IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

Chronic Obstructive Pulmonary Disease Prediction Using Machine Learning

Mr Ch. Papa Rao[1], K. Anuroopa Visalakshi[2], G. Naga Sai[3], Ch. Anji Reddy[4], K. Pavan Kumar[5]

Associate Professor, Department of CSE, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh, India[1]

B. Tech, Department of CSE, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh, India[2,3,4,5]

ABSTRACT: Chronic obstructive pulmonary disease is defined by the WHO(World Health 0rganization) as "common, preventable and treatable disease that is characterized by long term respiratory symptoms and air flow restrictions that is due to airway abnormality caused by exposure to harmful particles or gases(A group of lung diseases that blocked airflow and make it difficult to breath) IN the recent years, there are an increase in the death rate due to chronic obstructive pulmonary disease(COPD) Patients and it is estimated that it will increase in the coming years also Traditional methods take a longtime to identify this disease because a lot clinical test has to be perfumed for getting the conformation, however with the advent of intelligent techniques like AI ML as well as looking the potential of these techniques for predicting other critical diseases, it is believed that it would help to detect the chronic diseases at an early age in precise manner.

KEYWORDS: COPD, Random Forest, Respiratory Symptoms, chronic bronchitis.

I.INTRODUCTION

The COPD project encompasses a comprehensive effort to improve the understanding, diagnosis, and management of Chronic Obstructive Pulmonary Disease (COPD). This multifaceted initiative integrates various research and clinical strategies aimed at addressing the complexities of COPD, from its ethology to its impact on patients' lives. Key components of the project include investigating the role of genetic predisposition, environmental factors, and smoking cessation interventions in COPD development and progression. Additionally, the project focuses on enhancing diagnostic methods through advanced imaging techniques and biomarker discovery, facilitating earlier detection and intervention. Treatment optimization is another crucial aspect, involving personalized medication regimens, pulmonary rehabilitation programs, and novel therapies targeting specific disease mechanisms. Furthermore, the project emphasizes patient education and self-management strategies to empower individuals with COPD to actively participate in their care and improve their quality of life. By integrating cutting-edge research, clinical expertise, and patient-centred approaches, the COPD project aims to make significant strides in reducing the burden of this debilitating respiratory condition. The percentage of the affected people with chronic obstructive pulmonary disease is increasing compared to the past year, and it is expected that it will increase at a higher rate because the growth rate of our population is increasing day by day. Anyone who has struggled to breathe knows how important it is even for a short time. The World Health Organization (WHO) has said 64% of million people suffer from chronic obstructive pulmonary disease (COPD). Every year more than three million people die from chronic obstructive pulmonary disease (COPD). The World Health Organization estimates that by 2030 it will be the 3rd main cause of death worldwide and is expected to exceed 30% in the next 10 years. Chronic obstructive pulmonary disease (COPD) is a lung disease. This is chronic obstruction of the airways in the lungs that interferes with normal breathing and but cannot be reversed. The terms 'chronic bronchitis' and 'emphysema' are no longer used and are now included in the diagnosis of COPD. This is not a single disease. An application used to describe chronic lung disease. If the most common are chronic obstructive pulmonary disease, asthma, pulmonary hypertension, and occupational lung diseases.

II.LITERATURE REVIEW

[01]A Machine Learning-Based Predictive Model for 30 Day Hospital Readmission Prediction for COPD-Patients, Previous research in the development of machine learning-based predictive models for 30- day hospital readmission prediction in COPD patients has demonstrated promising results. Several studies have focused on leveraging electronic health record (EHR) data to identify factors associated with COPD exacerbations and subsequent hospital readmissions.[02] In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

Learning Tools, Previous research in the classification of saliva samples from COPD patients and healthy controls using machine learning tools has shown promise in identifying biomarkers and distinguishing between different disease states. Several studies have investigated the potential of saliva-based diagnostics for COPD, recognizing its non-invasive nature and potential for early disease detection.[03] Detection of different stages of COPD patients using machine learning techniques, Validation of these classification models involves assessing their performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUCROC). Cross-validation techniques and external validation cohorts are commonly utilized to ensure the robustness and generalizability of the developed models.

