HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

Developing a Novel, Highly Accurate, Chronic Obstructive Pulmonary Disease (COPD) Machine Learning (ML) Model

Bhadresh Amarnath[1] and Usha Soundarapandian[#]

[1]Enloe Magnet High School, Raleigh, North Carolina, USA
[#]Advisor

ABSTRACT

Introduction: Chronic Obstructive Pulmonary Disease (COPD), is a condition caused by damage to the airways or other parts of the lung that blocks airflow and makes it hard to breathe [1]. COPD is the third leading cause of death worldwide, and the seventh leading cause of poor health worldwide [2]. Studies have shown that 20–86% of people with COPD worldwide may be undiagnosed [3]. As there is currently no cure for COPD, early detection is the best option. Current Machine Learning (ML) models focus on using chest images (CT or X-ray scans) to detect COPD; however, the scanning process can be unsafe for patients with COPD [4]. Methods: This study utilized an open data set containing various physical tests of 100 patients with COPD. To train this model, the Random Forest (RF) classifier was used. The accuracy was then plotted on a graph. Results: The Random Forest classifier was able to achieve an accuracy of 92.41% with a perfect recall value of 1.00. This recall value indicates that the Random Forest classifier was able to correctly diagnose all of the patients with COPD in this dataset. Discussion: This study developed a novel ML model that can accurately provide a diagnosis for COPD. Further studies could use this code with a larger dataset to obtain a higher accuracy. White individuals have been reported to have a higher prevalence of COPD [5]. By developing a dataset that accounts for race, we will be able to obtain a more accurate diagnosis.

Introduction

What is COPD

Chronic Obstructive Pulmonary Disease (COPD), is a condition caused by damage to the airways or other parts of the lung that blocks airflow and makes breathing hard. Chronic Bronchitis and Emphysema are the two most common types of COPD [1]. COPD is usually caused by long-term exposure to irritants that damage the lungs and airways. These irritants include cigarette smoking, long-term exposure to dust and chemicals, exposure to fumes from burning fuel, and a genetic disorder that causes an alpha-1-antitrypsin (AAT) deficiency [2].

According to the World Health Organization COPD is the third leading cause of death worldwide, causing 3.23 million deaths in 2019, and the seventh leading cause of poor health worldwide [3]. In the United States, COPD is the sixth leading cause of death in the US, and more than half of those diagnosed are women [4]. It is estimated that COPD cases globally among those aged 25 years and older will increase by 23% from 2020 to 2050, approaching 600 million patients with COPD globally by 2050 [5]. However, even with the high prevalence of COPD globally, studies have shown that 20–86% of people with COPD worldwide may be undiagnosed [6].

Living with Chronic Obstructive Pulmonary Disease (COPD) can be challenging. Many patients with COPD experience persistent shortness of breath, even during simple activities like walking or climbing stairs, which can lead to constant fatigue and a lack of energy. The chronic cough and the need to clear mucus from the lungs are daily

struggles, as well. Emotionally, the disease can cause anxiety and depression due to its chronic nature and the limitations it imposes, often leading to social isolation. However, with the right management strategies, such as regular use of medications, participation in pulmonary rehabilitation programs, and lifestyle adjustments like quitting smoking and adopting a healthy diet, many people find ways to manage their symptoms and improve their quality of life. Support groups and online communities also provide valuable emotional support and practical advice, helping individuals with COPD navigate their daily challenges and find inspiration in the success stories of others who have learned to live well with the disease [2].

Current Diagnosis Methods

The primary test used to diagnose COPD currently is spirometry. It measures the amount of air you can exhale in one second (FEV1) and the total amount of air you can exhale after taking a deep breath (FVC). A ratio of FEV1/FVC less than 0.7 after using a bronchodilator indicates COPD [7-8]. Blood tests may also be used to diagnose COPD such as an arterial blood gas test or complete blood count (CBC). An arterial blood gas test measures the levels of oxygen and carbon dioxide in the blood to assess how well the lungs are functioning, while the CBC test focuses on helping to rule out infections that might be causing symptoms similar to COPD. Chest images such as CT or X-ray scans may also be used to check for any abnormalities in the lungs [9].

Artificial intelligence (AI) models have been developed by researchers, however they rely on using chest images. These models are also only up to 90% accurate [10-11]. Chest scans for individuals with chronic obstructive pulmonary disease (COPD) can have several negative effects. The use of contrast dye during a CT scan may lead to side effects such as itching, swelling, rash, or trouble breathing. Additionally, lung ventilation-perfusion scans (VQ scans) can be affected by the patient's position; for instance, changing from an upright position during the ventilation scan to a supine position during the perfusion scan can cause a mismatch. Other factors that can impact the accuracy of the scan include heart failure, obstructive lung disease, and inadequate distribution of the injectable agent. Furthermore, chest scans involve ionizing radiation, which poses a concern for both patients and healthcare providers. These potential negative effects highlight the importance of careful consideration and monitoring when utilizing chest scans for COPD patients [12].

