OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 03 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

MACHINE LEARNING MODELING FOR FORECASTING FORCE VELOCITY CAPACITY

Muhammad Mubashar Ali[1], Javed Iqbal[2], Sahar Iqbal[3], *Muhammad Awais[2]

[1, 3, 4]Govt. Graduate College, Civil Lines, Sheikhupura, Pakistan.

[2]Govt. Waris Shah Graduate College, Jandiala Sher Khan, Sheikhupura, Pakistan.

***Corresponding Author:** (Ovaisqarni15@mail.com)*

***DOI:** (https://doi.org/10.71146/kjmr1030)*

Article Info

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license https://creativecommons.org/licenses/by/4.0

Abstract

A spirometer is a device used to monitor people with respiratory diseases who are unable to meet their own needs due to their poor health. Thus, spirometry was used to measure significant parameters. The present study attempted to develop machine-learning models based on Random Forest, KNN regression, Decision Tree Regression, and Support Vector Regression to forecast force velocity capacity values when the rigorous end-of-test conditions were not fully met. Smoking has become the leading cause of reduces lung function among males. We assessed the effects of smoking on lung function. To perform this study, a sample of 120 industrial workers was used. Traditional machine-learning algorithms were used and compared their accuracies and other performance measures. We found that the best algorithm was K-nearest Neighbor with 78 percent Coefficient of determination. These forecasted outcomes play a crucial role in the evaluation of spirometry measures in clinical practice since they serve as reference values. Research advises industrial workers to keep a healthy body and abstain from smoking to survive their respiratory systems.

Keywords: *Spirometry analysis, Force Velocity Capacity, machine learning, forecasting*

1. Introduction

One of the biggest health issues facing the globe, chronic obstructive pulmonary disease (COPD) is predicted to rank third in terms of causes of mortality (Yao and Tao, 2021; Erdal, 2021; Subramaniyan, 2023). According to the World Health Organization (WHO), COPD is the world's fourth leading cause of death, accounting for 5% of all deaths in 2023 (Sharma et al., 2023; Singh et al.,2023). COPD is projected to become the world's third biggest cause of death by 2030 (Hashemi et al., 2020; Alqahtani, 2022). It was difficult to make an accurate diagnosis of COPD (Maldonado-Franco et al., 2023; Lin, 2023). Inadequate spirometry quality assurance may lead to misdiagnosis; therefore, a spirometry assessment should only be performed by competent and trained staff in an environment with a consistent quality assurance programme. Some patients find it difficult to achieve these requirements during spirometry. (Schermer et al 2003) studied that Half of the spirometry results were inadequate due to incomplete tests, increasing the risk of misdiagnosis and incorrect treatment. More precisely, in order to acquire a correct force velocity capacity (FVC) value, participants must exhaust themselves to the point where they are unable to exhale any more in order to meet the EOT requirements. This presents a number of difficulties for patients with impaired lung function (Ulmer, 2005). For the reasons stated above, there is a clinical need to create an FVC quantification method capable of obtaining appropriate FVC values even if the EOT criteria are not completely met. Multiple linear regression was employed by Mengesha and Mekonnen (1985) to forecast the results of spirometry based on a few demographic inquiries in Ethiopia. Different prediction models for FVC and FEV1 in the Greek population were found by Baltopoulos et al. (2000). A strong association was discovered between the anticipated FVC and FEV1 levels. A regression equation for estimating forced expiratory volumes in 2 seconds (FEV2) and 3 seconds (FEV3) was proposed by Ioachimescu et al., (2005). Nevertheless, there was a high rate of misclassification if the condition of a postbronchodilator FEV1/FVC less than 0.7 was used as the threshold to distinguish airflow limitations. Wang et al., (2019) proposed prediction equations relying on support vector regression with only four covariates to estimate FVC where the EOT criteria in traditional spirometry was not meet. Several studies developed equations for spirometry parameters. Do Amara and De Melo (2020) improved the diagnosis and management respiratory disease system by applied different machine-learning techniques and compared their performance evaluation. Bolourani et al., (2020) predicted respiratory failure using machine learning techniques. They identified the risk factor of respiratory failure. Zaman et al., (2022) predicted lung function, from 323 recorded voice files using random forest algorithm. Their research work could be implemented to aid decision making and self-monitoring in asthma. Kristensen et al., (2023) used random forest algorithm and improved clinical spirometry references as compared to previous studies. The purpose of this study is to assess the impact of smoking on lung function of industrial workers and also compare the machine learning models utilizing the same dataset.

