precision, recall, and F1 score.

Keywords: COPD, Machine Learning, Performance evaluation

I. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a progressive and long-term respiratory condition characterized by persistent breathing difficulties and limited airflow. It is strongly associated with smoking and was the fourth leading cause of death globally in 2010, with projections suggesting it would rise to third by 2020. The Global Burden of Disease Study estimated that in 2016, approximately 251 million people worldwide had COPD, and the disease caused around 3.17 million deaths in 2015.

In the United States, about 21% of COPD patients were readmitted to the hospital within 30 days of discharge, with readmission costs exceeding initial hospitalization costs by 18%. Due to its high prevalence and economic burden, the Centers for Medicare and Medicaid Services (CMS) has identified COPD as a priority for reducing hospital readmissions. As Purdy et al. highlighted, COPD is sensitive to ambulatory care, meaning effective primary or preventive care can help avoid hospitalizations. However, the factors influencing readmissions remain poorly understood.

While cigarette smoke (CS) is a well-known cause of COPD, the effects of smoking, such as airway wall thickening, reduced small airway function, and lung tissue damage (emphysema), vary among individuals, complicating studies of the link between smoking and COPD risk. Key symptoms of COPD include excessive mucus production, persistent coughing, shortness of breath, chest tightness, and wheezing. Early diagnosis and management are crucial for controlling COPD, although no medication currently exists to reverse lung damage caused by the disease.

The spirometry pulmonary function test is a cornerstone and highly efficient tool in primary care for diagnosing COPD among the available diagnostic methods. This test measures patients' lung capacity by repeatedly assessing their inhalation and exhalation. However, with a sensitivity range of only 64.5-79.9%, spirometry often leads to significant underdiagnosis of COPD-related morbidity and mortality. To address this limitation, implementing a reliable machine learning (ML) approach is crucial for accurately diagnosing, managing, and treating COPD. ML offers a powerful means of predicting medical conditions, enabling healthcare providers to make precise decisions. Among the various ML classification methods, eXtreme Gradient Boosting (XGB), Gradient Boosting (GB), and Support Vector Machines (SVMs) are some of the most prominent techniques used to analyze health-related data and identify disease-specific patterns.

Machine learning (ML)-based approaches excel at performing complex computations to identify diseases within large datasets. These models have recently proven effective in minimizing potential errors by healthcare professionals and facilitating the early and accurate diagnosis of various conditions, including Parkinson’s disease, heart disease, Alzheimer’s disease, cervical cancer, liver cancer, breast cancer, and others.

As a result, ML-driven methods can support medical professionals in making informed decisions about a wide range of health conditions, including COPD, while reducing their workload and enabling them to deliver accurate and timely treatments.

In this chapter we comprehensively examined the performance of machine learning models including Random Forest, Support Vector Machine (SVM), Naïve Bayes, Decision Tree and XGBoost and have given a comparison in terms of accuracy, precision, recall and F-measure.

II. Literature Survey

Numerous researchers have explored machine learning algorithms to aid in clinical decision-making for accurately categorizing the severity of various disease so also COPD patients.

In the context of prior research related to conventional machine learning models, Spathis and Vlamos utilized the random forest (RF) classification algorithm to predict COPD [16]. Their study involved 132 medical records containing 22 distinct patient-related attributes. After applying the RF classifier to foresee COPD patients, the authors achieved a precision rate of 97.7%., Fang et al. [17] introduced an integrated approach by combining direct search simulated annealing with SVM for diagnosing COPD using a knowledge graph based on the COPD dataset. This dataset consists of 1200 samples, wherein 750 samples belong to individuals with COPD, and the remaining 450 samples belong to those without the condition.

To identify the most suitable attributes from the input dataset, they employed an adaptive feature subset selection technique.2. Their diagnostic accuracy for COPD stood at an impressive 95.1%. Dhar et al. [18] introduced an innovative ensemble approach for the prompt identification of COPD [23]. The researchers employed two sets of 8 classifiers each. They employed a genetic algorithm to discover the best hyperparameters for each classifier.

The outcomes of their model surpassed the performance of numerous contemporary machine learning models used in early COPD detection. Porkodi et al. [19] have presented a feature extraction method for the structural representation of COPD images using the Gabor Filter.

Furthermore, they trained and assessed COPD-derived functions or categories using SVM classification. The findings indicated that the suggested approach exhibits greater accuracy, flexibility, and dependability.

