International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:9, No:4, 2015

Comparison of Multivariate Adaptive Regression Splines and Random Forest Regression in Predicting Forced Expiratory Volume in One Second

P. V. Pramila, V. Mahesh

***Abstract*—**Pulmonary Function Tests are important non-invasive diagnostic tests to assess respiratory impairments and provides quantifiable measures of lung function. Spirometry is the most frequently used measure of lung function and plays an essential role in the diagnosis and management of pulmonary diseases. However, the test requires considerable patient effort and cooperation, markedly related to the age of patients resulting in incomplete data sets. This paper presents, a nonlinear model built using Multivariate adaptive regression splines and Random forest regression model to predict the missing spirometric features. Random forest based feature selection is used to enhance both the generalization capability and the model interpretability. In the present study, flow-volume data are recorded for N= 198 subjects. The ranked order of feature importance index calculated by the random forests model shows that the spirometric features FVC, FEF_{25}, PEF, FEF_{25-75}, FEF_{50} and the demographic parameter height are the important descriptors. A comparison of performance assessment of both models prove that, the prediction ability of MARS with the `top two ranked features namely the FVC and FEF_{25} is higher, yielding a model fit of R^2= 0.96 and R^2= 0.99 for normal and abnormal subjects. The Root Mean Square Error analysis of the RF model and the MARS model also shows that the latter is capable of predicting the missing values of FEV_1 with a notably lower error value of 0.0191 (normal subjects) and 0.0106 (abnormal subjects) with the aforementioned input features. It is concluded that combining feature selection with a prediction model provides a minimum subset of predominant features to train the model, as well as yielding better prediction performance. This analysis can assist clinicians with a intelligence support system in the medical diagnosis and improvement of clinical care.

***Keywords*—**FEV_1, Multivariate Adaptive Regression Splines Pulmonary Function Test, Random Forest.

I. Introduction

PULMONARY disease changes the physiology of the lungs which manifests as changes in respiratory mechanics [1]. Impairment of Pulmonary function is a significant factor that may affect quality of life and can lead to increased risks of premature morbidity and mortality [2]-[4]. The assessment of respiratory function and mechanics is of crucial importance to understand the pathophysiology of the disease and to provide guidelines for therapeutic measures.

Pulmonary function tests (PFT) are non invasive diagnostics, provide quantifiable measures of lung function and are used to evaluate and monitor pulmonary function abnormality. Spirometry is the often used choice of PFT in assessment of airways impairment by measuring forced expiratory volumes and flow rates of respiratory system. These measurements enable physician to diagnose the presence and severity of airway obstruction, lung cancer, coronary artery disease and stroke [5]–[8]. It is an essential investigation method for diagnosis and management of lung diseases like asthma and Chronic Obstructive Pulmonary Disease (COPD) [9].

The significant flow volume parameters recorded with the spirometer are Forced Vital Capacity (FVC), Forced Expiratory Volume in one second (FEV_1), Forced expiratory flow at 50% (FEF_{50}) of forced vital capacity and Forced Expiratory Flow at 25% of FVC (FEF_{25}). The FVC is the maximal amount of air that the patient can forcibly exhale after taking a maximal inhalation. FEV_1 is forced expiration volume in 1 second during forced exhalation and is universally used as an important marker of asthma and a measure of severity in COPD [10].

However the test requires patient effort and co-operation and in particular elderly patients are at risk for misdiagnosis due to low performance expectations which meets the ATS criteria [11], [12]. Such misclassification may lead to inappropriate treatment and increased use of acute healthcare services [13], [14]. For this reason, reliable prediction of significant spirometric parameters such as FEV_1 using expert systems is essential to overcome this fuzziness in physicians' interpretation.

