Journal of Computational Science 63 (2022) 101768

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Classification of obstructive and non-obstructive pulmonary diseases on the basis of spirometry using machine learning techniques

Sudipto Bhattacharjee [a], Banani Saha [a], Parthasarathi Bhattacharyya [b], Sudipto Saha [c,*]

[a] Department of Computer Science and Engineering, University of Calcutta, JD-2, Sector-III, Salt Lake, Kolkata 700098, India
[b] Institute of Pulmocare and Research, DG-8, Action Area-1, New Town, Kolkata 700156, India
[c] Division of Bioinformatics, Bose Institute, EN 80, Sector V, Bidhan Nagar, Kolkata 700091, India

ARTICLE INFO

Keywords:
Spirometry
Pulmonary diseases
Random forest
Support vector machine
Naive Bayes
Neural network

ABSTRACT

Background: The symptomatic similarities between the two categories of pulmonary diseases, obstructive and non-obstructive, make the early diagnosis difficult for clinicians. Spirometry is a popular lung investigation that is performed in the early diagnostic stages to understand the mechanics of lungs. This work aims to develop machine learning models to classify obstructive and non-obstructive pulmonary diseases on the basis of spirometry data.

Method: Supervised learning models were developed with support vector machine (SVM), random forest (RF), Naive Bayes (NB) and multi-layer perceptron (MLP) algorithms. Models were trained with spirometry data of 1163 patients using 5-fold cross validation (CV) and further validated with a blind dataset of 151 patients for external validation.

Results: The MLP model performed optimally with an accuracy of 83.7% and Matthew's correlation coefficient of 0.682 with 5-fold CV. All the models performed well while validating the blind dataset. The disease-specific prediction of COPD and DPLD, as obstructive and non-obstructive respectively, achieved ~90% accuracy in the training dataset. The MLP model was stored in a web server for use in a web application.

Conclusions: The machine learning models were able to predict obstructive and non-obstructive pulmonary diseases with good accuracy, based on spirometry data. The web application can be used by clinicians and patients as a tool for early prediction.

1. Introduction

Pulmonary diseases are leading causes of death and disability in the world with 4 million premature deaths every year globally [1]. The decline in air quality with the rise of pollution makes people more vulnerable to lung diseases. Most lung diseases are chronic in nature and their treatment incurs a financial burden on the patient. The treatment mostly aims to cause symptomatic relief for diseases which are largely irreversible, like chronic obstructive pulmonary disease (COPD). Thus, an early diagnosis of such diseases is very important. Pulmonary diseases can be classified into two categories - obstructive and non-obstructive. The obstructive diseases are characterized by airflow limitations caused by structural and/or functional changes in the airway wall or lumen. The functional changes are mostly derived from airway hyper-responsiveness secondary to incitement by inhalation of aero-allergens or noxious agents or both [2]. The impact is influenced by several factors including genetic predisposition and host microbial interactions. Asthma is an obstructive airway disease developed from heightened airway hyper-responsiveness frequently from aero-allergens. It results in symptoms like wheezing, breathlessness, chest tightness and coughing, often more frequently at night or/and in the early morning. COPD is another disease of similar airflow limitation with similar symptoms but a different pathogenesis of the inflammation that is often progressive and accompanied by features of systemic involvement. There are other diseases related to airway involvement and airflow obstruction but asthma and COPD together with their proposed overlap state (ACOS: Asthma-COPD Overlap Syndrome).

Beyond the airways on their continuations the part of the lung that

* Correspondence to: Division of Bioinformatics, Bose Institute, Unified Academic Campus, Bose Institute, EN 80, Sector V, Bidhan Nagar, Kolkata 700091, WB, India.
E-mail addresses: ttsudipto@gmail.com (S. Bhattacharjee), bsaha_29@yahoo.com (B. Saha), parthachest@yahoo.com (P. Bhattacharyya), ssaha4@jcbose.ac.in, ssaha4@gmail.com (S. Saha).