Methods

Dataset

This study utilized an open data set containing various physical tests. These included the MWT1, MWT2, MWT1Best, FEV1, FEV1PRED, FVC, FVCPRED, CAT, HAD, and SGRQ tests (Figure.1). Other risk factors such as age, gender, smoking, diabetes, hypertension, and heart problems were also accounted for in this dataset. This dataset consisted of 100 patients, with 80 patients being used to train the model, and 20 patients were grouped to test the model.

Test Name	Test Description	Test Name	Test Description
MWT1 (6-Minute Walk Test 1)	This test measures the distance a patient can walk in six minutes. It's used to assess the functional exercise capacity of individuals with COPD	**SGRQ (St. George's Respiratory Questionnaire)**	A tool used to measure health status and quality of life in patients with diseases of airway obstruction, including COPD
MWT2 (6-Minute Walk Test 2)	Similar to MWT1, this is another instance of the 6-minute walk test, often used to compare results over time or after interventions	**AGEquartiles**	This refers to dividing the patient population into quartiles based on age, which helps in analyzing the impact of age on COPD outcomes
MWT1Best	This refers to the best performance recorded during the 6-minute walk tests, indicating the maximum distance walked in six minutes	**Gender**	Gender differences can influence the prevalence, symptoms, and outcomes of COPD
FEV1 (Forced Expiratory Volume in 1 second)	This measures the amount of air a person can forcefully exhale in one second. It's a key indicator of lung function and is used to diagnose and stage COPD	**Smoking**	A major risk factor for developing COPD. Smoking history is crucial in diagnosing and managing the disease
FEV1PRED (Predicted FEV1)	This is the predicted value of FEV1 based on a person's age, gender, height, and race. It helps in comparing an individual's FEV1 to the average values for similar individuals	**Diabetes**	A common comorbidity in COPD patients, which can complicate the management and outcomes of the disease
FVC (Forced Vital Capacity)	This measures the total amount of air exhaled during a forced breath. It's used alongside FEV1 to assess lung function in COPD patients3	**Muscular**	Refers to muscle strength and function, which can be affected in COPD patients due to reduced physical activity and chronic inflammation
FVCPRED (Predicted FVC)	This is the predicted value of FVC, similar to FEV1PRED, and is used for comparison with actual FVC values	**Hypertension**	High blood pressure, another common comorbidity in COPD patients
CAT (COPD Assessment Test)	A questionnaire that measures the impact of COPD on a patient's life. It includes questions about symptoms and daily activities, with scores ranging from 0 to 40	**AtrialFib (Atrial Fibrillation)**	A type of irregular heartbeat that can occur in COPD patients, increasing the risk of stroke and heart failure
HAD (Hospital Anxiety and Depression Scale)	This scale assesses anxiety and depression levels in patients, which are common comorbidities in COPD	**IHD (Ischemic Heart Disease)**	A condition characterized by reduced blood flow to the heart, commonly seen in COPD patients

Figure 1. Table depicting the various tests compiled in the data set

Training and Testing the Model

To train this model, the Random Forest (RF) classifier was used. RF is a machine learning algorithm that combines the results of multiple decision trees to reach a single conclusion, which was perfect, given the various tests used in this dataset. The accuracy was then plotted on a graph.

Figure 2. Schematic Diagram of Methodology

Results

Overview

The Random Forest classifier was able to achieve an accuracy of 92.41%.

Statistical Significance

In this experiment, the confusion matrix was utilized to assess the performance. A confusion matrix is a table that compares the predicted labels by the models to the actual labels, detailing the number of true positives, true negatives, false positives, and false negatives. These metrics are crucial for evaluating the effectiveness of the COPD models in accurately predicting patient outcomes.

This model successfully achieved a perfect recall of 1.00. Recall value indicates that the Random Forest classifier was able to correctly diagnose all of the patients with COPD in this dataset (Figure 3). Additionally, the model was able to achieve an accuracy of 92.41% when randomly choosing COPD patients (Figure 4).

