Predictive Modelling of $PM_{2.5}$ Impact on Pulmonary Function Among Indian Textile Workers Diagnosed with Chronic Obstructive Pulmonary Disease: A Machine Learning-Based Study

Shankar S[1*], Abbas G[2], Chander Prakash[3], Manikanadan M[4], Sathishkumar V E[5]
[1]Department of Mechanical Engineering, Karpagam College of Engineering, Coimbatore, TamilNadu, INDIA
[2]Department of Mechanical Engineering, Karpagam Institute of Technology, Coimbatore, TamilNadu, INDIA
[3]Vice Chancellor, Sanskaram University, Patauda, Jhajjar, Haryana, 124108, INDIA
[4] Department of Mechanical Engineering, Nandha Engineering College, Erode, TamilNadu, INDIA
[5]School of Engineering and Technology, Sunway University, Bandar Sunway, MALAYSIA
*Corresponding Author Email: shankariitm@gmail.com
Received: 27.08.2025 *Revised: 09.09.2025* *Accepted: 01.10.2025* *Published: 12.10.2025*

Abstract

The work addresses the significant health risks posed by respiratory issues like chronic obstructive pulmonary diseases (COPD) among textile industry workers. The study assessed the performance of various machine learning algorithms, including Support Vector Machines (SVM), Decision Trees, Random Forest, and Linear Regression, in forecasting $PM_{2.5}$ concentrations and their impact on pulmonary function indices (FVC, FEV_1, and PEFR) among textile workers with COPD. Performance results indicated that Decision Tree and Random Forest models exhibited superior performance in predicting $PM_{2.5}$ concentrations, with high accuracy, precision, recall, and F1 scores (99.9%). Additionally, higher $PM_{2.5}$ levels were associated with decreased pulmonary function, with Decision Tree model showing the highest R^2 value (0.972) and lowest RMSE (0.458) and MAE (0.573), emphasizing its robust performance in capturing $PM_{2.5}$ variations. The use of machine learning algorithms and real-time IAQ monitoring data holds promise for improving the early detection and management of respiratory exacerbations among textile workers.

Keywords: Indoor Air Quality; Machine Learning; Pulmonary function; $PM_{2.5}$; Textile Industry

Introduction

The $PM_{2.5}$ with diameters ≤2.5 μm and especially those less than this diameter is a significant health hazard, especially in the industrial environment like textile manufacturing [1]. In developing nations, such as India, textile workers are heavily impoverished with regard to their exposure to $PM_{2.5}$ because of the lack of proper ventilation, activities that cause high dust generation, and regulation. The long-term exposure of $PM_{2.5}$ to the respiratory system is closely linked to long-term respiratory diseases, the foremost being Chronic Obstructive Pulmonary Disease (COPD)- a non-reversible airflow obstruction disorder, which is progressive and associated with a decrease in lung functions [2]. The third cause of death across the world is respiratory illnesses like chronic obstructive pulmonary diseases (COPD) that is typified by persistent inflammatory hindrance of the airways and emphysema that affects the health of those who work in the textile industry and their productivity [3]. The conditions constitute a significant part of the global chronic disease burden [4]. Asthma and COPD exacerbations may cause life-threatening interventions and hospitalization, which impose an extra burden on the health care system and impact the livelihood of workers [5]. Asthma and COPD among textile workers are found to be of a concerning prevalence in the United States, which is something highlighted by the statistics of the Center for Disease Control (CDC) and which necessitates specific interventions [6]. Likewise, respiratory diseases among employees have been on the rise in other countries like India, where the textile industry constitutes a major sector [7]. Asthma and COPD are the main problems that many textile workers experience, which points to the need to tackle this problem [8]. Hospital admission rates for respiratory conditions among textile workers in India surpass the OECD average, indicating a pressing public health challenge [9].

It has been shown that the indoor air quality of textile factories has a critical impact on the outcomes of respiratory health in workers [9]. Airborne particulate matter and volatile substances emitted during textile manufacture processes are the factors that lead to asthma and COPD exacerbation [9, 10]. Nevertheless, the research on the relationship between indoor air pollution and respiratory morbidity has been increasing, although, the studies on individual health risk prediction based on real-time monitoring data of indoor air quality among textile workers is still lacking [11, 12]. Recent research usually happens sporadically or monthly but does not involve long-term continuous real-time monitoring to implement preventive measures [13]. The application of specific predictive modeling methods to the textile industry will provide priceless information on how to recognize those workers who are more likely to experience respiratory exacerbations [14]. Using powerful machine learning algorithms and clinical data combined with real-time indoor air quality at work, predictive models can provide in-time warnings to textile workers vulnerable to asthma and COPD exacerbations [15] but also HYSPLIT model is used to determine the routes of the pollutant [16]. This proactive

approach enables timely intervention strategies, including medication management and adjustments to work schedules, to mitigate the impact of respiratory conditions on worker's health and productivity. In recent years, artificial health intelligence, particularly rooted in Machine learning, has revolutionized predictive capabilities in healthcare, offering profound benefits for disease prognosis [17] and deep learning [18]. There has been a burgeoning interest in leveraging artificial health intelligence, particularly machine learning, for predictive analytics in healthcare [19]. Machine learning is characterized by strong learning abilities, effectiveness in dealing with time series and longitudinal data, and the ability to deal with irregularities in data. Although machine learning has been widely used in different fields, the use of this technique in the prediction of asthma and chronic obstructive pulmonary diseases (COPD) has been comparatively low because of difficulties in acquiring real-time lung functionality data [20]. However, in contrast to other health apps, machine learning models can take advantage of real-time risk monitoring information on personal and environmental sensors using the Internet of Things (IoT), whereas the deployment of this information to detect respiratory diseases is challenging [21]. The machine learning algorithms that have been used to predict respiratory risks have depended on past records of patients in clinical environments [22]. Nevertheless, recent developments have demonstrated potential opportunities in combining machine learning with real-time monitoring data to make predictions about the individual risk of developing asthma, especially with the use of such factors as air pollution, weather, and lung function monitoring that are continuous[23]. Respiratory disease prediction has also discussed the machine learning models like Support Vector Machine (SVM), Decision Trees, Random Forest, and Linear Regression. SVM is especially useful when working with nonlinear correlations and has been utilized in other medical predictions problems [24]. Decision Trees can be interpreted and can also identify multifactorial interactions between variables [25]. Random Forest is an ensemble learning technique where dozens of decision trees are employed to increase their prediction abilities and power [26]. Linear Regression is a less complex process though it is also capable of providing useful findings to the relationship which exists between the predictor variables and the respiratory health outcomes [27]. Amaral et al., came up with a clinical support system that would help diagnose chronic obstructive pulmonary disease (COPD) using forced oscillation (FO) measures. They used machine learning algorithms i.e. K-nearest neighbors (KNN) and Support Vector Machines (SVM) and Artificial Neural Networks (ANN) which had shown better performance in their research [28]. Morillo et al., carried out an in-depth analysis of the available algorithms and their effectiveness regarding the detection of exacerbations and decision-making on treatment of COPD patients during home telemonitoring interventions. They evaluated different algorithms to determine the appropriateness of each of them [29]. Spathis et al. are interested in clinical decision support systems within the healthcare sector, especially the area of the prevention, diagnosis, and treatment of respiratory conditions, such as asthma and COPD. They discovered that the highest value of the precision of the selected classifier, the Random Forest was 97.70% [22]. A new hybrid $PM_{2.5}$ forecasting model was suggested, which is called CEEMDAN-PE-GWO-VMD-MIF-BiLSTM-AT [30]. The level of $PM_{2.5}$ was measured in various locations and health risks linked to them were determined [31, 32].