Feature subset selection technique for reducing the attribute space of a feature set, has shown to be very effective in increasing efficiency in learning tasks [15], [16]. Random forest (RF) [17] widely used in many research fields for its improved prediction and ability to provide variable importance measures that can be used to identify the most important predictor variables. Random forest is an ensemble of decision trees encompassing the advantages of low bias, ease of interpretation of variables in decision trees. The problem of over fitting in decision trees is overcome by averaging the outcomes across different decision trees [18]. RF has been used in numerous biological applications comprising of identification of cancer biomarkers [19], cardiac arrhythmia diagnosis [20] and predictive models to diagnose asthma patients based on respiratory sound signals [21]. Random forest is well-liked to identify the feature that contribute most to prediction using importance scores, but relatively a little

P. V. Pramila is with Dhanalakshmi College of Engineering, Chennai, India (e-mail: pvpramila@gmail.com).

V. Mahesh is Associate Professor with Department of Biomedical Engineering, SSN college of Engineering, Chennai, India (corresponding author, e-mail: maheshv@ssn.edu.in)

research has been done on predictive models exploring their effectiveness in the contribution of predictive accuracy [22].

Statistical modelling technique, the Multivariate adaptive regression splines (MARS) is used to fit an interpretable model and study the involvement of feature importance in prediction accuracy. MARS a non parametric and flexible modelling approach [23] and can be considered as generalization of Classification and Regression Trees. The regression model is constructed by fitting spline basis function to distinct intervals of the independent variables [24]. The occurrence of disjoint sub regions and discontinuity of the approximating function at the boundaries of the intervals as in binary recursive partitioning has been eliminated in MARS [25]. In recent years MARS has been increasingly applied in bioinformatics that includes predictive modelling of binary outcomes [26], nonlinear modelling of time series analysis and disease risk research [27].

The objective of this analysis is to build an expert system based model for the prediction of missing spirometric parameter Forced expiration in one second. This study comprises of two stage hybridized approach: In the first stage, the learning algorithm random forest is employed to estimate feature ranking and to obtain a subset of significant spirometric features. In stage two, performance assessment of RF and MARS based predictive models is evaluated on the selected spirometric feature set that generalizes well when tested with an independent validation data set.

II. Materials and Methods

The spirometer recordings are carried out on adult volunteers (N = 198) for the present study. The portable InspireX spirometer with a gold standard transducer is used for the acquisition of the data. The patients were advised on the test procedure and instructed forced expirations by trained technician. PFT was performed in a sitting position according to American Thoracic Society (ATS) guidelines [22]. Spirometry data were recorded electronically from the subjects along with their demographic parameters which include age, weight, height and smoking history.

Random forest, is an ensemble many individual decision trees that combines Breiman's [11] bagging idea and random selection of features, to construct the collection of decision trees using a two stage randomization procedure. Each tree is built on a bootstrap sample of the input spirometric training dataset introducing the first level of randomization. With the bootstrap sampling, each tree in the forest is constructed with a different subset of the training data resulting in a collection of different trees. A second layer of randomization is introduced at the node level. A subsample of the spirometric features is used at each tree node split, introducing further variation among trees. This two step randomization de-correlates the trees, so that the forest ensemble will have low variance. For a tree grown on a bootstrap data, the remaining training data called the out of bag (OOB) data is used as test set for that tree. Further, the OOB data is also used to estimate the importance index of each spirometric feature. The ranking of features based on the importance index is then used to build a predictive model. The accuracy of the random forest's prediction can be estimated from the OOB data as [26]

$$OOB_{MSE(t)} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_{k\,OOB}})^2 \quad (1)$$

where y_k is the prediction for the k[th] observation and $\widehat{y_{k\,OOB}}$ denotes the average prediction for the k[th] observation from all trees for which this observation has been the out of bag data.