[04] Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population, -Studies have typically involved the analysis of diverse datasets containing clinical, demographic, and environmental factors relevant to COPD development. Feature selection and engineering play crucial roles in model development, with researchers identifying the most informative predictors of COPD risk within the Chinese population. These predictors may include smoking history, age, gender, occupational exposures, genetic variants, respiratory symptoms, and comorbidities.[05] Developing an ml pipeline for asthma and COPD: - The case of a Dutch primary care service, The overall goal of developing such an ML pipeline is to provide a tool that enhances the precision of asthma and COPD management in the Dutch healthcare system, improving patient outcomes through tailored treatment plans and preventive measures based on individual risk profiles and disease characteristics. This initiative not only addresses the clinical needs but also aims to reduce healthcare costs through improved disease management efficiency.[06] Machine learning Algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease, Previous work on utilizing machine learning algorithms alongside forced oscillation technique (FOT) measurements for the automatic identification of Chronic Obstructive Pulmonary Disease (COPD) represents a significant advancement in non-invasive respiratory diagnostics. The forced oscillation technique, which measures respiratory mechanics using external oscillations applied to the respiratory system, offers detailed insights into lung function that are not captured by conventional spirometry.

III.SUMMARY OF LITERATURE REVIEW

Studying all of the aforementioned publications revealed gaps in knowledge, limits in current methods, and lower accuracy when compared to one another. Their obtained results show a lack of accuracy. Furthermore, some work encounters difficulties with regional variability due to various disease prediction algorithms. The dataset may have an impact on the accuracy. Support vector machines and Random Forest are employed in this work. The accuracy of the two methods differs significantly. The parameters that the study works inside are specified by the delimitations. For example, the delimitations in a study on Random Forest-based COPD prediction might comprise it's frequently necessary to combine domain expertise, data preprocessing methods, model tuning, and continual assessment and improvement of the predictive system to address these delimitations.

IV.PROPOSING METHODS

Employing machine learning (ML) techniques holds immense promise in revolutionizing the management of Chronic Obstructive Pulmonary Disease (COPD). By leveraging diverse datasets encompassing clinical records, imaging, biomarkers, and environmental factors, ML algorithms can facilitate early detection, predict disease progression, and optimize treatment strategies. Through sophisticated data preprocessing, feature engineering, and model selection, this approach aims to extract meaningful insights from complex data, enabling personalized interventions tailored to individual patient needs.

4.1 Methodology

1.Data Collection: Gather a comprehensive dataset comprising clinical data of patients, including demographics, smoking history, lung function tests (spirometry), medical history, medication usage, and other relevant variables. You can obtain this data from hospitals, clinics, or research databases.

2.Data Preprocessing: Impute missing data using techniques like mean imputation or predictive imputation. Scale numerical features to a similar range to prevent any particular feature from dominating the model training. Identify and select relevant features using techniques like correlation analysis or feature importance ranking.

3.Feature Engineering: Extract additional features that may provide valuable information for COPD prediction, such as derived lung function parameters, smoking pack-years, or comorbidity indices.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

4.Model Selection: Experiment with various machine learning algorithms suitable for classification tasks, such as
- XGBOOST
- Random Forest
- Support Vector Machines (SVM)
- Gradient Boosting Machines (GBM)

5.Model Training and Evaluation: Split the dataset into training and testing sets (e.g., 70% training, 30% testing). A Train each model on the training set and fine-tune hyperparameters using techniques like cross-validation. Evaluate model performance on the test set and compare results across different algorithms.

6. Validation: Validate the models using an independent dataset to assess their generalizability and robustness.

4.2 Algorithms

Classification algorithms and regression algorithms are the two primary categories of algorithm models used in machine learning.

1.Random Forest: The forecasts from many decision trees are combined using an ensemble learning method known as a random forest to get a single, more accurate forecast. Both regression and classification tasks can benefit from the application of this kind of supervised learning technique.