2. Material and Methods

2.1 Data Collection

Data used in this study was experimental data. Data was obtained from the industrial sites in district Sheikhupura. Factory workers were our respondents. A sample of 120 workers was selected using convenient sampling technique. The FVC and FEV1 measurements met the American Thoracic Society's requirements and were considered acceptable.16 Miller et al. (2005) provide standards.

2.2 Predictors Selection

The selection of predictors was collected from research works conducted by (Jaakkola et al., 2019; Kumar et al., 2022; Khan et al., 2022; Enayatjazi et al., 2023) that illustrate harm effect of smoking on men’s respiratory

system. The selected predictors were: Age, Height, Weight, and Smoking. The predictors would be examined to determine whether there was a significant relationship between them and FVC; those that did not have a relationship would not be considered in the process of developing a prediction model.

Table 1: Descriptive Statistics of the selected variables

Descriptive Statistics						
	N	**Mean**	**SD**	**SE**	**Grubbs' test**	**p-value**
FVC	120	2.961	1.118	0.102	3.56	0.031
Age(year)	120	34.200	10.897	0.995	2.37	0.998
weight(kg)	120	75.069	9.418	0.860	2.75	0.636
Height(inch)	120	5.526	0.213	0.019	2.51	0.997
Smoking	Smokers:69	Non-smokers: 51				

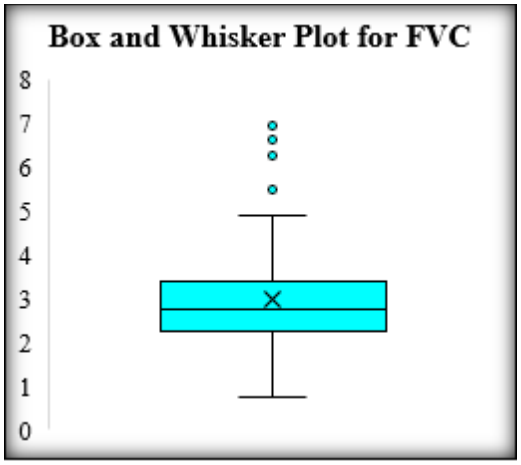


figure 1

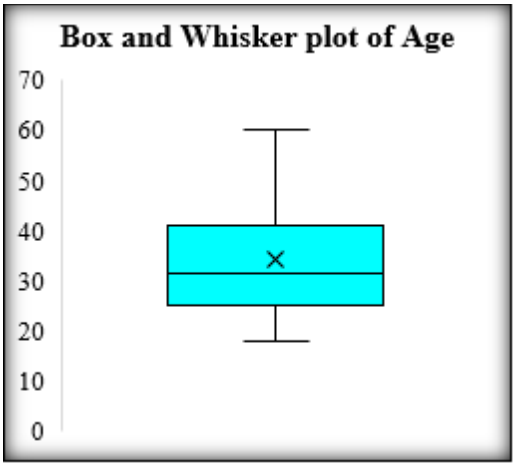


figure 2

Boxplot is used to visually represent the dataset.. Boxplots of all continuous variables are made to identify outliers in figure 1 to 4 respectively. Outliers occur in study variable.