MARS regression model is constructed by fitting piecewise linear basis function for each input feature x_i at distinct intervals. These functions are called hinge function $h(x)$ and is of the form $\max(0, x_i - t)$, $\max(0, t - x_i)$. The joining points of the piecewise polynomials are the knots or nodes (t). MARS uses two-sided truncated power functions as spline basis function described by [25],

$$[-(x-t)]_+^q = \begin{cases} (t-x)^q & if\ x < t \\ 0 & otherwise \end{cases} \quad (2a)$$

$$[-(x-t)]_+^q = \begin{cases} (x-t)^q & if\ x >= t \\ 0 \end{cases} \quad (2b)$$

where q is the power to which the splines are raised and which determines the degree of smoothness of the resultant function estimate. The final MARS model of has the form

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \quad (3)$$

where $\hat{y}$ is the predicted response, x is the explanatory variable, a_0 and a_m are estimated coefficients to yield the best fit of data, M is the number of basis functions included into the model and $B_m(x)$ is the m[th] basis function.

III. Results and Discussion

In the present study, flow-volume data are recorded for N= 198 subjects and a total of 9 parameters are derived from them. The statistical analysis such as mean and standard deviation on the spirometric pulmonary function parameters are presented in Table I. The mean values of significant parameters such as FVC, FEV_1, PEF and FEF_{25-75}% in normal subjects are distinctly higher than that of the abnormal cases.

TABLE I
Spirometric Pulmonary Data Description of Attributes

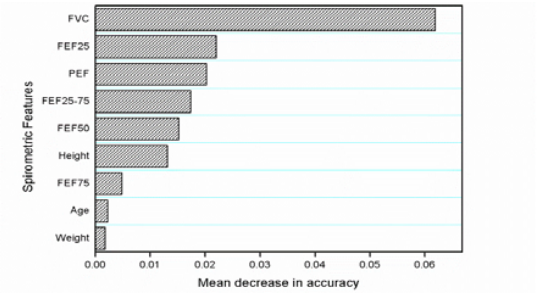
Attribute description	Normal Subjects Mean ±SD	Abnormal Subjects Mean ± SD
Forced Vital Capacity (FVC)	2.63 ± 0.74	1.95 ± 0.70
Forced Expiratory flow at 25% (FEF_{25})	4.02 ± 1.36	3.30 ± 1.59
Forced expiratory flow at 50% (FEF_{50})	2.54 ± 0.92	2.02 ± 0.94
Forced expiratory flow at 25% - 75% (FEF_{25-75})	2.34 ± 0.85	1.85 ± 0.83
Peak Expiratory flow (PEF)	5.24 ± 1.79	4.22 ± 1.94
Height(cms)	152.71 ± 12.60	153.13 ± 13.50
Weight(Kgs)	54.08 ± 15.88	57.27 ± 17.90

To obtain the ordered list of features, the feature importance index calculated by the Random forests model was recorded. The procedure was repeated 10 times due to the stochastic

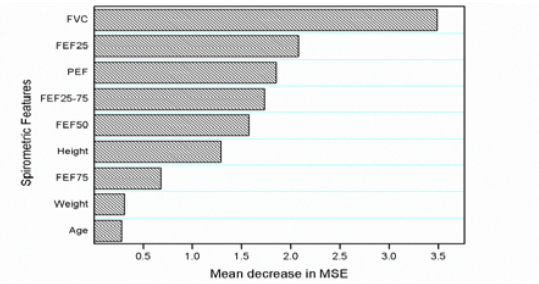
nature of the random forest approach and averaged for each feature. The averaged values of a total of 9 spirometric features were then sorted and ranked in descending order (Figs. 1 (a) & (b)).

The results depicts that the spirometric features FVC, FEF $_{25}$, PEF, FEF $_{25-75}$, FEF_{50} and the demographic parameter height are the important descriptors while the other features are moderately important. Random Forest and MARS predictive model were fitted with aforementioned dominant spirometric features where each feature was included in an incremental order based on their ranking and the statistical results of their performance were analysed.

The coefficient of determination R^2 by the model and the Root mean squared error (RMSE) are performed to determine the external predictive ability of the model. The coefficient of determination R^2 very close to 1 exhibit a well-fitted regression model with prediction outcome close to the observed data values .The results of statistical analysis obtained (Tables II and III) show that both the methodologies presented high values of coefficient of determination $R^2 > 0.90$ in the prediction of the spirometric parameter Forced expiration in one second.