https://doi.org/10.1016/j.jocs.2022.101768
Received 19 May 2022; Accepted 4 July 2022
Available online 6 July 2022

takes part in gaseous exchange is called the lung parenchyma. Diseases involving lung parenchyma primarily can be described as non-obstructive lung diseases. The category also involves diseases of the pleura (the sheath of the lungs), the pulmonary vessels and also some disorders related to the neuromuscular control of breathing and ventilation. The classical example of parenchymal lung disease is DPLD (diffuse parenchymal lung disease). The non-obstructive diseases of parenchymal or pleural origin result in impediment to pulmonary expansion and result in reduction of lung volume [3]. DPLD, a group of heterogeneous diseases with some common impact on physiology, is the most common parenchymal lung disease. They show mostly non-infectious and non-malignant diffuse interstitial scarring to result in ventilation-perfusion mismatch and impediment in gaseous exchange. The common causes of DPLD are chronic hypersensitivity pneumonitis, pulmonary sarcoidosis, CTD-ILD (connective tissue disease related interstitial lung disease), idiopathic interstitial pneumonias, occupational inorganic dust related lung diseases and others. All of them lead to restriction of the lung expansion and some of them also show an element of obstruction.

The diagnosis of these two classes of lung diseases is often difficult as both the obstructive and non-obstructive lung diseases often share similar symptoms. Moreover, this important class specific diagnosis (obstructive vs non-obstructive) may be the first step in finalizing a therapeutic regimen as there are enormous number of conditions that can fit largely (not perfectly) in each group. Moreover, apart from the complexity, the standard interpretations of the commonly used tests are subjective to a large extent, and they show various degrees of accuracy with experience and involvement of the interpreter, time given for interpretation, and other factors that often remain the reasons of inadequacies and mistakes. These subjective components can only be reduced or removed if the process is executed through artificial intelligence. Hence, the diagnostic predictions, based on the features from an investigation, by application of AI and subsequently by an expert may prove to be efficient.

The aim of the study is to apply machine learning (ML) on spirometry for classification of obstructive and non-obstructive pulmonary diseases. Spirometry is a popular non-invasive test used for the purpose [4]. It is simple to perform, fairly inexpensive for the patient and does not require massive infrastructure.

A lot of work is being done in recent times to predict pulmonary diseases using ML. HRCT images of lungs are used to detect emphysema [5]. Deep neural nets are used to detect interstitial lung abnormalities on the basis of HRCT [6]. Chest X-ray images are used to detect pneumonia and pulmonary tuberculosis [7,8]. Bronchoscopy investigation findings are used for classifying lung cancer subtypes [9]. All these imaging investigations require sophisticated apparatus and are costly for patients. In a logistically constrained scenario as in the developing world, a ML-based approach of initial classification of lung diseases using inexpensive investigation data can help to form a practical therapeutic algorithm to evaluate the patients. This may also prove useful to avert unnecessary expenditure on the part of the patients.

ML algorithms are found to be effective in prediction of diseases with numerical and categorical clinical data also. Lynch CM et al. used random forest and support vector machine models to predict lung cancer based on demographic, diagnostic and procedural data [10]. Swaminathan et al. developed and compared different ML models to detect COPD exacerbations using a combination of demographic, clinical, comorbidity-related and symptomatic information of 101 patients [11]. Aneja & Lal provided an approach for fusion of Naive Bayes with neural networks to improve accuracy of predicting asthma on the basis of symptom information of over 1000 patients [12]. Mani et al. used different ML algorithms to “assess the risk of development of type 2 diabetes” using electronic medical record (EMR) data containing demographic and clinical variables of over 2000 patients [13]. Hussain et al. used support vector machine and Naive Bayes models along with other feature extraction algorithms in detection of prostate cancer on the basis of magnetic resonance imaging (MRI) scans of 682 patients [14].

The patients in this study were divided into 2 groups - Group A and B. Four popular supervised machine learning models were created for the classification. The models were trained with the spirometry data of patients in Group A. Stratified 5-fold cross validation was used. Data of patients in Group B were used as blind dataset for validation of the models.

2. Materials and methods

2.1. Creation of datasets

The data contained spirometry investigation reports of 1314 patients from Institute of Pulmocare and Research (IPCR), Kolkata diagnosed with obstructive diseases viz. asthma, chronic obstructive pulmonary disease (COPD) and non-obstructive diseases consisting of diffuse parenchymal lung disease (DPLD), obstructive sleep apnea (OSA), pulmonary sarcoidosis and chest pain. This work was approved by the Institutional Ethics Committee of IPCR, Kolkata. The patients were divided in 2 groups - Group A and Group B consisting of 1163 and 151 patients respectively. The reports of the patients diagnosed with obstructive diseases were labelled as positive and those with non-obstructive diseases were labelled as negative. The patient groups are summarized in Table 1. The summary of diagnosis of the patients in different groups is given in Fig. SF1 and Table ST1 of Supplementary material. There were 1172 spirometry reports of patients that constitute Group A. Among them, 1006 were positive and 166 were negative. There were 154 spirometry reports of patients belonging to Group B with 103 positive and 51 negative. The reports in Group A were used for training and testing with cross validation. The reports in Group B were not part of the training dataset and used as blind dataset.