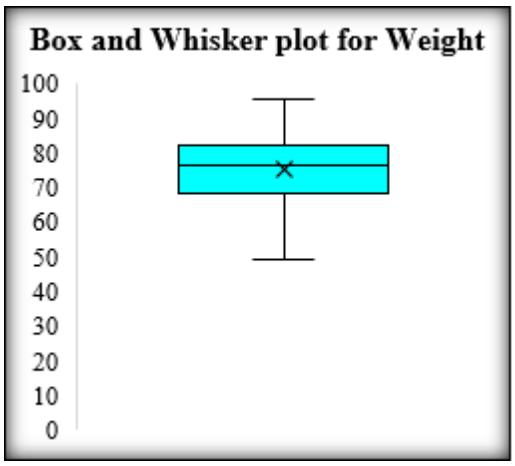


figure 3

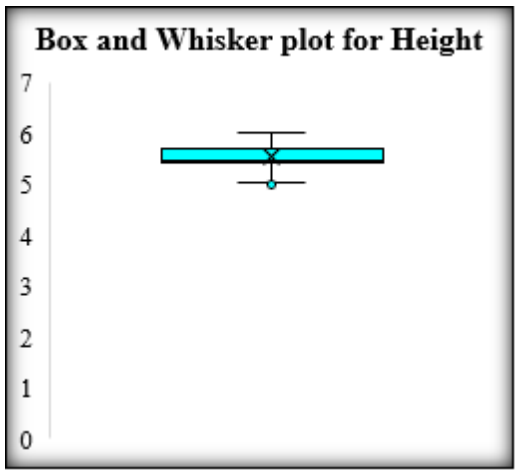


figure 4

Table 2: Diagnosis tests for study variable

Hypothesis	Test name	Test statistic	P-value
FVC follows Normal distribution	Kolmogorov-Smirnov	KS = 0.123	0.981
There is outlier in FVC dataset	Grubbs' test	G = 2.61	0.01

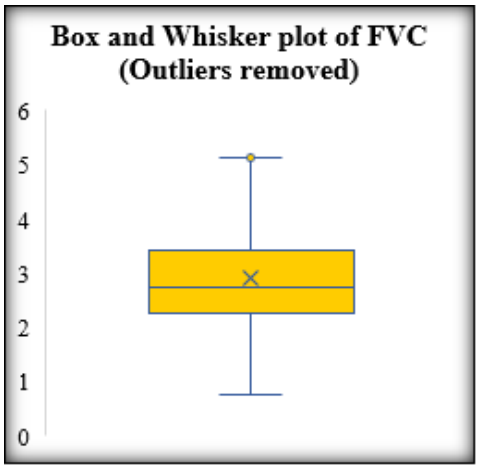


figure 5

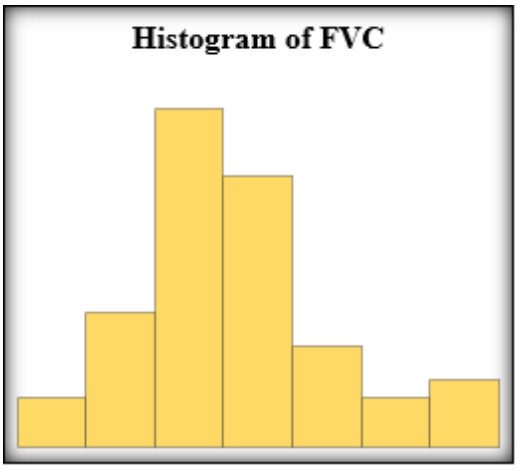


figure 6

To different diagnosis tests are applied after removing the outliers. Figure 5 and Grubbs' test value in table 2 indicate that there are no outliers in the values of FVC. Figure 6 and Kolmogorov-Smirnov test value indicate that FVC follows normal distribution.