(a) Mean decrease in accuracy



(b) Mean decrease in Mean square error

Fig. 1 Bar plots depicting the featuring ranking of input spirometric features

TABLE II
ASSESSMENT OF FEATURE RELEVANCE USING RANDOM FOREST ALGORITHM AND MULTIVARIATE ADAPTIVE REGRESSION SPLINES OF NORMAL SUBJECTS

S.No	Features	RF			MARS		
		R^2	RMSE	F-value	R^2	RMSE	F-value
1	FVC, FEF $_{25}$	0.92	0.0222	0.8894	0.96	0.0191	1.0742
2	FVC,FEF_{25}, PEF	0.92	0.0231	2.32136	0.96	0.0200	1.0760
3	FVC,FEF_{25},PEF,FEF_{25-75},	0.94	0.0224	2.6485	0.96	0.0135	0.9905
4	FVC,FEF_{25},PEF,FEF_{25-75}, FEF_{50}	0.95	0.0219	1.8156	0.98	0.0135	0.9905
5	FVC,FEF_{25},FEF_{50},PEF, FEF_{25-75}, Height	0.95	0.0219	2.388	**0.98**	0.0135	0.9905

TABLE III
ASSESSMENT OF FEATURE RELEVANCE USING RANDOM FOREST ALGORITHM AND MULTIVARIATE ADAPTIVE REGRESSION SPLINES OF ABNORMAL SUBJECTS

S.No	Features	RF			MARS		
		R^2	RMSE	F-value	R^2	RMSE	F-value
1	FVC, FEF $_{25}$	0.98	0.0192	3.4923	0.99	0.0106	1.1041
2	FVC,FEF_{25}, PEF	0.98	0.2052	1.57034	0.99	0.0114	1.1017
3	FVC,FEF_{25}, PEF,FEF_{25-75},	0.99	0.0185	1.5860	0.99	0.0082	1.1068
4	FVC,FEF_{25}, PEF,FEF_{25-75}, FEF_{50}	0.98	0.0229	1.4599	0.99	0.0082	1.1084
5	FVC,FEF_{50},PEF, FEF_{25-75}, Height	**0.99**	0.0192	1.6035	**0.99**	0.0082	1.1084

Hence the performance of both the regression models is robust even in the presence of highly correlated input features. While comparing the prediction ability of individual models, it is observed that the MARS model outperforms the Hence the performance of both the regression models is robust even in the presence of highly correlated input features. While comparing the prediction ability of individual models, it is observed that the MARS model outperforms Random forest model for both normal as well as the abnormal subjects. The results show a higher value of R^2= 0.96 and R^2= 0.99 with the first two ranked input features namely the FVC and FEF $_{25.}$

The computed root mean square error value between the predicted and observed FEV_1 values is plotted in Fig. 2 for the normal subjects and in Fig. 3 for the abnormal subjects. In both the models it is observed that lowest error rate is obtained with the input features FVC, FEF_{25}, PEF and $FEF_{25-75.}$ The error analysis of the RF model and the MARS model also show that the latter is capable of predicting the missing values of FEV_1 with a notably lower error value of RMSE = 0.0191 for normal subjects and RMSE = 0.0106 for abnormal subjects with a minimal subset of two input features. Hence the MARS model generalizes well even with a high dimensional dataset and in the presence of highly correlated variables predicting near accurate values of FEV_1. It is then concluded that

combining feature selection step with a prediction model we can obtain a minimum subset of important features to train a faster and more robust model, yielding better prediction performance.