Raja and Babu [20] have proposed an ensemble classification model considering a feature selection as a pre-processing step using the image dataset to classify the disease severity. They have applied five different classifier method for the validation in terms of false positive measures. With the aid of fundamental indicators, comorbidities, and inflammation after admission, Peng et al. [21] applied the C5.0 decision classifier to quickly detect the deterioration and death risk of AECOPD patients. This paper's C5.0 decision classifier successfully predicted 80.3% of occurrences. In a systematic investigation, Min et al. [22] developed a variety of machine learning models, both deep and shallow, to forecast the probability of readmission for COPD patients. On a real-world database containing the medical claims of 111,992 patients from the Geisinger Health System from January 2004 to September 2015, they have assessed those various ways.

They have based their machine learning models on both knowledge-driven and data-driven patient features, which are features that have been derived from the patient's actual data and are based on clinical knowledge that may be connected to COPD readmission. In their study, Wu et al [23] addressed the issue of forecasting readmissions among COPD patients by introducing a fresh scoring system called CORE (COPD – Readmission). This score predicts patient readmissions by taking into account five key predictors: eosinophil count, lung function, triple inhaler therapy, past hospitalization history, and the presence of neuromuscular disease.

III. Dataset

In this chapter, employs an open access Exasens dataset [11-16], which is available in the UCI ML repository for implementing our proposed model. The researchers utilize eight features in this dataset to precisely classify and recognize COPD patients' saliva samples and healthy people [11]. There are 239 samples collected as demographic information for detecting COPD in which dielectric characterizations were performed on 80 samples out of the available 239 samples [9,10] because of the biosensor's limited life-cycle [1-9]. However, in this study, for highlighting the vital function of demographic attributes for detecting COPD, analyses are conducted on 239 samples of this dataset with dielectric properties. In this study, two groups of saliva samples, such as 160 samples for healthy controls and 79 samples for COPD sufferers [17-19], are used for investigating the performance of our proposed ensemble model.

IV. Classifiers

In this study, the following classifiers are considered.

1. Decision Tree: A decision tree is a tree-like structure resembling a flowchart, where internal nodes represent features, branches represent rules, and leaf nodes represent the algorithm's outcomes.

2. Naïve Bayes: Naive Bayes is a probabilistic machine learning algorithm used for classification. It is based on Bayes' theorem and the assumption of feature independence, which is why it's called "naive." Naive Bayes calculates the probability of a given data point belonging to a particular class based on the probabilities of its features or attributes.

It calculates the probability of an input belonging to different classes and assigns it to the most likely class. It's known for its simplicity, speed, and effectiveness in many real-world applications, especially when dealing with text data and high-dimensional feature spaces

3. Support Vector Machine (SVM): A Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm used for classification and regression tasks. It's particularly well-suited for binary classification problems but can be extended to handle multi-class problems as well. The main idea behind SVM is to find a hyperplane that best separates the data points belonging to different classes in a way that maximizes the margin between the classes. The "support vectors" are the data points that are closest to this hyperplane and play a crucial role in defining the decision boundary.

4. Random Forest: Random Forest (RF) is an ensemble learning technique where numerous decision trees are constructed and then combined to achieve a more precise and reliable prediction

5. XGBoost: XGBoost represents an open-source implementation of the gradient boosted trees algorithm. It serves as an efficiently designed, distributed gradient boosting library that prioritizes high performance, adaptability, and versatility. XGBoost encompasses a range of machine learning algorithms within the Gradient Boosting framework.

V. Results and Discussion

Decision tree, Naïve Bayes, SVM, Random forest and XGBoost are considered for the classification purpose. And their performance is analyzed using accuracy, precision, recall and F1 score. Table 1 depicts the performance comparison.

Table 1: *Performance Comparison*

Methods	Accuracy	Precision	Recall	F measure
Decision Tree	75.40	73	74	74.49
Naïve Bayes	50.37	52.12	43.50	47.16
SVM	56.16	47.28	57.45	52.17
Random Forest	78.41	79.25	76.24	78.47
XGBoost	80.47	80.10	79.17	79.85

From table 1 it is observed that out of the considered classifiers, XGBoost gives a better performance in all aspect in compared to the others. XGBoost is an Extreme Gradient Boosting approach which combines the predictions of multiple weak models to produce a stronger prediction. Although Random Forest is a robust and easy-to-use ensemble algorithm, XGBoost often provides better predictive accuracy and faster training. Moreover, in some cases Random Forest is sensitive to outliers in the data, which can result in biased predictions.

On the other hand XGBoost is generally more robust to outliers due to its gradient boosting framework, which can adapt and learn from these data points more effectively. In the similar manner, the performance of DT, Naïve Bayes and SVM can be analyzed as being single classifier, their performance is less as compared to the XGBoost. XGBoost outperforms Decision Trees, Naïve Bayes, and SVMs due to its ensemble-based nature and ability to capture complex patterns in the data.