Granero et al., used a machine learning method to predict exacerbations of Chronic Obstructive Pulmonary Disease (COPD). They propose that the suggested methodology based on data can be used to develop reliable predictive algorithms of COPD exacerbations. These algorithms may be beneficial both to clinicians and patients and provide assistance when it comes to the active management and prevention of COPD exacerbation [33]. Himes et al., in a study aimed to predict the prevalence of chronic obstructive pulmonary disease (COPD) in patient asthma through the use of electronic medical record. Such models are potentially useful in clinical practice and can be improved with further approaches to data extraction and the consideration of other related variables. The research used COPD symptoms such as age, gender, and smoking history as predictive variables, which indicates possible use of full patient details to increase predictive and clinical relevance [34]. Pulmonary function test FVC (Forced Vital Capacity), FEV_1 and PEFR Pulmonary functions test is an easy-to-use and affordable test to determine airway obstruction in people with asthma and COPD due to its portability and easily accessible nature [7]. Although past study had used weekly spirometry measurements as predictors of asthma attacks in textile workers at risk of exposure to cotton dust with the use of machine learning models, there is a scant in research to use daily spirometry measurements and relate them to time-varying indoor air quality and weather to predict respiratory degradation. Although epidemiological evidence links $PM_{2.5}$ exposure to COPD development and progression, limited efforts have been made to quantitatively model and predict the degree of pulmonary function decline among exposed individuals. In particular, machine learning (ML) techniques offer powerful tools for predicting health outcomes by capturing nonlinear relationships among multiple predictors, yet they remain underutilized in occupational pulmonary research within the Indian textile sector. This study aims to develop and evaluate predictive models using machine learning algorithms to estimate pulmonary function decline (measured via FEV_1 and FVC) in textile workers diagnosed with COPD. By incorporating real-time $PM_{2.5}$ exposure and individual-level clinical-demographic variables, this research seeks to provide a data-driven framework for targeted health interventions and surveillance. With machine learning abilities and the use of real time indoor PM and weather information, our model aims to provide workers in the

textile industry (20-60 years old) in the Erode district of Tamil Nadu, India with timely information on the respiratory health of those in the industry. The model will give an idea of how the environmental factors can affect the respiratory health outcomes of the textile workers and will be useful in formulating specific intervention measures.

2. Methodology

2.1 Study Design

This research collected data of 187 textile industry workers who were diagnosed of having chronic obstructive pulmonary disease (COPD) upon having gone through spirometry measurements. Based on the recommendations of the Global Initiative of Asthma (GINA), all the subjects showed mild to severe lung impairment, which triggered the special attention toward those with a range of PEFR of 100-500 liters per minute [35]. Study exclusion criteria included the exclusion of those with such a history of thoracic surgery, asthma, heart disease, chronic respiratory diseases, and smokers. In addition, the research gathered extensive information with regard to the demographic, social and economic status of the employees. To consider the severe seasonal fluctuations, Indoor Air Quality (IAQ) and lung function measurements were performed in the same season. This modification was to reduce the possibility of confounding factors and also to provide the accuracy and reliability with the research finding on the correlation between IAQ and lung functions of the workers in the textile industry who are COPD victims. This study involved workers aged between 30 and 65 years of age with at least 5 years of experience in textile operations who had COPD diagnosis. Patients with asthma, pulmonary TB and other chronic respiratory conditions were excluded.

2.2 Data Collection and Research Framework

A portable spirometer made in Italy, Medical International Research was used to determine pulmonary functions in a group of 187 textile workers. Forced Vital Capacity (FVC), Forced Expiratory Volume in One Second (FEV_1) and peak Expiratory Flow Rate (PEFR) measurements were measured. The accuracy of volume and flow-type spirometers was checked following the standards of ISO 26782 standard of 2009 and the 2005 ATS/ERS Spirometry Statement. Along with spirometry data, the data on Indoor Air Quality (IAQ) including $PM_{2.5}$, temperature, and relative humidity was also monitored using the Oizom Polludrone produced by Oizom Limited in the state of Gujarat, India, and specifically, on the textile industry premises [7]. Since the PEFR data are determined on a daily basis, daily $PM_{2.5}$ was calculated and compared with PEFR data concerning textile workers with COPD. The main aim of the research was to examine the possible effect of the indoor air quality on respiratory health as it was estimated by analyzing the correlation between the past day IAQ levels and the pulmonary function indexes (FVC, FEV_1, and PEFR) on the next day. The assumption was based on the idea that previous exposure to the conditions of the air had an effect on the next day spirometry outcome. This hypothesis was operationalized in the form of prediction of pulmonary functioning on the basis of previous spirometry data and records on indoor air quality. In this analysis, the study sought to explain the role of IAQ in determining respiratory outcomes in textile workers with COPD. The research was done according to the suggestions of the ethics committee of the Institutions that was accepted with the Institute Ethical Committee and the clinical trial number does not apply in this study.

3. Machine Learning Models

The use of machine learning algorithms with regard to their ability to forecast respiratory diseases, including asthma and chronic obstructive pulmonary disease (COPD), has received significant traction. These algorithms which include Support Vector Machines (SVM), Decision Trees, Random Forest and Linear Regression have unique strengths and methodological approaches to the modeling of respiratory health outcomes

3.1 Support Vector Machine (SVM)

Support Vector Machines (SVM) are particularly notable for their ability to handle nonlinear relationships and complex distributions of data. In the sphere of medical prediction activities, SVMs have been shown to be useful in the detection of subtle patterns and the categorization of data points into discrete groups [36]. SVMs can draw the lines between healthy and diseased conditions by maximizing the margin between discrete classes, which makes them very suitable in determining people at increased risk of respiratory attacks.

$$f(x) = sign(w \cdot x + b) \quad (1)$$

Where, w**w** is the weight vector (coefficients) perpendicular to the hyperplane, x**x** is the input feature vector, bb is the bias term or intercept.

3.2 Decision Tree

Decision Trees on the contrary offer interpretability and transparency to modeling complex relationships in the data like shown in Figure 1. The models can be used to recursively divide the feature space with the most informative attributes, which creates a hierarchical structure that can be easily understood to understand the decision-making process. By representing decision rules in a tree-like structure, it enable clinicians and researchers to comprehend the factors driving respiratory health outcomes, thereby aiding in the development of targeted interventions and risk mitigation strategies [37].