2.2. Attributes of spirometry

In spirometry, patients are asked to take a maximal inspiration and then to forcefully expel air as quickly as possible into a mouthpiece [15]. The test is repeated following the administration of a bronchodilator. The pre and post bronchodilator values of the following three metrics along with the degree of bronchodilator induced changes (responsiveness) tested by spirometry were used as input.

- *Forced Vital Capacity (FVC)*: It is the volume of air exhaled forcefully and quickly after full inhalation.
- *Forced Expiratory Volume in one second (FEV1)*: It is the volume of air expired during the first second of performing the FVC test.
- *Forced Expiratory Flow (FEF25-75)*: It is the flow (or speed) of air coming out of the lung during the middle portion of a forced expiration. It is measured by taking the mean of the flow during the interval 25–75% of FVC.

The result of each test consists of 4 attributes. Thus, there are a total of 12 attributes.

Table 1
Summary of patient groups.

	Group A Used for training and testing with 5-fold cross validation	**Group B** Used as blind dataset for validation
Patient count	1163	151
Total number of spirometry reports	1172	154
Number of obstructive spirometry reports	1006	103
Number of non-obstructive spirometry reports	166	51

- *Pre-bronchodilator (Pre-BD) Value*: It is the value of the corresponding metric tested before the administration of bronchodilator.
- *Pre-BD Predicted Ratio*: It is the percent ratio of pre-bronchodilator (pre-BD) value to the predicted value. There is no normal range of values for the observed metric that is applicable to all individuals in a population. Instead, comparison is made with an expected value for a patient of a particular gender, age and physical characteristics. These are called the predicted values.
- *Post-bronchodilator (Post-BD) Value*: It is the value of the corresponding metric tested after the administration of bronchodilator.
- *Post-BD Predicted Ratio*: It is the percent ratio of post-bronchodilator (post-BD) value to the predicted value.

ANOVA *f*-test was performed to compute statistical significance of the attributes and the results are shown in Table ST-2 of Supplementary material.

2.3. Creation of machine learning models

Supervised machine learning models were developed for the classification task using Support Vector Machine (SVM) [16], Random Forest (RF) [17], Naive Bayes (NB) [18] and Multi-Layer Perceptron (MLP) [19] algorithms. The model training was performed with the spirometry data of patients in Group-A. This dataset was highly imbalanced where the positive to negative ratio (P:N) was 6:1. The imbalance often results in a skewed performance of the models with high sensitivity and low specificity values. To overcome this problem, an under-sampling method was used in which the majority (positive) class samples in Group-A were randomly divided into six disjoint (and, exhaustive) subsets. Then the minority (negative) class samples were concatenated with each positive class subset to obtain six under-sampled datasets with P:N:: 1:1. Two sets of models were trained with each algorithm: (a) one *whole dataset model* which was trained with all the samples in Group-A using stratified random 5-fold cross validation; and (b) six *under-sampled models*, each one trained with the respective under-sampled dataset using stratified random 5-fold cross validation and the performance metrics were averaged. Then, the Group-B dataset was applied to the trained models for external validation. The workflow is shown in Fig. 1. A recursive feature elimination (RFE) using 5-fold cross validation was performed, with random forest as the underlying model, to understand the importance of spirometry features. The ML models were implemented using the Python library *scikit-learn* [20].

Each ML algorithm uses several hyperparameters whose values are determined before the start of the training process. The tuning of hyperparameters was performed to improve the performance of the models using grid search technique, which is an exhaustive search using a parameter grid created by taking the cartesian product of pre-specified sets of values for each hyperparameter. Hyperparameter optimization was performed separately for both sets of models - trained with the whole training set and with the under-sampled datasets.

There were 3 hyperparameters of the RF model that were tuned - *number of trees*, *max depth* and *max features*. The radial basis function (RBF) kernel was used in the SVM models. There were 2 hyperparameters - *C* and *gamma*. For the MLP models, "adam" weight optimizer and a rectified linear unit activation function with constant learning rate were used. There were two hyperparameters that were tuned - *hidden layer size* and *learning rate*. There was only one hyperparameter, *smoothing*, that was tuned for the NB models. The parameter grids used for different models are given in Table - ST3 of Supplementary material.