It can also capture non-linear relationships and interactions between features. Better performance in terms of precision illustrate that XGBoost has an almost accurate classification as the number of false positives are comparatively less. Similarly in terms of recall XGBoost shows that it has an almost accurate classification as the number of false negatives are comparatively less. And thus from the application point of view high value of recall is very much essential for healthcare and medical purpose.

Figure 1 depicts the graphical performance comparison of Decision tree, Naïve Bayes, SVM, Random forest and XGBoost in terms of accuracy, precision, recall and F1 score.

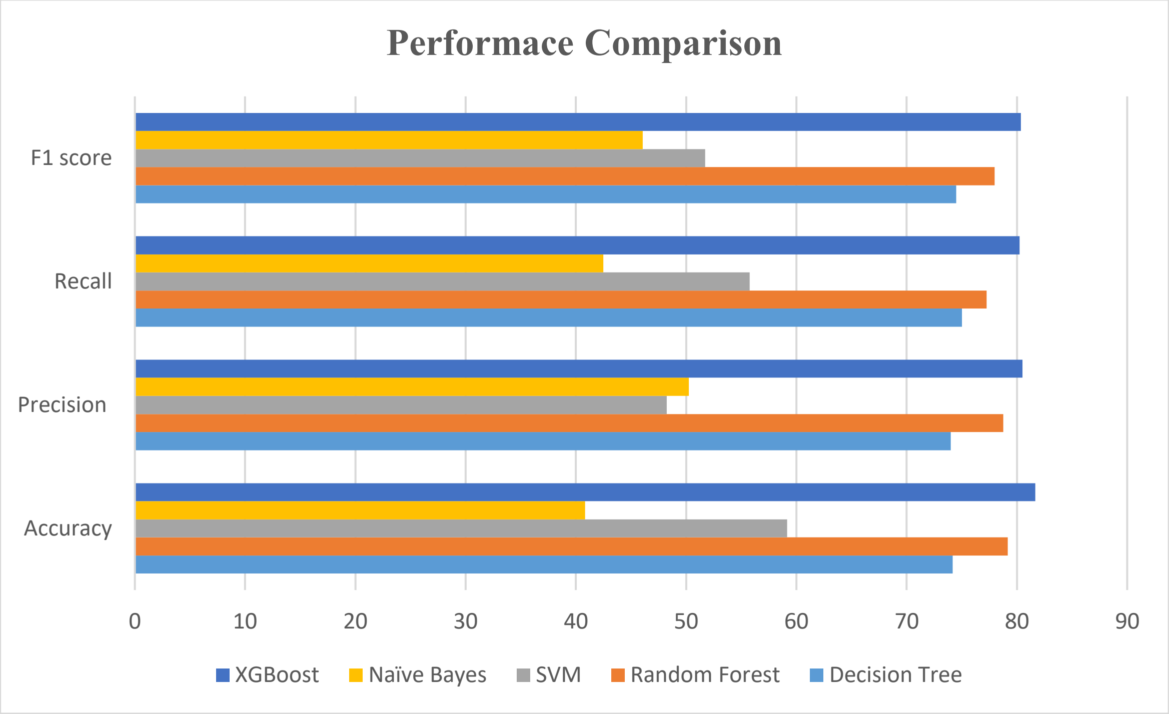

Figure 1: *Graphical performance comparison*

VI. Conclusion

Chronic Obstructive Pulmonary Disease (COPD) is a significant and progressive respiratory disorder impacting millions globally. This study focuses on evaluating various machine learning techniques, including Decision Tree, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and XGBoost, to predict COPD. Their performance is compared based on accuracy, precision, recall, and F1 score.

Experimental results reveal that XGBoost consistently outperforms the other methods, achieving an accuracy of 80.47%. Its superior performance can be attributed to its ensemble learning framework, which effectively captures complex patterns, non-linear relationships, and feature interactions in the data.

The study highlights the importance of classifier performance and emphasizes that data preprocessing steps such as outlier detection, addressing class imbalance, and feature selection can further improve results. Future research could explore these enhancements and leverage larger clinical datasets to achieve even better predictive outcomes.

References

[1] Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380(9859):2095–2128. doi:10.1016/S0140-6736(12) 61728-0

[2] Global Initiative for Chronic Obstructive Lung Disease G. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease. Updated 2013; 2019.

[3] Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006;3(11):e442. doi:10.1371/ journal.pmed.0030442

[4] Chronic obstructive pulmonary disease (COPD). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disea e-(copd). October 27, 2020

[5] Elixhauser, A. et al. Readmissions for chronic obstructive pulmonary disease. Rockville, MD: Agency for Heal. Care Res. Qual. (2011).