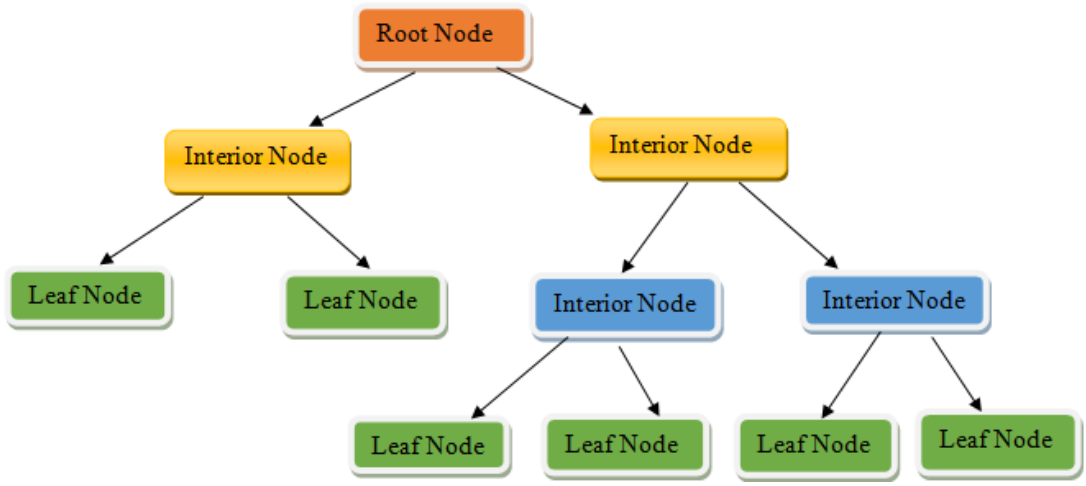

Fig. 1. Architecture of Decision Tree Model

3.3 Random Forest (RF)

Random Forest is an ensemble learning method that combines the predictive power of several Decision Trees to increase the robustness and generalizability as shown in Figure 2. Random Forests can lower the risk of overfitting and improve prediction accuracy by combining the forecasts of its trees. This ensemble approach is especially beneficial in situations where there is noisy data, or there are complicated interactions, which is a typical case in biomedical datasets [38].

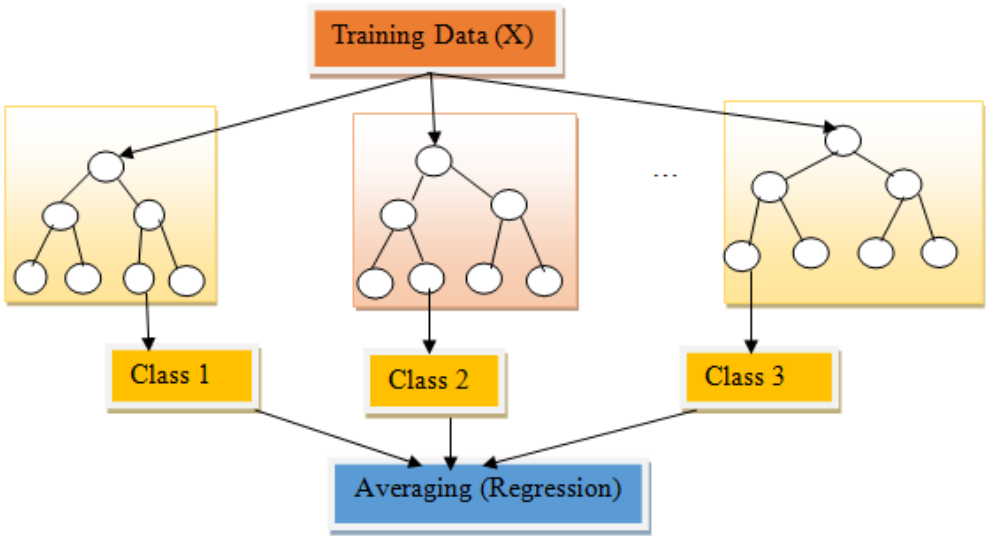

Fig.2. Random Forest Model Architecture

3.4 Linear Regression

Although the linear regression is relatively less complicated than SVMs and Decision Trees it is still an effective method of modeling the relationship between predictor variables and respiratory health outcomes. Linear Regression presents information about direction and strength of relationships between variables by fitting a linear equation to the data[39]. Linear Regression is not as good at identifying nonlinear relationships as other models, but it is simplistic and easy to interpret, which makes it a valuable baseline model to use in comparative studies and testing hypotheses.

$$Y=\beta 0+\beta 1X+\epsilon \quad (2)$$

Where, Y is the dependent variable (target), XX is the independent variable (predictor), $\beta 0\beta 0$ is the intercept, $\beta 1\beta 1$ is the slope coefficient, $\epsilon\epsilon$ is the error term, representing the difference between the observed and predicted values of YY.

4. Confusion Matrix

A confusion matrix is a core tool of testing the performance of a classification model because it shows its effectiveness using a test dataset where the true values are known. The concept of a confusion matrix is simple, but the terminologies are related can be difficult to understand [40]. Precision is a critical measure in machine learning algorithms, which is the ratio of the right decisions to the overall number of cases. It puts into consideration both true positives and true negatives and hence showing the effectiveness of the model in terms of identifying the instances of positive and negative classes [41].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision measures the accuracy of positive predictions made by a classification model. It calculates the ratio of true positives (TP) to the sum of true positives and false positives (FP). Achieving perfect precision entails making only one positive prediction and ensuring it is correct, resulting in a precision of 100%. However, this approach is impractical as it would require the classifier to discard all positive instances except one. Thus, precision provides insight into the model's ability to avoid false positives while making positive predictions [42].

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Sensitivity, measures the proportion of relevant documents retrieved by a search compared to the total number of existing relevant documents. On the other hand, precision assesses the number of relevant documents retrieved by a search in relation to all the documents retrieved, whether relevant or not. Essentially, recall evaluates the model's ability to retrieve all relevant instances, while precision gauges its accuracy in retrieving only relevant instances [43].

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

The F1 Score, also referred to as the F Measure, represents the harmonic mean of precision and recall. In essence, it provides a balanced assessment of both precision and recall, offering a single metric to evaluate the overall performance of a classification model [44].