2.4. Performance metrics

The performance metrics, which were computed, included both threshold-dependent metrics and threshold-independent metrics as follows-

- *Accuracy*: It is the ratio of the number of correctly predicted samples to the number of total samples (Eq. 1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP=true positives, TN=true negatives, FP=false positives and FN=false negatives

- *Sensitivity*: It is the ratio of the number of correctly predicted positive samples to the number of positive samples (Eq. 2).

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

- *Specificity*: It is the ratio of the number of correctly predicted negative samples to the number of negative samples (Eq. 3).

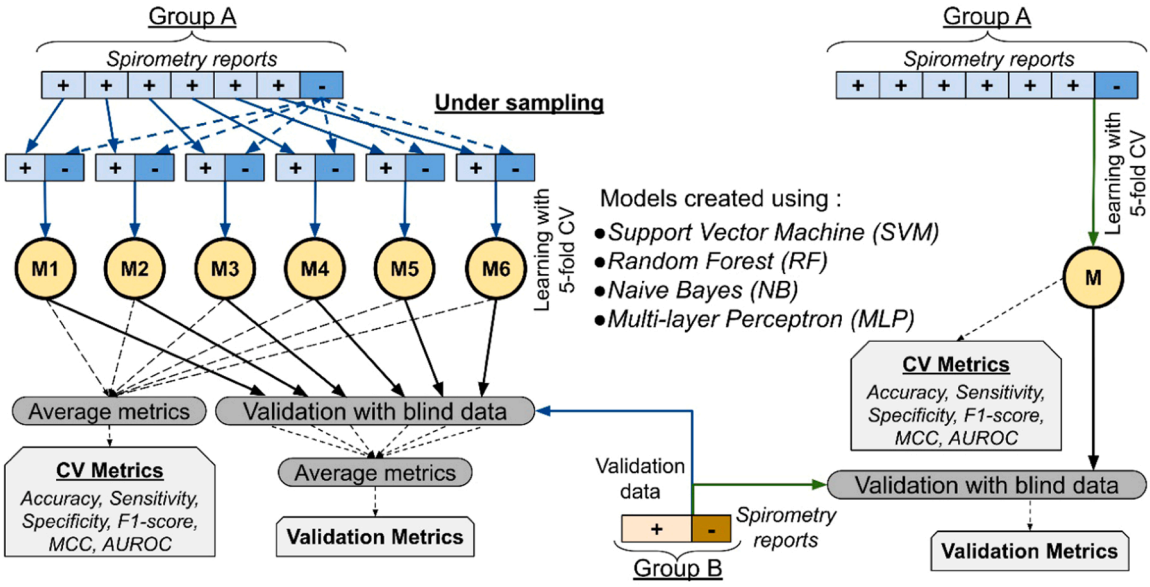

Fig. 1. Workflow for development of machine learning models.

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

- *F1-score*: It is the harmonic mean of precision and sensitivity (Eq. 4).

$$F1 \quad score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

- *Matthew's correlation coefficient (MCC)*: It is a correlation coefficient whose value is in the range [− 1, + 1]. The value of + 1 represents perfect prediction; 0 represents a prediction that is no better than random choice; and − 1 represents inverse prediction (Eq. 5).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

- *Area Under Receiver Operator Characteristic curve (AUROC)*: The ROC curve plots the *sensitivity* against *(1-specificity)* at different thresholds. It represents the ability of a binary classification model to perform at various discrimination thresholds. The area under the ROC curve denotes the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample by the model.

The optimal model was chosen on the basis of the highest MCC value as it accounts for all values of the confusion matrix and is suitable for imbalanced datasets [21,22].

2.5. Development of web application

A web application, named PulmoPred, was developed that acts as a user-friendly interface for using the ML models for prediction purposes. The Python instance representing the optimal model was serialized as a set of files and stored in a web server using the Python library *Joblib*. The user-input facility was implemented as a hypertext mark-up language (HTML) form. The scripts for server-side processing were written using PHP language. These PHP scripts are responsible for relaying the user input to the python scripts performing the ML based prediction (invoking the stored models) and sending the output back to the client-side. The data transmission between PHP and Python scripts were done using a popular format of data-interchange, JavaScript Object Notation (JSON). The final predicted class and the total probability are computed by averaging the prediction probabilities of the six under-sampled models. Let $p_i\ (+)$ and $p_i\ (-)$ be the probabilities that classifier i predicts a given sample as positive or negative respectively such that $p_i\ (+) = 1 - p_i\ (-)$ (for all $i \in [1, 6]$). Then, the total probability that predicted sample is positive $P\ (+)$ or negative $P\ (-)$ is given by Eq. (6) and Eq. (7) respectively. The final predicted class is positive, if $P\ (+) > P\ (-)$; and negative, otherwise.