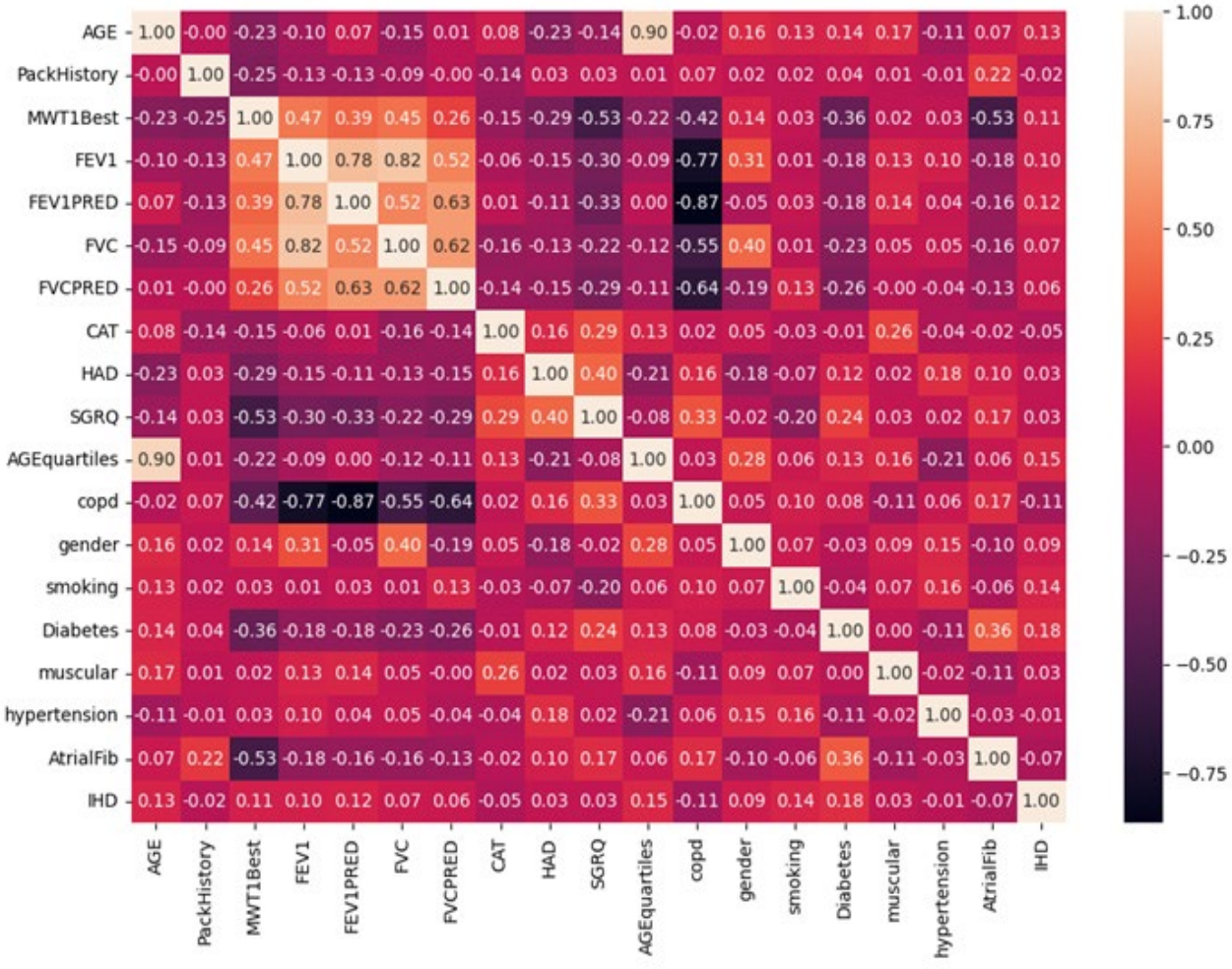

Figure 3. Confusion Matrix

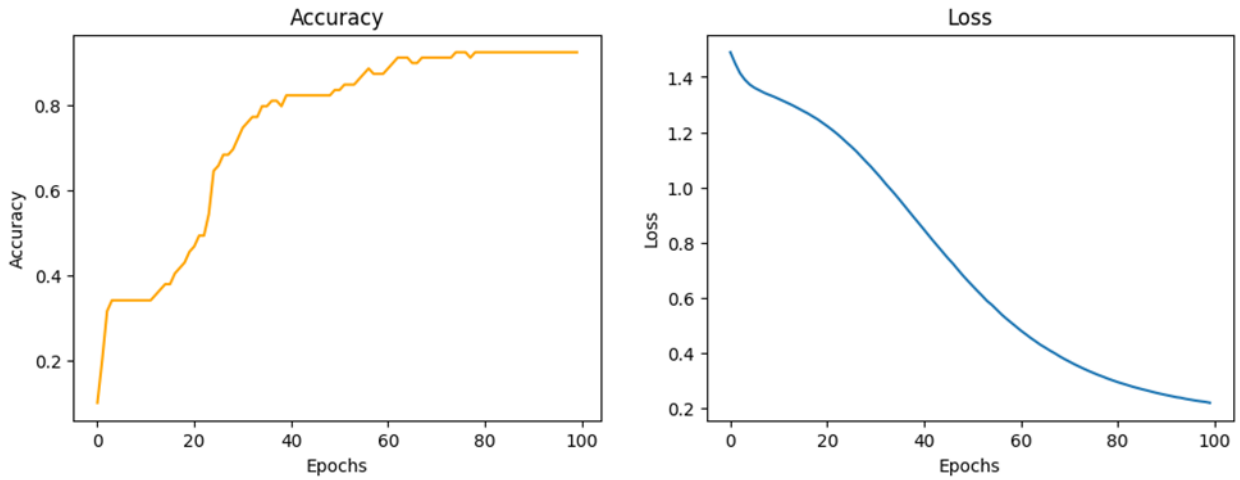

Figure 4. Graphs of accuracy and prediction

Discussion

Through this work, this study presents a novel Machine Learning (ML) model that can accurately diagnose COPD patients. Previous AI models have primarily focused on using advanced imaging techniques, such as CT scans, to identify and stage COPD. For instance, deep learning models have been developed to analyze radiomics features from CT images, achieving high diagnostic accuracy by integrating multiple data sources [13]. Another approach utilized natural language processing (NLP) to predict COPD exacerbations from clinical notes and vital signs [14]. These methods, while effective, often require expensive and sophisticated equipment, limiting their accessibility in many clinical settings.

In contrast, this study emphasizes the importance of using various physical tests, including MWT1, MWT2, MWT1Best, FEV1, FEV1PRED, FVC, FVCPRED, CAT, HAD, and SGRQ tests. These tests, along with other risk factors such as age, gender, smoking, diabetes, hypertension, and heart problems, provide a comprehensive dataset that is both cost-effective and practical for widespread use. Physical tests like these are generally less expensive and more accessible than spirometry, making them a viable option for early COPD diagnosis in resource-limited settings [15,16].

The inclusion of these diverse tests is crucial because they capture a wide range of physiological and clinical parameters, offering a holistic view of the patient's health status. For example, the FEV1 and FVC tests measure lung function, while the CAT and HAD tests assess the impact of COPD on the patient's quality of life and mental health. By accounting for these variables, the model can provide a more accurate and personalized diagnosis, improving patient outcomes.