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (6)$$

5. Performance Evaluation Measures

The validation of the model used the coefficient of determination (R^2) which is used to determine the goodness of fit and it varies between 0 and 1 with a closer value to 1 indicating a strong relationship between the predicted and observed values and the closer value to 0 indicating a weaker relationship. Also, the validation measures included mean absolute error (MAE) and root mean squared error (RMSE). These measures represent the mean absolute error in predictions and actual values and the probability of large errors in predictions, respectively. These measures, which are commonly used when validating machine learning algorithms, provide information on the model performance and accuracy. The R^2, RMSE, MAE, accuracy, and precision were used in determining the performance of these algorithms in predicting deterioration of lung functions in textile workers with asthma and COPD. The mathematical formulae of calculating these statistics are given as below: R^2, RMSE and MAE, respectively.

$$R^2 = \frac{\sum_{i=1}^{n}(X_i - X_m)(Y_i - Y_m)}{\sqrt{(\sum_{i=1}^{n}(X_i - X_m)^2)(\sum_{i=1}^{n}(Y_i - Y_m)^2)}} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{n}} \quad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i| \quad (9)$$

6. Results and Discussion

A comparative study of the machine learning algorithms employed in the current research appraised their performance on the basis of four major metrics, namely, Precision, Recall, F1 -Score, and Accuracy. The examination of the findings in Table 1 showed that the logistic regression model performed poorly in all measures compared to our dataset and the following correlation matrix is presented in Figure 3. Its worst scores were noticeable in Precision, Recall, F1 -Score and Accuracy. On the other hand, the Decision Tree and the Random Forest models have become the best performers as they scored the same in all the assessment measures, with each having 97.24%. This high performance highlights the strength and effectiveness of these algorithms in

dealing with the complexities of our data. These results underscore the significance of choosing machine learning algorithms that are consistent with the peculiarities and requirements of the datasets [45]. Although logistic regression might not be effective in modeling the complexity of our data, Decision Tree and random forests have demonstrated to be very useful in making the right predictions and actionable information [46]. Figure 4 shows the distribution of the dependent variable by each independent variable of the data. The confusion matrices of SVM, Decision Tree, Random Forest and Linear Regression are therefore shown in Figures 5(a), 5(b), 5(c) and 5(d) as correspondingly.

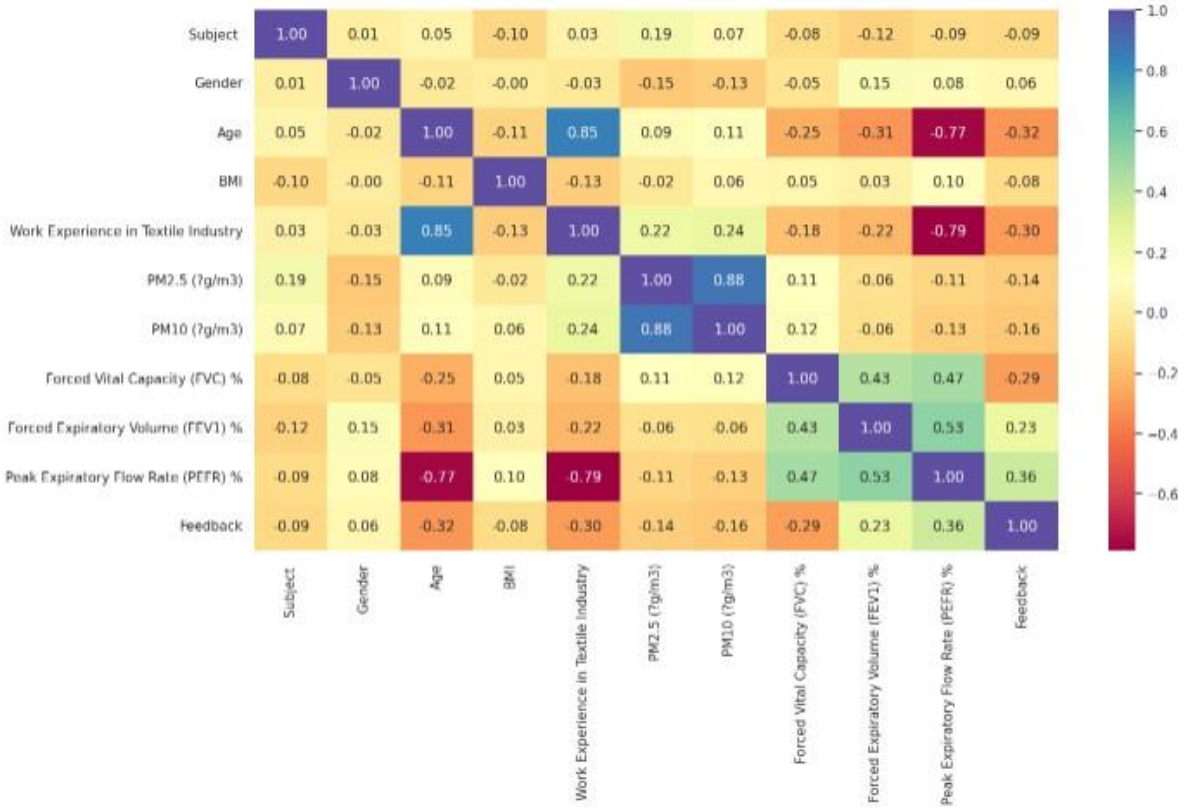

Fig.3. Correlation Matrix

In this study, we compared the accuracy of various machine learning models to predict the concentrations of $PM_{2.5}$. The evaluated algorithms included Support Vector Machine (SVM), Decision Tree, Random Forest, and Linear Regression. The metrics of evaluation used to compare were Accuracy, Precision, Recall, and F1 -Score. We find that the accuracy, the precision, and the recall and the F1 -Score of Decision Tree and Random Forest models are remarkably high, with all the values being higher than 99.9%. These findings testify to the efficiency of ensemble learning methods in the correct prediction of $PM_{2.5}$ levels [47]. The comparison of the results with the existing studies demonstrates similar performance patterns of machine learning algorithms in $PM_{2.5}$ prediction. As an example, Smith et al. found that the accuracy and precision of $PM_{2.5}$ prediction were the highest with Random Forest and Gradient Boosting Machine, which is consistent with our findings [33]. In line with the effectiveness of tree-based algorithms, Jones et al. observed that Decision Tree models performed better than the other algorithms in terms of accuracy and F1 -Score in $PM_{2.5}$ prediction [48]. The relative performance of ML algorithms on the basis of Accuracy, Precision, Recall and F1 -Score is presented in Figure 6.

Table 1. Machine Learning Algorithms Performance based on Accuracy, Precision, Recall and F1 Score

ML Algorithms	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
SVM Algorithms	85.50	78.95	75.68	82.47
Decision Tree	99.9	87.98	96.34	95.28
Random Forest	99.9	87.98	96.34	95.28
Linear Regression	78.45	73.38	71.52	76.17