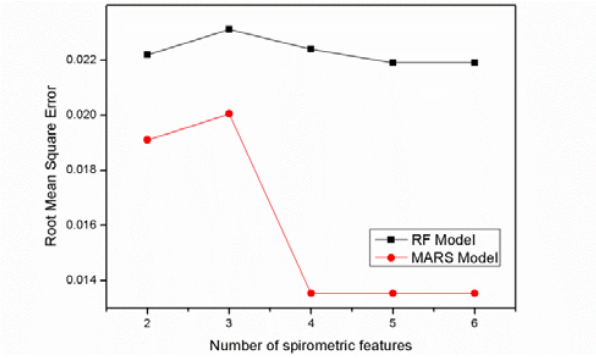


Fig. 2 Variations in Root mean square error with varying number of input features for normal subjects

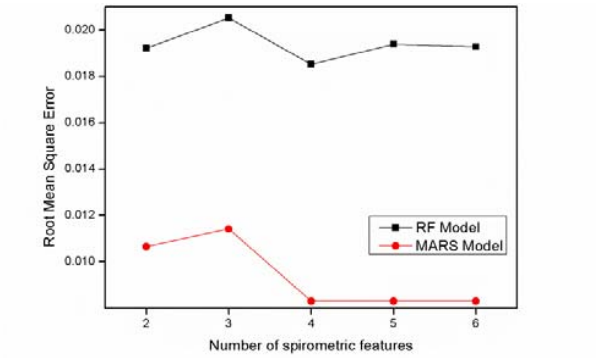


Fig. 3 Variations in Root mean square error with varying number of input features for abnormal subjects

IV. Conclusion

Spirometry is the most frequently performed clinical pulmonary function test to assess the dynamics of respiratory system in human subjects. It evaluates lung disease measuring airflow that moves air out of the lungs after taking the deepest breath possible. Spirometric tests may sometimes result in incomplete data set due to lack of ability to follow instructions of the test in particular patients with lung impairment.

In this work, investigation to apply methodology to able to reduce the feature space while increasing the model prediction capabilities and hence reducing the redundancy and correlation between variables is analyzed. The performance of two non- linear regression strategies Random forest (RF) and Multivariate adaptive regression splines for the prediction of most significant pulmonary function parameter Forced Expiratory Volume in one second (FEV_1) is presented. Flow volume spirometric parameters for N=198 inclusive of both normal and abnormal subjects was considered for analysis. It is observed that the MARS model outperforms the Random forest model with higher value of R^2= 0.96 and R^2= 0.99 for the top two ranked input features namely the FVC and FEF_{25}. A significant outcome of this model is its capability to predict the missing FEV_1 values of the subjects with minimal subset of input features, yet achieving a comparable result with [3] evidencing the benefits of feature selection performed by the RF model. The analysis provides a better understanding of the underlying relationship of the feature space, improved prediction performance with descriptors diminished in their numbers. This improved hybridized feature selection and prediction model based decision support system can also aid in clinical care tests enhancing medical diagnosis for pulmonary impairments.