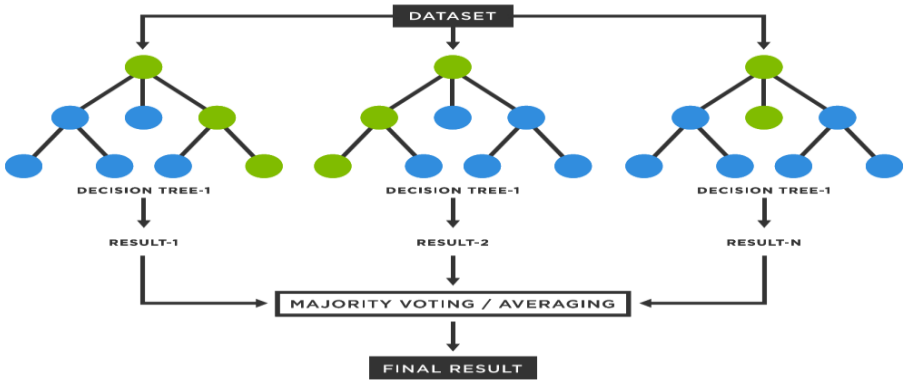

FIG: 4.2 – RANDOM FOREST ALGORITHM

2.Support Vector Machine: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

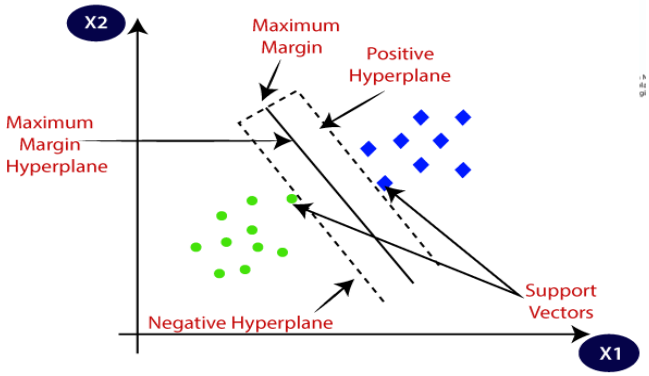

FIG :4.3 – SVM ALGORITHM

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

3.XGBOOST: Efficiency and scalability are given priority in the design of XGBOOST, a toolset for training machine learning models using distributed gradient boosting. Through the process of ensemble learning, multiple weak models' predictions are combined to produce a more robust forecast.

4.Light GBM: Light GBM uses a leaf-wise tree growth strategy instead of level-wise (depth-first) like other boosting algorithms. In leaf-wise growth, the algorithm splits nodes vertically (choosing the best split based on the maximum reduction in loss) rather than horizontally, which tends to lead to deeper trees. This approach can result in faster training times but may be prone to overfitting if not carefully controlled.

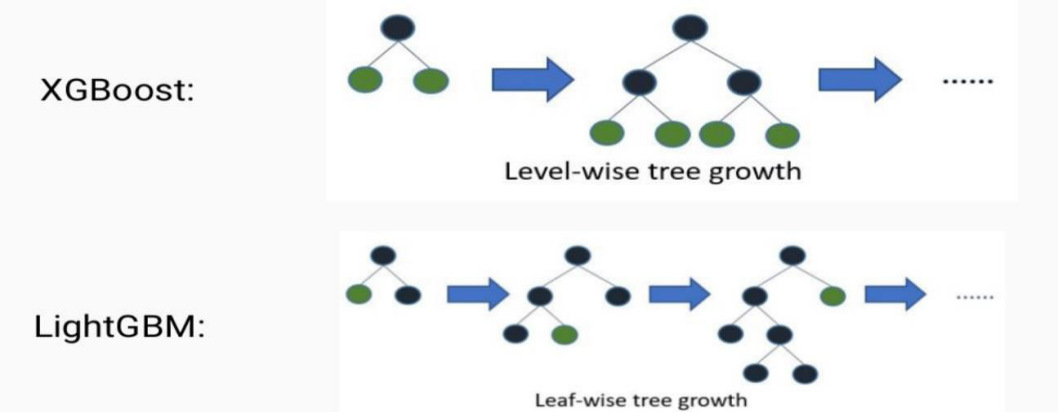

FIG: 4.4 – XGBOOST ALGORITHM&LIGHT GBM ALGORITHM

V.RESULT

```
[37]: #SVM Algorithm
svc = SVC(kernel = 'linear')
svc_model = svc.fit(X_train_scaled,y_train)
svc_pred = svc_model.predict(X_test_scaled)
print(f'Support Vector Classifier  : {accuracy_score(svc_pred,y_test)*100}%')

Support Vector Classifier  : 81.82608695652173%
```

Fig5.1: Support vector machine accuracy

```
[42]: from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, classification_report
xgb_model = XGBClassifier()
xgb_model.fit(X_train_scaled, y_train)
xgb_pred = xgb_model.predict(X_test_scaled)
print(f'Accuracy of XGBClassifier  : {accuracy_score(xgb_pred,y_test)*100}%')

Accuracy of XGBClassifier  : 82.52173913043478%
```

Fig5.2: XGBOOST accuracy

```
[23]: #Random Forest Algorithm
rfc = RandomForestClassifier(n_estimators=500,random_state=42,)
rfc_model = rfc.fit(X_train_scaled,y_train)
rfc_pred = rfc_model.predict(X_test_scaled)
print(f'Accuracy of Random Forest Classifier  : {accuracy_score(rfc_pred,y_test)*100}%')

Accuracy of Random Forest Classifier  : 83.1304347826087%
```