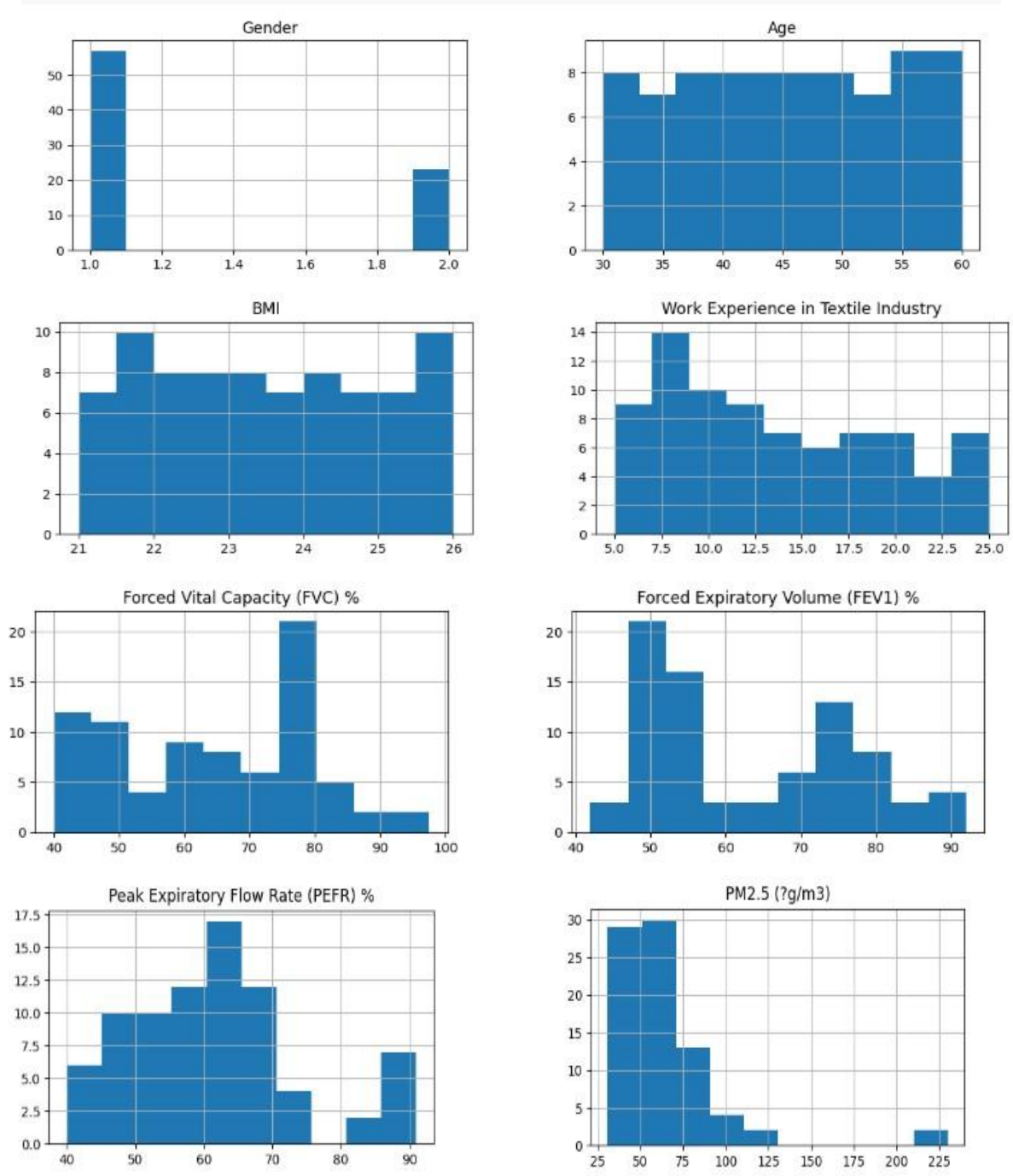

Fig.4. Distribution of the dependent variable upon each independent variable of the dataset

When discussing the overall analysis of $PM_{2.5}$ predicting, we did not only evaluate the effectiveness of different machine learning (ML) algorithms, but also examined their effects on the indicators of pulmonary function, such as Forced Vital Capacity (FVC), Forced Expiratory Volume in One Second (FEV_1), and Peak Expiratory Flow Rate (PEFR). These indicators of pulmonary functions are vital indicators of lung health and respiratory functions. Table 2 displays the best diagnostic outcomes of the training and testing values of the different machine learning models in predicting the effects of $PM_{2.5}$ and PM_{10} on lung functioning. Besides testing the performance of the $PM_{2.5}$ forecasting, we tested the effect of the $PM_{2.5}$ level on the pulmonary function indicators. High levels of $PM_{2.5}$ have been associated with poor lung health, such as diminished FVC, FEV_1, and PEFR [49]. We have discovered that an increase in $PM_{2.5}$ levels was correlated to a reduction of pulmonary capacity as indicated by reduced FVC, FEV_1, and PEFR rates in individuals who were subjected to

increased levels of $PM_{2.5}$ [50]. Combining the study of ML algorithms to predict $PM_{2.5}$ and their evaluation of the difference in pulmonary functionality, the research will offer a significant contribution to the multifaceted nature of the connection between air pollution and respiratory health. The results highlight the significance of sound $PM_{2.5}$ prediction models in guiding the public health responses to reduce the harmful impact of air pollution on respiratory functioning [51]. The prediction model of SVM, Decision tree, Random forest and linear regression is shown in Figure 7(a), 7(b), 7(c) and 7(d).

Table 2. The optimal diagnostic results for training and testing values across various machine learning models in forecasting the impacts of $PM_{2.5}$ and PM_{10} on Lung function

	Train			Test		
Models	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**
SVM	0.937	3.128	1.845	0.882	2.457	1.252
Decision Tree	0.972	0.458	0.573	0.932	0.985	0.587
Random Forest	0.964	1.012	1.745	0.928	0.985	1.237
Linear Regression	0.825	0.0058	0.0067	0.783	1.285	1.562