Table 3: Significance of the Predictors

Predictor	Coefficient	p-value	R_square
Age	-0.0246	0.0080	5.75
Height	2.584	0.0000	24.33
Weight	0.06216	0.0000	27.42
Smoking	-1.54	0.0000	46.78

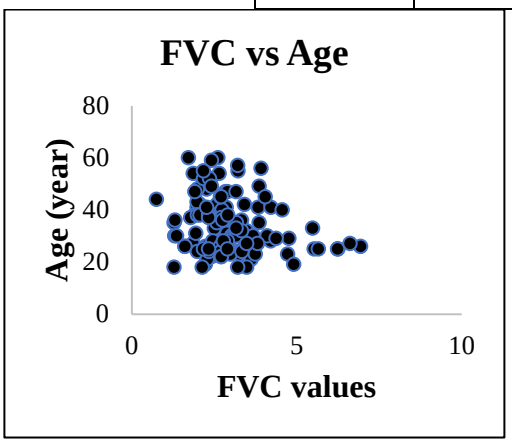


figure 7

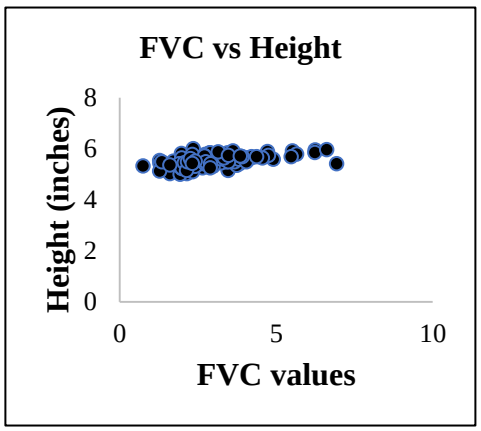


figure 8

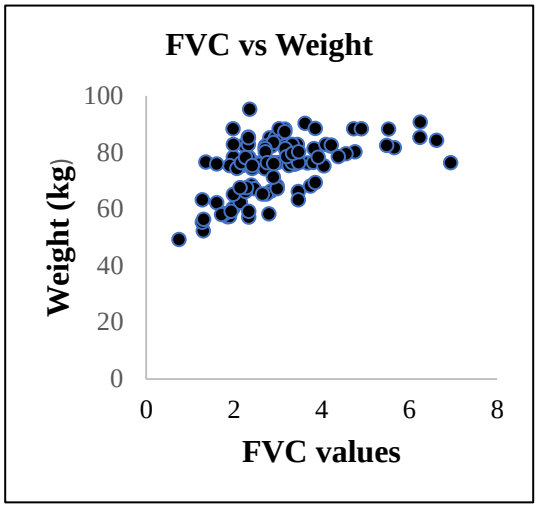


figure 9

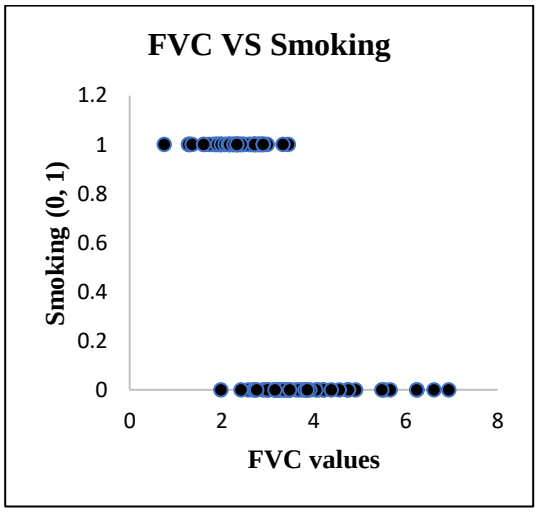


figure 10

Four variables: age (A), Height (H), Weight (W) and Smoking (S) are selected as independent variables. They are are significant and highly effect the HVC. Table 3 depicts that they are all are significant but age and smoking are negatively correlated with HVC where weight and height are positively correlated.