Moreover, the cost-effectiveness of using physical tests over spirometry is significant. Spirometry, while considered the gold standard for COPD diagnosis, can be costly and requires specialized equipment and trained personnel [17]. In contrast, physical tests can be administered more easily and at a lower cost, making them a practical alternative for large-scale screening and diagnosis.

Conclusion

In conclusion, this study demonstrates the potential of using a comprehensive set of physical tests and risk factors to develop an accurate and cost-effective ML model for COPD diagnosis. By leveraging accessible and affordable diagnostic tools, this approach can facilitate early detection and management of COPD, particularly in settings where

resources are limited. Future research should continue to explore and validate these methods in diverse populations and clinical environments to enhance their generalizability and impact.

Limitations

Firstly, this study used a small dataset containing only 100 patients. With a small dataset, the statistical power is limited, which may affect the generalizability of the findings. Additionally, the sample size may not adequately represent the broader population, potentially introducing bias. Future studies should aim to include a larger and more diverse sample to validate these results. As a result, the model's performance may be influenced by the quality and completeness of the data. Any missing or inaccurate data can lead to suboptimal model training and predictions. Ensuring high-quality data collection and preprocessing is crucial for improving model accuracy. The model may not perform as well in different settings or with different populations. External validation in varied environments is necessary to confirm the model's robustness and adaptability.

Secondly, White individuals have been reported to have a higher prevalence of COPD [18]. By developing a dataset that accounts for race, we will be able to obtain a more accurate diagnosis. Including diverse racial and ethnic groups in the dataset will help ensure that the model is more representative and can provide reliable predictions across different populations.

Acknowledgments

I would like to thank my mentor for the valuable insight provided to me on this topic.