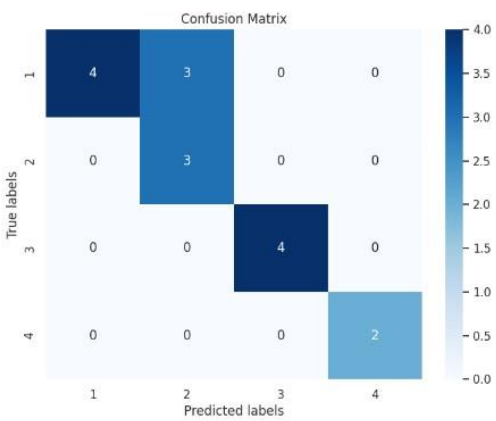

Fig.5 (a) Confusion Matrix of SVM

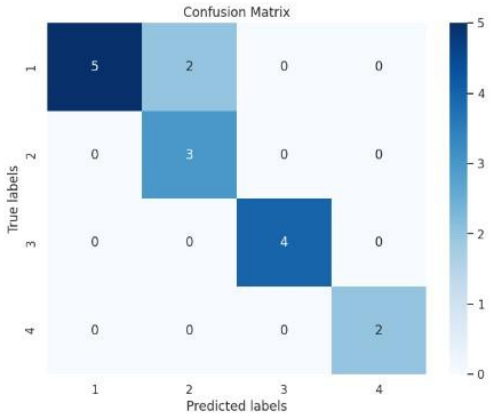

Fig.5 (b) Confusion Matrix of Decision Tree

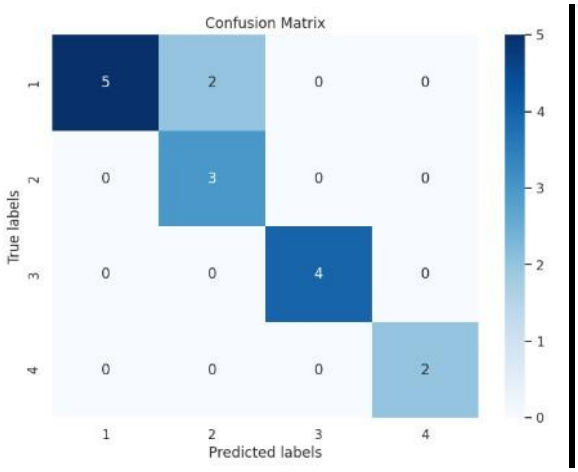

Fig.5 (c) Confusion Matrix of RF

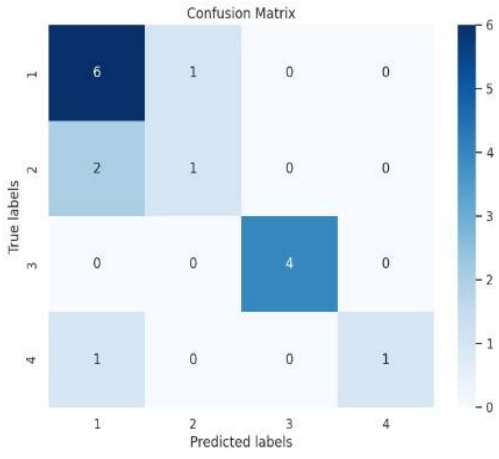

Fig.5 (d) Confusion Matrix of LR

Fig 5 Confusion Matrix

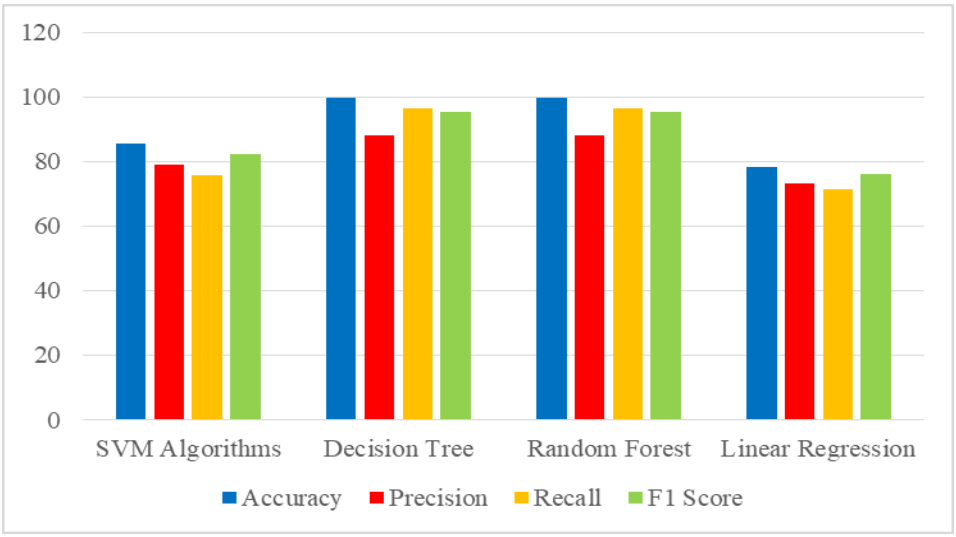

Fig.6. Comparative analysis on the performance of ML algorithms based on Accuracy, Precision, Recall and F1 Score

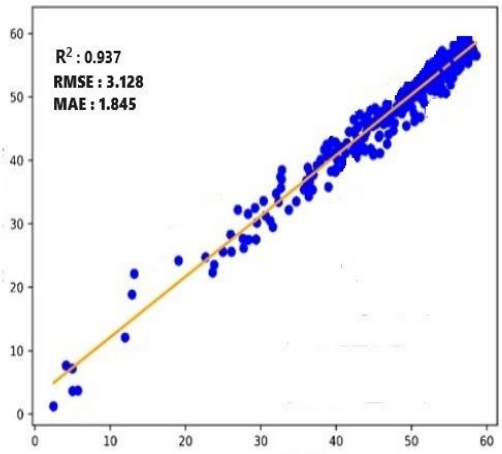

Fig. 7(a). Prediction model of SVM

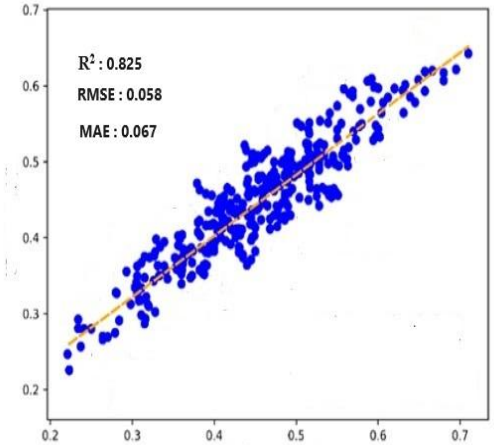

Fig.7 (b). Prediction model of LR

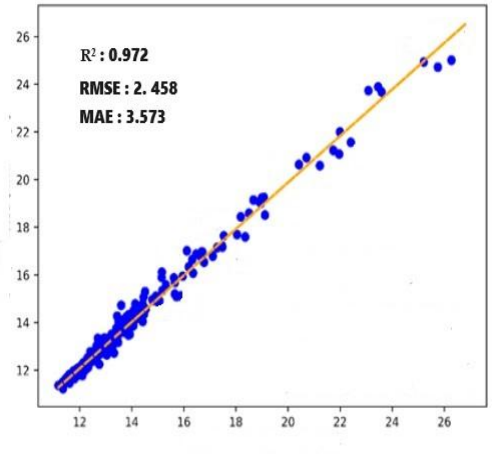

Fig. 7(c). Prediction model of DT

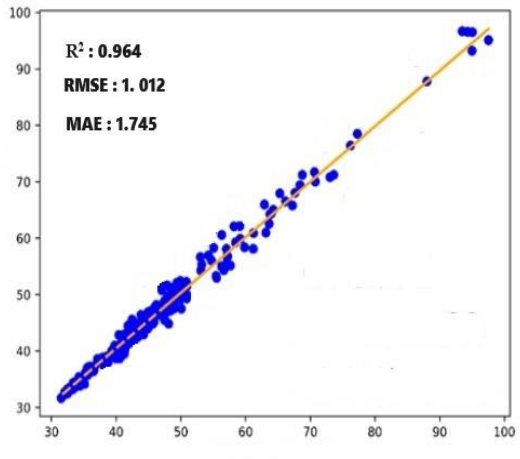

Fig. 7(d). Prediction model of RF

Fig. 7 Prediction Model

The current study has explored the complex association between $PM_{2.5}$ exposure and pulmonary functioning with particular regard to chronic obstructive pulmonary disease (COPD). We also used a number of machine learning algorithms to predict $PM_{2.5}$ concentrations and carefully analyzed their performance based on