[6] Purdy, S., Griffin, T., Salisbury, C. & Sharp, D. Prioritizing ambulatory care sensitive hospital admissions in england for research and intervention: a delphi exercise. Prim. Heal. Care Res. & Dev. 11, 41–50 (2010).

[7] D. Goodman, E. Fisher, and C. Chang, "The Revolving Door: A Report on US Hospital Readmissions," Princeton, NJ Robert Wood Johnson Found., 2013.

[8] P. Jain, Prognostic COPD healthcare management system, no. May. FLORIDA ATLANTIC UNIVERSITY, 2014.

[9] R. Behara, A. Agarwal, F. Fatteh, and B. Furht, "Predicting Hospital Readmission Risk for COPD Using EHR Information," in Handbook of Medical and Healthcare Technologies, Springer, 2013, pp. 297– 308

[10] Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, Williams A, Lynch DA, Make BJ, Crapo JD, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax. 2014;69 (5):415–22

[11] P. Soltani Zarrin, N. Roeckendorf, and C. Wenger, ''In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools,'' IEEE Access, vol. 8, pp. 168053–168060, 2020, doi: 10.1109/ACCESS.2020.3023971.

[12] Q. Wang, H. Wang, L. Wang, and F. Yu, ''Diagnosis of chronic obstructive pulmonary disease based on transfer learning,'' IEEE Access, vol. 8, pp. 47370–47383, 2020, doi: 10.1109/ACCESS.2020.2979218

[13] D. Price, A. Crockett, M. Arne, B. Garbe, R. Jones, A. Kaplan, A. Langhammer, S. Williams, and B. Yawn, ''Spirometry in primary care case-identification, diagnosis and management of COPD,'' Primary Care Respiratory J., vol. 18, no. 3, pp. 216–223, Aug. 2009.

[14] S. Haroon, R. Jordan, Y. Takwoingi, and P. Adab, ''Diagnostic accuracy of screening tests for COPD: A systematic review and meta-analysis,'' BMJ Open, vol. 5, no. 10, Oct. 2015, Art. no. e008133

[15] P. S. Zarrin, F. Zahari, M. K. Mahadevaiah, E. Perez, H. Kohlstedt, and C. Wenger, ''Neuromorphic on-chip recognition of saliva samples of COPD and healthy controls using memristive devices,'' Sci. Rep., vol. 10, no. 1, Dec. 2020, Art. no. 19742, doi: 10.1038/s41598-020-76823-7.

[16] D. Spathis and P. Vlamos, ''Diagnosing asthma and chronic obstructive pulmonary disease with machine learning,'' Health Informat. J., vol. 25, no. 3, pp. 811–827, Sep. 2019, doi: 10.1177/1460458217723169

[17] Y. Fang, H. Wang, L. Wang, R. Di, and Y. Song, ''Diagnosis of COPD based on a knowledge graph and integrated model,'' IEEE Access, vol. 7, pp. 46004–46013, 2019, doi: 10.1109/access.2019.2909069

[18] Dhar J. Multistage Ensemble Learning Model With Weighted Voting and Genetic Algorithm Optimization Strategy for Detecting Chronic Obstructive Pulmonary Disease, in IEEE Access, vol. 9, pp. 48640-48657, 2021, doi: 10.1109/ACCESS.2021.3067949.

[19] Porkodi V., Karuppusamy S. A., Classification of Chronic Obstructive Pulmonary Disease (COPD) using Gabor Filter With SVM Classifier, International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, 2019

[20] Raja B. S. and Babu T. R., A Novel Feature Selection based Ensemble Decision Tree Classification Model for Predicting Severity Level of COPD Disease, Biomedical & Pharmacology Journal, vol 12(2), 875-886, 2019.

[21] Peng, J., Chen, C., Zhou, M. et al. A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators. Sci Rep 10, 3118 (2020). https://doi.org/10.1038/s41598-020-60042-1

[22] Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. Sci Rep. 2019 Feb 20;9(1):2362. doi: 10.1038/s41598-019-39071-y. PMID: 30787351; PMCID: PMC6382784.

[23] Wu YK, Lan CC, Tzeng IS, Wu CW. The COPD-readmission (CORE) score: A novel prediction model for one-year chronic obstructive pulmonary disease readmissions. J Formos Med Assoc. 2021 Mar;120(3):1005-1013. doi: 10.1016/j.jfma.2020.08.043. Epub 2020 Sep 12. PMID: 32928614.