References

1. COPD - What Is COPD? | NHLBI, NIH. (2023, October 25). Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/copd#:~:text=COPD%2C%20or%20chronic%20obstructive%20pulmonary
2. COPD - Symptoms and causes. (n.d.). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679#:~:text=In%20the%20vast%20majority%20of
3. World. (2023, March 16). Chronic obstructive pulmonary disease (COPD). Who.int; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)#:~:text=16%20March%202023
4. National Heart, Lung, and Blood Institute. (2022, March 24). COPD - What Is COPD? | NHLBI, NIH. Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/copd#:~:text=In%20the%20United%20States%2C%20COPD
5. Boers, E., Barrett, M., Su, J. G., Benjafield, A. V., Sinha, S., Kaye, L., Zar, H. J., Vuong, V., Tellez, D., Gondalia, R., Rice, M. B., Nunez, C. M., Wedzicha, J. A., & Malhotra, A. (2023). Global Burden of Chronic Obstructive Pulmonary Disease Through 2050. JAMA Network Open, 6(12), e2346598. https://doi.org/10.1001/jamanetworkopen.2023.46598
6. Johnson, K. M., Bryan, S., Ghanbarian, S., Sin, D. D., & Sadatsafavi, M. (2018). Characterizing undiagnosed chronic obstructive pulmonary disease: a systematic review and meta-analysis. Respiratory Research, 19(1). https://doi.org/10.1186/s12931-018-0731-1
7. Stephen D. Cagle, J., Landrum, L. S., & Kennedy, A. M. (2023). Chronic Obstructive Pulmonary Disease: Diagnosis and Management. American Family Physician, 107(6), 604–612. https://www.aafp.org/pubs/afp/issues/2023/0600/chronic-obstructive-pulmonary-disease.html

8. Diagnosing COPD. (2019, July 8). National Institutes of Health (NIH). https://www.nih.gov/news-events/nih-research-matters/diagnosing-copd
9. MSN. (n.d.). Www.msn.com. https://www.msn.com/en-us/health/condition/Chronic-obstructive-pulmonary-disease/hp-Chronic-obstructive-pulmonary-disease?source=conditioncdx
10. Bian, H., Zhu, S., Zhang, Y., Fei, Q., Peng, X., Jin, Z., Zhou, T., & Zhao, H. (2024). Artificial Intelligence in Chronic Obstructive Pulmonary Disease: Research Status, Trends, and Future Directions --A Bibliometric Analysis from 2009 to 2023. International Journal of Chronic Obstructive Pulmonary Disease, 19, 1849–1864. https://doi.org/10.2147/COPD.S474402
11. Wu, Y., Xia, S., Liang, Z., Chen, R., & Qi, S. (2024). Artificial intelligence in COPD CT images: identification, staging, and quantitation. Respiratory Research, 25(1). https://doi.org/10.1186/s12931-024-02913-z
12. Washko, G. (2010). Diagnostic imaging in COPD. Seminars in Respiratory and Critical Care Medicine, 31(03), 276–285. https://doi.org/10.1055/s-0030-1254068
13. Zhu, Z., Zhao, S., Li, J., Wang, Y., Xu, L., Jia, Y., Li, Z., Li, W., Chen, G., & Wu, X. (2024). Development and application of a deep learning-based comprehensive early diagnostic model for chronic obstructive pulmonary disease. Respiratory Research, 25(1). https://doi.org/10.1186/s12931-024-02793-3
14. Robertson, N. M. (n.d.). Integrating Artificial Intelligence in the Diagnosis of COPD Globally: A Way Forward. Chronic Obstructive Pulmonary Diseases:Journal of the COPD Foundation, 11(1), 114–120. https://journal.copdfoundation.org/jcopdf/id/1456/Integrating-Artificial-Intelligence-in-the-Diagnosis-of-COPD-Globally-A-Way-Forward
15. Roland, J. (2018, November 7). COPD Tests and Diagnosis. Healthline; Healthline Media. https://www.healthline.com/health/copd/tests-diagnosis
16. American Lung Association. (2023). How COPD Is Diagnosed. Www.lung.org. https://www.lung.org/lung-health-diseases/lung-disease-lookup/copd/symptoms-diagnosis/diagnosing
17. Qu, S., You, X., Liu, T., Wang, L., Yin, Z., Liu, Y., Ye, C., Yang, T., Huang, M., Li, H., Fang, L., & Zheng, J. (2021). Cost-effectiveness analysis of COPD screening programs in primary care for high-risk patients in China. Npj Primary Care Respiratory Medicine, 31(1). https://doi.org/10.1038/s41533-021-00233-z
18. Liu, Y. (2023). Trends in the prevalence of Chronic Obstructive Pulmonary Disease among adults aged ≥18 years — United States, 2011–2021. MMWR. Morbidity and Mortality Weekly Report, 72(46). https://doi.org/10.15585/mmwr.mm7246a1