2.3 Performance Measures

Performance measures for the evaluation of ML-regression techniques are described below:

2.3.1 Mean Square Error (MSE)

Mean square error (MSE) calculates the average of the squared deviations between the estimated and actual performance values. This is determined as the mean of the squared residuals between the predicted and true values, as illustrated in equation (1):

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_{predicted} - Y_{actual}\right)^2}{n} \quad \text{(1)}$$

2.3.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error derived from the MSE, calculates the average separation between forecasted value and actual values, maintaining the identical units as the original data. RMSE is favored for its ability to diagnose significant errors and extreme values, as it involves a square root operation applied to the MSE, as illustrated in Equation (2):

$$MAE = \frac{\sum_{i=1}^{n}\left(Y_{predicted} - Y_{actual}\right)}{n} \quad \text{(2)}$$

2.3.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) calculates the average absolute variance between estimaeted and observed values. It is calculated as the average of the absolute residuals between estimated and actual values, as illustrated in Equation (3):

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{predicted} - Y_{actual}\right)^2}{n}} \quad \text{(3)}$$

2.3.4 R_square (Coefficient of Determination)

R-squared (R^2): R^2 assesses the absolute goodness of fit of the models and ranges from zero to one. A value of 1 indicates a perfect fit, while a value of 0 indicates no relationship between the variables. It is calculated as follows in Equation 4:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{predicted} - Y_{actual}\right)^2}{n}} \quad \text{(4)}$$

2.4 Machine Learning Regression Algorithms

Regression algorithms are a subset of machine learning algorithms that predict a continuous output variable based on one or more input features (Sarker, 2021). Regression aims to model the relationship between the dependent variable: output and one or more independent variables: inputs (Kamel and Abonazel, 2023). These algorithms attempt to find the best-fit line, curve, or surface that minimizes the difference between predicted and actual values (Roy et al., 2022; Viering and Logo, 2022). In this study, to achieve the goals of the study we applied four different ML-regression algorithms.

2.4.1 Random Forest Regression

Random Forests are a collection of decision trees (Bernard et al. 2009). They are a supervised learning algorithm that is used for classification and regression. The input data goes through several decision trees. The algorithm works by constructing a different number of decision trees during training time and then outputting the class, which is the mode of the classes (for classification) or the mean prediction of the individual trees.

2.4.2 Support Vector Regression

Support vector regression (SVR) also uses the same idea as SVM, but it tries to predict the real values here (Zhang and O'Donell, 2020). This algorithm uses hyperplanes to segregate the data (Boswell, 2002). If this separation is impossible, it uses the kernel trick, where the dimension is increased, and then the data points become separable by a hyperplane.

2.4.3 Decision Tree Regression

Tree models can be applied to any data with numerical and categorical features (Gunluk et al., 2021). Decision trees are effective at capturing the nonlinear interaction between features and the target variable (Mishra, 2021). Understanding the data is very intuitive because decision trees are similar to human-level thinking.

2.4.4 K-Nearest Neighbor Regression

KNN regression is a non-parametric method that intuitively approximates the relationship between independent variables and continuous outcomes by averaging observations in the same neighborhood (Oswald et al., 2000; Sabry, 2023).

2.5 Machine Learning Modeling

The study cohort was randomly divided into two datasets: a training dataset for generating ML models, and a validation dataset for applying and testing the model to obtain an unbiased estimate of its performance. The training dataset was then divided into 70% training and 30% testing.

Table 4: Performance Comparison of Machine-Learning Models

Regression	MSE	RMSE	MAE	MAPE (%)	R_square
Random Forest	0.816	0.903	0.637	20.22	0.75
SVP	0.902	0.95	0.637	18.75	0.71
KNN	0.246	0.496	0.398	15.13	0.78
Decision Tree	0.486	0.697	0.499	14.91	0.72