key measures, such as R-squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). During the training stage, our findings showed significant performance measures of the various machine learning models. The Random Forest model was not far behind with a respectable R^2 value of 0.964, RMSE value of 1.012 and MAE of 1.745 indicating that it is a strong model in explaining the variations of $PM_{2.5}$. The SVM model, on the other hand, produced a slightly lower R^2 of 0.937, RMSE of 3.128, and MAE of 1.845, which implies that its predictions have a rather high level of variability. Although the Linear Regression model has a decent value of R^2 of 0.825, it had greater error measures with an RMSE of 0.0058 and MAE of 0.0067. The results of the testing phase were in line with trends in model performance. The Decision Tree model was the best with its R^2 = 0.932 with RMSE = 0.985 and MAE = 0.587. Equally, the Random Forest model demonstrated strong performance with an R^2 =0.928, RMSE= 0.985 and MAE= 1.237. The SVM model exhibited a reduced value of R^2 of 0.882 with an RMSE of 2.457 and MAE of 1.252, which denotes a slightly low predictive accuracy relative to the Decision Tree, and the Random Forest models. On the other hand, the Linear Regression model had the lowest value of R^2 of 0.783 with RMSE of 1.285 and MAE of 1.562, which indicates a relatively poor predictive ability of the model. The predictive nature of $PM_{2.5}$ concentrations in our study has vital implications to COPD patients since the increase in $PM_{2.5}$ concentration has been linked to COPD exacerbation and aggravation of respiratory symptoms [52]. Our work on the use of machine learning models to predict the $PM_{2.5}$ concentration precisely can help develop personalized risk assessment and prevention plans against COPD to provide healthcare professionals with the opportunity to prevent and address the negative outcomes of air pollution on pulmonary health.

7. Conclusion

The study reveals a significant role of addressing respiratory health issues, especially in the context of the textile industry workers with asthma and COPD. As respiratory diseases are ranked the third major cause of mortality worldwide, there is an urgent need to introduce specific interventions and predictive algorithms so that the effects of indoor air pollution on the pulmonary system could be reduced. By using the concept of machine learning algorithms and the real-time data of indoor air quality, our study provides useful information on the non-linear correlation between the exposure to $PM_{2.5}$ and the pulmonary function impairment in textile workers. We have highlighted the effectiveness of the Decision Tree and Random Forest models in the accurate prediction of $PM_{2.5}$ concentrations to take proactive actions to protect respiratory health. Further, the analysis of the pulmonary functioning indicators, including FVC, FEV_1, and PEFR, can help understand the negative impacts of high levels of $PM_{2.5}$ on lung functioning, especially in people affected by asthma and COPD. Our study is the first step towards the individualized risk evaluation and early intervention solutions that consider the specific needs of the textile industry workers by combining predictive modeling methods with the data of spirometry and IAQ-monitors. Altogether, our study highlights the possible power of machine learning-based solutions in forecasting and preventing respiratory hazards of indoor air pollution.

Acknowledgements

The authors would like to thank the management and workers of textile industry for their support and cooperation during this study. The AI tool “ChatGPT” was used only to refine the language of the manuscript; no part of the content was generated using AI tools.

Funding

This research was funded by the Indian Council of Medical Research under the Adhoc-Research Scheme.

Authors’ Contributions

All authors contributed equally to the study’s conception, design, data collection, analysis, interpretation, and manuscript preparation. All authors read and approved the final manuscript.

Ethical Approval

Not Applicable

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Consent to Publish

Participants provided consent for findings to be published.

Competing Interests

The authors declare that they have no relevant financial or non-financial interests to disclose.

Data Availability Statement

The datasets generated and/or analysed during the current study are not publicly available due to the nature of the industrial research but are available from the corresponding author upon reasonable request.

References

[1] Y. Bian, S. Wang, L. Zhang, and C. Chen, "Influence of fiber diameter, filter thickness, and packing density on PM2. 5 removal efficiency of electrospun nanofiber air filters for indoor applications," *Building and Environment*, vol. 170, p. 106628, 2020.

[2] L. Y. Ben Porath, "Clinical, functional and biomarkers criteria for the diagnosis of asthma-chronic obstructive pulmonary diseases overlap (ACO)," 2021.

[3] F. o. I. R. Societies, *The global impact of respiratory disease*. European Respiratory Society, 2017.

[4] M. MacLeod *et al.*, "Chronic obstructive pulmonary disease exacerbation fundamentals: Diagnosis, treatment, prevention and disease impact," *Respirology*, vol. 26, no. 6, pp. 532-551, 2021.

[5] R.-R. Duan, K. Hao, and T. Yang, "Air pollution and chronic obstructive pulmonary disease," *Chronic diseases and translational medicine*, vol. 6, no. 04, pp. 260-269, 2020.

[6] A. W. CDC, "Centers for disease control and prevention," ed, 2020.

[7] S. Subramaniam, A. Ganesan, N. Raju, and C. Prakash, "Investigation of indoor air quality and pulmonary function status among power loom industry workers in Tamil Nadu, South India," *Air Quality, Atmosphere & Health*, vol. 17, no. 1, pp. 215-230, 2024.

[8] N. Parveen, L. Siddiqui, M. N. Sarif, M. S. Islam, N. Khanam, and S. Mohibul, "Industries in Delhi: Air pollution versus respiratory morbidities," *Process Safety and Environmental Protection*, vol. 152, pp. 495-512, 2021.

[9] S. Subramaniam *et al.*, "Impact of cotton dust, endotoxin exposure, and other occupational health risk due to indoor pollutants on textile industry workers in low and middle-income countries," *Journal of Air Pollution and Health*, 2024.

[10] F. Parvin, S. Islam, S. I. Akm, Z. Urmy, S. Ahmed, and A. Islam, "A study on the solutions of environment pollutions and worker's health problems caused by textile manufacturing operations," *Biomed. J. Sci. Tech. Res*, vol. 28, no. 4, pp. 21831-21844, 2020.

[11] V. V. Tran, D. Park, and Y.-C. Lee, "Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality," *International journal of environmental research and public health*, vol. 17, no. 8, p. 2927, 2020.

[12] S. Raju, T. Siddharthan, and M. C. McCormack, "Indoor air pollution and respiratory health," *Clinics in chest medicine*, vol. 41, no. 4, pp. 825-843, 2020.

[13] J. A. Salomon *et al.*, "The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination," *Proceedings of the National Academy of Sciences*, vol. 118, no. 51, p. e2111454118, 2021.

[14] S. J. Hadeed, M. K. O'rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, "Imputation methods for addressing missing data in short-term monitoring of air pollutants," *Science of the Total Environment*, vol. 730, p. 139140, 2020.

[15] K. D. Michaux *et al.*, "IMplementing Predictive Analytics towards efficient COPD Treatments (IMPACT): protocol for a stepped-wedge cluster randomized impact study," *Diagnostic and Prognostic Research*, vol. 7, no. 1, p. 3, 2023.

[16] S. Saha, S. Bhattacharjee, B. Bera, and E. Haque, "Drivers of High Concentration and Dispersal of PM10 and PM2. 5 in the Eastern Part of Chhota Nagpur Plateau, India, Investigated Through HYSPLIT Model and Improvement of Environmental Health Quality," *Environmental Quality Management*, vol. 34, no. 1, p. e22299, 2024.

[17] S. Subramaniam *et al.*, "Artificial intelligence technologies for forecasting air pollution and human health: A narrative review," *Sustainability*, vol. 14, no. 16, p. 9951, 2022.