Fig5.3: Random Forest accuracy

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

```
[46]: x=[0,1,0.0,0,0,0,0,0,0,0,2.4766,0.77,2.921,3.805,2.622000]
      x=np.array(x)
      x=x.reshape(1,-1)
      r=rfc_model.predict(x)
      c[r[0]]

[46]: 'COPD not present'
```

Fig5.4: COPD present or not present

```
[50]: x=[1,1,0.0,1,2,1,2,2,1,1,-1.0000,0.51,1.906,3.732,2.002000]
      x=np.array(x)
      x=x.reshape(1,-1)
      r=rfc_model.predict(x)
      c[r[0]]

[50]: 'COPD present'
```

Fig5.5: COPD present or not present

In this section, we compared the efficacy of the previously discussed strategies. Table 5.4,5.5 displays a comparison of the calculated accuracy for the "COPD present" and "COPD not present" classes. We found that the Random Forest classifier performed better in forecasting COPD than the other machine learning methods, with an accuracy of 83.13%.

IV.CONCLUSION

In conclusion, our study demonstrates the potential of machine learning algorithms in predicting COPD risk with high accuracy and reliability. By leveraging a comprehensive dataset encompassing demographic, clinical, and environmental factors, we developed robust predictive models capable of identifying individuals at risk of developing COPD. The performance metrics of our models indicate their effectiveness in clinical settings, offering healthcare providers a valuable tool for early detection and intervention. Furthermore, our analysis highlights the significance of specific risk factors such as smoking history, air pollution exposure, and respiratory symptoms in COPD prediction. By incorporating these factors into predictive models, we can enhance the precision of risk assessment and facilitate targeted preventive strategies.

VII.FUTURE SCOPE

"In future directions for COPD prediction using machine learning, the emphasis on algorithm accuracy stands out as a pivotal consideration. While advancements in feature selection techniques, ensemble learning methods, deep learning architectures, and validation on external datasets all hold promise, their ultimate impact hinges on their ability to improve the accuracy of predictive models. Refined feature selection methods aim to identify the most informative predictors, while ensemble learning techniques leverage diverse models to enhance predictive accuracy**".**

REFERENCES

[1] Tina Shah, Valerie G. Press, Megan Huisingh-Scheetz, and Steven R. White, "COPD Readmissions Addressing COPD in the Era of Value-based Health Care," Recent Advances in Chest Medicine, PP.916- 1926, 2016.
[2] Anna Spathis, and Sara Booth, "End of life care in chronic obstructive pulmonary disease: in search of a good death," Int J Chron Obstruct Pulmon Dis., vol. 3(1), 2008, pp.11–29, doi: 10.2147/copd.s698.
[3] Yosef Berlyand, et al., "How artificial intelligence could transform emergency department operations," Am. J of Em. Med., pp.1515-1517, 2018.
[4] Tadahiro Goto, C. A. Camargo Jr., M. K. Faridi, B. J. Yun, K. Hasegawa, "Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED," American Journal of Emergency Medicine, pp. 1650-1654, 2018.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305307 |

[5] B , et al., "A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem," J of Community Hospital Internal Medicine Perspectives, vol.2(1), 2012.
[6] Wen-Yen Lin, Vijay Kumar Verma, Ming-Yih Lee and Chao-Sung Lai, "Activity Monitoring with a Wrist-Worn, Accelerometer-Based Device," Micromachines 2018.
[7] LM. Osman, DJ Godden, JA Friend, et al., "Quality of life and hospital re-admission in patients with chronic obstructive pulmonary disease," Thorax BM Journal, pp. 67-71, 1997.
[8] R. M. Effros, B. Peterson, R. Casaburi, J. Su, M. Dunning, J. Torday, J. Biller, and R. Shaker, Epithelial lining uid solute concentrations in chronic obstructive lung disease patients and normal subjects, J. Appl. Physiol., vol. 99, no. 4, pp. 12861292, Oct. 2005.
[9] P.SoltaniZarrin,F.Jamal,S.Guha,J.Wessel,D.Kissinger,andC. Wenger, Design and fabrication of a BiCMOS dielectric sensor for viscosity mea surements: A possible solution for early detection of COPD, Biosensors, vol. 8, no. 3, p. 78, Aug. 2018.
[10] S.Aryal, E.Diaz-Guzman, and D.M.Mannino, COPD and genderdifferences: An update, Transl. Res., vol. 162, no. 4, pp. 208218, Oct. 2013.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details