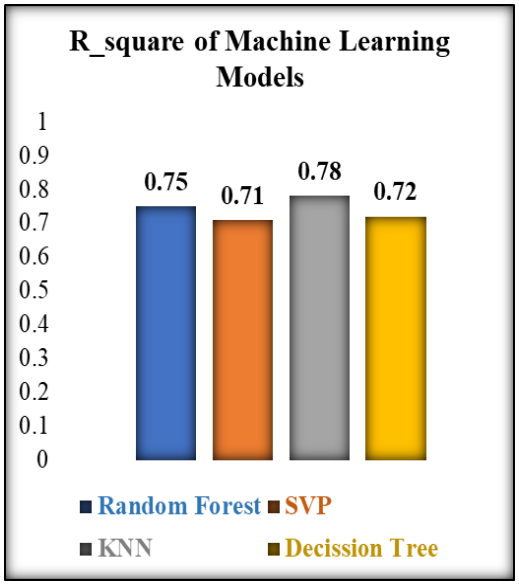


figure 11

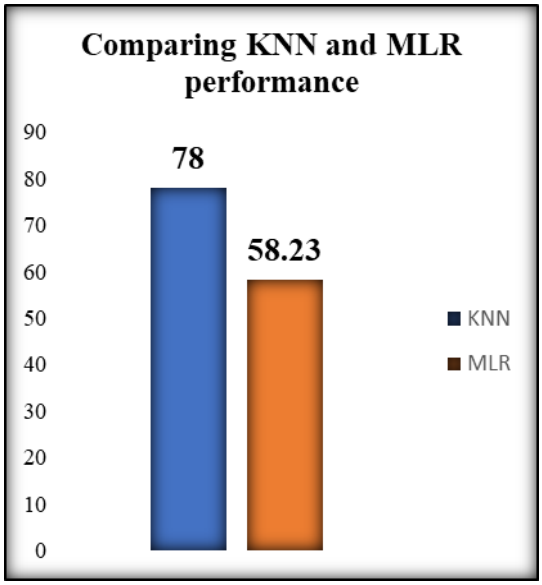


figure 12

3. Discussion

To our knowledge, this is the first time that spirometry data has been investigated in an industrial area in district Sheikhupura to estimated HCV values using ML-regression models. Figure 5 and 6 showed a clear difference between the performance of difference algorithms. As shown in table 5, The accuracy of 4 ML-regression techniques. Every parameter was analyzed by proposed ML-regression techniques. And the best performance model was selected to complete the evaluation. As described in table 5, the evaluation of the best performance model depended on the effect of the R_square. K-nearest neighbor is proved to be the best model on the basis of its performance measures: R_square: 78%, MSE: 0.246, RMSE: 0.496, MAE: 0.398 and MAPR: 15.13%. In

our data, results from K-nearest neighbor showed a higher R-squared value for FVC than other applied ML-regression techniques and also proved to be an improved model as compared to multiple linear regression (78% and 58.23% respectively).

4. Conclusion

Primary care physicians can quickly and accurately identify patients who are likely to have a restrictive lung disease by using our algorithm to predict FVC from spirometry data. For instance, the ML algorithm might notify the patient and their healthcare provider that the patient needs to see a doctor for additional assessment and possible treatment if the patient's predicted FVC drops below a predetermined threshold.

In addition to ensuring that patients receive timely and appropriate care, this can help to avoid delays in diagnosis. This study has several significant limitations, despite the fact that our method for model construction and parameter tuning benefited from a large and diverse data set. First, the majority of the patients in our ML models' training and testing were Caucasian industrial patients. As a result, it is impossible to guarantee that the model will function similarly on other populations. Secondly, whole body measurements were used to obtain most of the FVC measurements in the training data.

It has been demonstrated that this approach, which is frequently regarded as the gold standard, overestimates FVC in patients with obstructive diseases (O'Donnell et al., 2010; Tang et al., 2016). The developed MLR model and the developed existing machine-learning models outperform the experimental data, according to the study's findings. These results suggest that using ML-models as the basis for a clinical model can be beneficial in predicting FVC. To lessen the size of the intercepts, more work with these models is required.

References

Alam, M. Z., Simonetti, A., Brillantino, R., Tayler, N., Grainge, C., Siribaddana, P., ... & Rezwan, F. I. (2022). Predicting pulmonary function from the analysis of voice: a machine learning approach. *Frontiers in digital health*, *4*, 750226.