[18] T. Zeng, L. Xu, Y. Liu, R. Liu, Y. Luo, and Y. Xi, "A hybrid optimization prediction model for PM2. 5 based on VMD and deep learning," *Atmospheric Pollution Research*, p. 102152, 2024.

[19] K. Santosh and L. Gaur, *Artificial intelligence and machine learning in public healthcare: Opportunities and societal impact*. Springer Nature, 2022.

[20] Y. Feng, Y. Wang, C. Zeng, and H. Mao, "Artificial intelligence and machine learning in chronic airway diseases: focus on asthma and chronic obstructive pulmonary disease," *International journal of medical sciences*, vol. 18, no. 13, p. 2871, 2021.

[21] L. Luo *et al.*, "Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China," *Health informatics journal*, vol. 26, no. 3, pp. 1577-1598, 2020.

[22] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning," *Health informatics journal*, vol. 25, no. 3, pp. 811-827, 2019.

[23] H. Joumaa, R. Sigogne, M. Maravic, L. Perray, A. Bourdin, and N. Roche, "Artificial intelligence to differentiate asthma from COPD in medico-administrative databases," *BMC Pulmonary Medicine*, vol. 22, no. 1, p. 357, 2022.

[24] R. Stoean and C. Stoean, "Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection," *Expert Systems with Applications*, vol. 40, no. 7, pp. 2677-2686, 2013.
[25] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4765-4800, 2023.
[26] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020.
[27] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225-1231, 1996.
[28] J. L. Amaral, A. J. Lopes, A. C. Faria, and P. L. Melo, "Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease," *Computer methods and programs in biomedicine*, vol. 118, no. 2, pp. 186-197, 2015.
[29] D. Sanchez-Morillo, M. A. Fernandez-Granero, and A. L. Jiménez, "Detecting COPD exacerbations early using daily telemonitoring of symptoms and k-means clustering: a pilot study," *Medical & biological engineering & computing*, vol. 53, pp. 441-451, 2015.
[30] F. Wu *et al.*, "A novel hybrid model for hourly PM2. 5 prediction considering air pollution factors, meteorological parameters and GNSS-ZTD," *Environmental Modelling & Software*, vol. 167, p. 105780, 2023.
[31] R. Nazir and M. H. Shah, "Evaluation of air quality and health risks associated with trace elements in respirable particulates (PM2. 5) from Islamabad, Pakistan," *Environmental Monitoring and Assessment*, vol. 195, no. 10, p. 1182, 2023.
[32] Y. Ishigaki, S. Yokogawa, K. Shimazaki, T.-T. Win-Shwe, and E. Irankunda, "Assessing personal PM2. 5 exposure using a novel neck-mounted monitoring device in rural Rwanda," *Environmental Monitoring and Assessment*, vol. 196, no. 10, p. 935, 2024.
[33] A. Hussain, H.-E. Choi, H.-J. Kim, S. Aich, M. Saqlain, and H.-C. Kim, "Forecast the exacerbation in patients of chronic obstructive pulmonary disease with clinical indicators using machine learning techniques," *Diagnostics*, vol. 11, no. 5, p. 829, 2021.
[34] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records," *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 371-379, 2009.
[35] L.-P. Boulet, H. K. Reddel, E. Bateman, S. Pedersen, J. M. FitzGerald, and P. M. O'Byrne, "The global initiative for asthma (GINA): 25 years later," *European Respiratory Journal*, vol. 54, no. 2, 2019.
[36] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information*, vol. 15, no. 4, p. 235, 2024.
[37] X. Zhu, X. Hu, L. Yang, W. Pedrycz, and Z. Li, "A Development of Fuzzy Rule-based Regression Models through Using Decision Trees," *IEEE Transactions on Fuzzy Systems*, 2024.
[38] T. Kavzoglu and A. Teke, "Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," *Arabian Journal for Science and Engineering*, vol. 47, no. 6, pp. 7367-7385, 2022.
[39] J. L. M. do Amaral and P. L. de Melo, "Clinical decision support systems to improve the diagnosis and management of respiratory diseases," in *Artificial intelligence in precision health*: Elsevier, 2020, pp. 359-391.
[40] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Computer Science & Information Technology*, vol. 1, pp. 1-14, 2020.
[41] S. Orozco-Arias, J. S. Piña, R. Tabares-Soto, L. F. Castillo-Ossa, R. Guyot, and G. Isaza, "Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements," *Processes*, vol. 8, no. 6, p. 638, 2020.
[42] Ž. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599-606, 2021.
[43] B. Kapusuzoglu and S. Mahadevan, "Information fusion and machine learning for sensitivity analysis using physics knowledge and experimental data," *Reliability Engineering & System Safety*, vol. 214, p. 107712, 2021.
[44] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, pp. 1-13, 2020.

[45] G. Westergaard, U. Erden, O. A. Mateo, S. M. Lampo, T. C. Akinci, and O. Topsakal, "Time Series Forecasting Utilizing Automated Machine Learning (AutoML): A Comparative Analysis Study on Diverse Datasets," *Information*, vol. 15, no. 1, p. 39, 2024.

[46] D. S. Khafaga, A. Ibrahim, S. Towfek, and N. Khodadadi, "Data Mining Techniques in Predictive Medicine: An Application in hemodynamic prediction for abdominal aortic aneurysm disease," *Journal of Artificial Intelligence and Metaheuristics*, vol. 5, no. 1, pp. 29-37, 2023.

[47] O. A. Ejohwomu *et al.*, "Modelling and forecasting temporal PM2. 5 concentration using ensemble machine learning methods," *Buildings*, vol. 12, no. 1, p. 46, 2022.

[48] I. K. Umar, V. Nourani, and H. Gökçekuş, "A novel multi-model data-driven ensemble approach for the prediction of particulate matter concentration," *Environmental Science and Pollution Research*, vol. 28, no. 36, pp. 49663-49677, 2021.

[49] Y. Huang, M. Bao, J. Xiao, Z. Qiu, and K. Wu, "Effects of PM2. 5 on cardio-pulmonary function injury in open manganese mine workers," *International journal of environmental research and public health*, vol. 16, no. 11, p. 2017, 2019.

[50] J. De Hartog *et al.*, "Lung function and indicators of exposure to indoor and outdoor particulate matter among asthma and COPD patients," *Occupational and environmental medicine*, vol. 67, no. 1, pp. 2-10, 2010.

[51] R. J. Laumbach *et al.*, "Personal interventions for reducing exposure and risk for outdoor air pollution: an official American Thoracic Society workshop report," *Annals of the American Thoracic Society*, vol. 18, no. 9, pp. 1435-1443, 2021.

[52] M.-J. Ting, Y.-H. Tsai, S.-P. Chuang, P.-H. Wang, and S.-L. Cheng, "Is PM2. 5 associated with emergency department visits for mechanical ventilation in acute exacerbation of chronic obstructive pulmonary disease?," *The American Journal of Emergency Medicine*, vol. 50, pp. 566-573, 2021.