References

[1] Daniel C Ginnan and Jonathon Dean Truwit, "Clinical review: Respiratory mechanics in spontaneous and assisted ventilation," Critical Care, vol. 9, no.5, pp. 472–484, 2005.
[2] R. L. Mulder, N. M. Thonissen, J. H. H. Vander Pal, P. Bresser, W. Hanselaar, C. C. E. Koning, F. Oldenburger, H. A. Heij, H. N. Caron, "Pulmonary function impairment measured by pulmonary function tests in lon g-term survivors of childhood cancer," Thorax, vol. 66, pp. 1065-1071, 2011.
[3] A. Mythili, C. M. Sujatha , S. Srinivasan and S. Ramakrishnan, "Prediction Of Forced Expiratory Volume In Spirometric Pulmonary Function Test Using Adaptive Neuro Fuzzy Inference System," Biomedical Sciences Instrumentation, vol. 48, pp.508-15, 2012.
[4] D. Ozerkis-Antin, J. Evans, A. Rubinowitz, R.J. Horner, R.A. Matthay, "Pulmonary manifestations of rheumatoid arthritis," Clinical Chest Medicine, vol.31, no.3, pp. 451-78, 2010.
[5] Thomas A Barnes, Len Fromer, "Spirometry use: detection of chronic obstructive pulmonary disease in the primary care setting," International Journal of CODP, 2011.
[6] R.E. Dales, K.L. Vandemheen, J. Clinch, et al. "Spirometry in the primary care setting: influence on clinical diagnosis and management of airflow obstruction," Journal of Chest, vol.128, no. 4, pp. 2443–2447, 2005.
[7] N. Chavannes, T. Schermer, R. Akkermans, et al. "Impact of spirometry on GPs' diagnostic differentiation and decision-making," Respiratory Medicine, vol.98, no.11, pp.1124–1130, 2004.
[8] R.P.Young, R. Hopkins, T.E. Eaton, "Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes," European Respiratory Journal, vol. 30, no.4, pp.616–622, 2007.
[9] "Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease," American Thoracic Society, American Journal of Respiratory and Critical Care Med, vol.152, pp.77-121, 1995.
[10] D.C. Richter , J.R. Joubert , H. Nell, M.M. Schuurmans, E.M. Irusen, "Diagnostic value of post-bronchodilator pulmonary function testing to distinguish between stable, moderate to severe COPD and asthma, International journal of chronic obstructive pulmonary disorder, vol. 3, no.4, pp. 693-699, 2008.
[11] Jeffrey M. Haynes, "Pulmonary Function Test Quality in the Elderly: A Comparison with Younger Adults," Respiratory care, vol.59, no.1, jan 2014.
[12] American Thoracic Society, Standardization of spirometry: a summary of recommendations from the American Thoracic Society. 1987 update, Ann Intern Med, vol.108, pp. 217–220, 1988.
[13] V. Bellia, R. Pistelli, F. Catalano, R. Antonelli-Incalzi, V. Grassi, G. Meillo, et al . "Quality control of spirometry in the elderly: the SARA study," Am J Respir Crit Care Med, vol. 161, no.4, pp. 1094-1100, 2000.
[14] L. Pezzoli, G. Giardini, S. Consonni , I. Dallera, C. Bilotta, G. Ferrario G et al. "Quality of sprirometric performance in older people. Age Ageing," vol. 32, no. 1, pp. 43-46, 2003.
[15] Xu, Ruo, "Improvements to random forest methodology," Graduate Thesis and Dissertations, Paper 13052, 2013.
[16] Mark R. Segal, "Machine Learning Benchmarks and Random Forest Regression," Kluwer Academic Publishers, 2003.
[17] L. Breiman, "Random forests," Machine Learning, vol. 45, 2001, pp. 5–32.
[18] Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, Inke R. Konig, "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics," available at: http://epub.ub.uni-muenchen.de/13766/1/TR.pdf

[19] M. Hilario, A. Kalousis, C. Pellegrini, M. Muller, "Processing and classification of protein mass spectra, Mass Spectrom Rev, vol.25, pp. 409-449.
[20] Akin Özçift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," Computers in Biology and Medicine, vol.41, no.5, pp.265-271, 2011.
[21] Benjamin A Goldstein, Alan E Hubbard, Adele Cutler and Lisa F Barcellos, "An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings," BMC genetics, vol.11, no.49, 2010.
[22] Nahit Emanet, Halil R Öz, Nazan Bayram and Dursun Delen, "A comparative analysis of machine learning methods for classification type decision problems in healthcare," Decision Analytics, vol.1, no.6, pp.1-20, 2014.
[23] K. J. Archer and R.V. Kimes, "Empirical characterization of random forest variable importance measures," Computational Statistics and Data Analysis, vol. 52, pp. 2249-2260.
[24] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat., vol. 19, pp. 1–141, 1991.
[25] Dani Guzmán, Francisco Javier de Cos Juez, Fernando Sánchez Lasheras, Richard loop adaptiev Myers and Laura Young, "Deformable mirror model for open- optics using multivariate adaptive regression splines," Optics Express, vol.18, no.7, pp. 6492 – 6505, 2013.
[26] P. A. W. Lewis and J. G. Stevens, "Nonlinear modeling of time series using multivariate adaptive regression splines (mars)," Journal of the American Statistical Association, vol. 86, no. 416, pp. 864-877, 1991.
[27] Peter C. Austin, "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality," Statistics in Medicine, vol. 26, pp. 2937–2957,