Alqahtani, J. S. (2022). Prevalence, incidence, morbidity and mortality rates of COPD in Saudi Arabia: Trends in burden of COPD from 1990 to 2019. *PLoS One*, *17*(5), e0268772.

Bolourani, S., Wang, P., Patel, V. M., Manetta, F., & Lee, P. C. (2020). Predicting respiratory failure after pulmonary lobectomy using machine learning techniques. *Surgery*, *168*(4), 743-752.

do Amaral, J. L. M., & de Melo, P. L. (2020). Clinical decision support systems to improve the diagnosis and management of respiratory diseases. In *Artificial intelligence in precision health* (pp. 359-391). Academic Press.

Erdal, M. (2021). Chronic obstructive pulmonary disease (COPD): exacerbations and costs.

Hashemi, S. Y., Momenabadi, V., Faramarzi, A., & Kiani, A. (2020). Trends in burden of chronic obstructive pulmonary disease in Iran, 1995–2015: findings from the global burden of disease study. *Archives of Public Health*, *78*, 1-7.

Ioachimescu et al. "Estimating FVC from FEV2 and FEV3: Assessment of a surrogate spirometric parameter." Chest128.3(2005):1274-1281

Kavitha et al. "Prediction of forced expiratory volume (FEV6) in flow volume spirometry using support vector machines and radial basis function neural networks." Journal of Mechanics in Medicine & Biology 10.04(2010): 683-693.

Kristensen, K., Olesen, P. H., Roerbaek, A. K., Nielsen, L., Hansen, H. K., Cichosz, S. L., ... & Hejlesen, O. (2023). Using random forest machine learning on data from a large, representative cohort of the general population improves clinical spirometry references. *The Clinical Respiratory Journal*, *17*(8), 819-828.

Licia Pezzoli et al. "Quality of spirometric performance in older people. " Age & Ageing 32.1(2002): 43-46.

Lin, C. H., Cheng, S. L., Chen, C. Z., Chen, C. H., Lin, S. H., & Wang, H. C. (2023). Current progress of COPD Early detection: key points and novel strategies. *International Journal of Chronic Obstructive Pulmonary Disease*, 1511-1524.

Maldonado-Franco, A., Giraldo-Cadavid, L. F., Tuta-Quintero, E., Bastidas Goyes, A. R., & Botero-Rosas, D. A. (2023). The Challenges of Spirometric Diagnosis of COPD. *Canadian Respiratory Journal*, *2023*(1), 6991493.

Mengesha, Y. A., & Mekonnen, Y. (1985). Spirometric lung function tests in normal non-smoking Ethiopian men and women. *Thorax*, *40*(6), 465-468.

Mythili et al. "Prediction of FEV1 and FEV6 in normal and obstructive abnormality using ELM regression and spirometric investigations." International Conference on Informatics 2015.

Sharma, S. K., Narain, J. P., & Mohan, A. (2023). Chronic Obstructive Lung Diseases. In *Global Health Essentials* (pp. 163-167). Cham: Springer International Publishing.

Schermer TR et al. "Validity of spirometric testing in a general practice population of patients with chronic obstructive pulmonary disease (COPD)." Thorax 58.10(2003): 861-6.

Wang, C., Chen, X., Zhao, R., He, Z., Zhao, Z., Zhan, Q., ... & Fang, Z. (2019). Predicting forced vital capacity (FVC) using support vector regression (SVR). *Physiological measurement*, *40*(2), 025010.

Yao, L. P., & Tao, R. (2021). Does Chronic Obstructive Pulmonary Disease Affect Workers' Health? *Frontiers in Public Health*, *9*, 711629.

Subramaniyan, V., Fuloria, S., Sekar, M., Shanmugavelu, S., Vijeepallam, K., Kumari, U., ... & Fuloria, N. K. (2023). Introduction to Lung Disease. In *Targeting Epigenetics in Inflammatory Lung Diseases* (pp. 1-16). Singapore: Springer Nature Singapore.

Singh, A., Prakash, N., & Jain, A. (2023). A review on prevalence of worldwide COPD situation. *Proceedings of Data Analytics and Management: ICDAM 2022*, 391-405.

Ulmer WT. "Lung function - Clinical importance, problems and new results." Journal of Physiology and Pharmacology, 54.1(2005): 11-13.

Jaakkola, J. J., Hernberg, S., Lajunen, T. K., Sripaijboonkij, P., Malmberg, L. P., & Jaakkola, M. S. (2019). Smoking and lung function among adults with newly onset asthma. *BMJ open respiratory research*, *6*(1), e000377.

Kumar, J., Khuhawar, S. M., Chandio, B. A., Sanghro, K. A., & Channa, S. M. (2022). Effect of Smoking on Chest Expansion, Respiratory and Lung Functions. *Pakistan Journal of Medical & Health Sciences*, *16*(06), 340-340.

Enayatjazi, M., Enayatjazi, S., Ghadiri, N., & Fathi, A. (2023). Effect of 8 weeks of Endurance Training on Volume and Capacity of Lungs (FEV1/FVC, FEV1, FVC) and Its Relationship with BMI in Male Smoking students. *New Approaches in Exercise Physiology*, *5*(9), 32-43.

Khan, F. A., Stewart, I., Moss, S., Fabbri, L., Robinson, K. A., Johnson, S. R., & Jenkins, R. G. (2022). Three-month FVC change: a trial endpoint for idiopathic pulmonary fibrosis based on individual participant data meta-analysis. *American journal of respiratory and critical care medicine*, *205*(8), 936-948.

Miller MR et al. "Standardisation of spirometry." European Respiratory Journal 26.2(2005): 319-338

Tang Y, Zhang M, Feng Y, Liang B. The measurement of lung volumes using body plethysmography and helium dilution methods in COPD patients: a correlation and diagnosis analysis. Sci Rep. (2016) 6:37550. doi: 10.1038/srep37550

O'Donnell CR, Bankier AA, Stiebellehner L, Reilly JJ, Brown R, Loring SH. Comparison of plethysmographic and helium dilution lung volumes: which is best for COPD? Chest. (2010) 137:1108–15. doi: 10.1378/chest.09-1504

Roy, A., Al Zubayer, T., Tabassum, N., Islam, M. N., & Sattar, M. A. (2022). CurFi: An automated tool to find the best regression analysis model using curve fitting. *Engineering Reports*, *4*(12), e12522.

Viering, T., & Loog, M. (2022). The shape of learning curves: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7799-7819.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Kamel, A. R., & Abonazel, M. R. (2023). A simple introduction to regression modeling using R. *Computational Journal of Mathematical and Statistical Sciences*, *2*(1), 52-79.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, 587-604.

Bernard, S., Heutte, L., & Adam, S. (2009, June). On the selection of decision trees in random forests. In *2009 International joint conference on neural networks* (pp. 302-307). IEEE.

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In *Machine learning* (pp. 123-140). Academic Press.

Mishra, P. (2021). Explainability for Non-Linear Models. In *Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks* (pp. 93-127). Berkeley, CA: Apress.

Boswell, D. (2002). Introduction to support vector machines. *Departement of Computer Science and Engineering University of California San Diego*, *11*, 16-17.

Günlük, O., Kalagnanam, J., Li, M., Menickelly, M., & Scheinberg, K. (2021). Optimal decision trees for categorical data via integer programming. *Journal of global optimization*, *81*, 233-260.

Oswald, R. K., Scherer, W. T., & Smith, B. L. (2000). *Traffic flow forecasting using approximate nearest neighbor nonparametric regression* (No. UVA-CE-ITS_01-4).

Sabry, F. (2023). *K Nearest Neighbor Algorithm: Fundamentals and Applications* (Vol. 28). One Billion Knowledgeable.