$$P(+) = \frac{1}{6}\sum_{i=1}^{6} p_i(+) \tag{6}$$

$$P(-) = \frac{1}{6}\sum_{i=1}^{6} p_i(-) \tag{7}$$

3. Results

3.1. Results of hyperparameter optimization

The optimal hyperparameters obtained for the SVM models were *C= 5* and *gamma= 0.0001*. For NB models, the optimal *smoothing* hyperparameter was obtained as *0*. MLP architecture with two hidden layers (100 nodes followed by 30 nodes) was obtained as the optimal architecture for the MLP model trained with whole dataset. For the MLP models trained with the under-sampled datasets, architecture with two hidden layers (100 nodes followed by 100 nodes) was obtained as the optimal. The optimal learning rate was 0.001 for both sets of models. The optimal hyperparameters obtained for RF models trained with under-sampled datasets were: *number of trees= 20*, *max depth= 4* and *max features= 0.4*. For RF models trained with the whole dataset, the parameters were: *number of trees= 30*, *max depth= 8* and *max features= 0.7*. The detailed results of hyperparameter optimization using grid-search is given in Supplementary material (Tables ST4-ST11) The architecture of the optimal MLP model is shown in Fig SF2 in Supplementary material.

3.2. Results of 5-fold cross validation

The performances of different models with 5-fold CV using optimal hyperparameters are given in Table 2. The specificity of the models trained with the imbalanced datasets was low which shows the inability of these models to correctly predict non-obstructive diseases. In case of the models trained on under-sampled datasets, the increase in specificity represents that the models are not biased towards one class. The MLP model showed optimal performance with MCC of 0.68 and accuracy of 83.7%. The receiver operator characteristic (ROC) curves were plotted and the areas under the curves (AUC) were computed. The ROC plot of the models trained with under-sampled datasets is given in Fig. 2. The MLP model achieved the highest AUC of 0.91. The ROC plot of the models trained with whole dataset is given in Fig SF3 in Supplementary material. The accuracy of predicting the spirometry reports of disease-specific diagnosis (asthma, COPD or DPLD) as either obstructive or non-obstructive by the models are given in Table 3. It is the ratio of the number of correctly classified spirometry reports of each diagnosis to the total number of reports of that diagnosis. In this case, correct classification refers to the prediction of asthma & COPD as obstructive, and DPLD as non-obstructive. The RFE model selected 6 features- namely *FVC pre-BD value*, *FVC post-BD value*, *FVC post-BD predicted ratio*, *FEF pre-BD predicted ratio*, *FEF post-BD value* and *FEF post-BD predicted ratio* - to be the most predictive features (see Table ST2 in Supplementary material) and achieved a MCC score of 0.604 in the classification task (see detailed output in Supplementary material).

3.3. Performance on validation dataset

The performance of different models on validation dataset is given in Table 4. An increase of accuracy was achieved in comparison to 5-fold cross validation for all models trained with under-sampled datasets. The MCC values were similar to that in cross validation. The trend of low specificity also persisted for validation dataset prediction with models trained on the imbalanced datasets. The accuracy of predicting the spirometry reports of each diagnosis in the blind dataset as either obstructive or non-obstructive by the models are given in Table 5.

Table 2
Performance of models with 5-fold cross validation.

Dataset	Model	Accuracy	Sensitivity	Specificity	F1-score	MCC
Whole training dataset	SVM	0.835	0.837	0.826	0.897	0.532
	RF	0.906	0.955	0.609	0.946	0.597
	NB	0.870	0.915	0.602	0.924	0.495
	MLP	0.918	0.966	0.626	0.953	0.645
Under-sampled datasets	SVM	0.823	0.825	0.821	0.824	0.650
	RF	0.822	0.832	0.811	0.824	0.647
	NB	0.800	0.864	0.737	0.813	0.607
	MLP	0.837	0.853	0.822	0.841	0.682

Abbreviations: MCC - Matthews correlation coefficient; SVM - Support Vector Machine; RF - Random Forest; NB - Naive Bayes; MLP - Multi-Layer Perceptron.

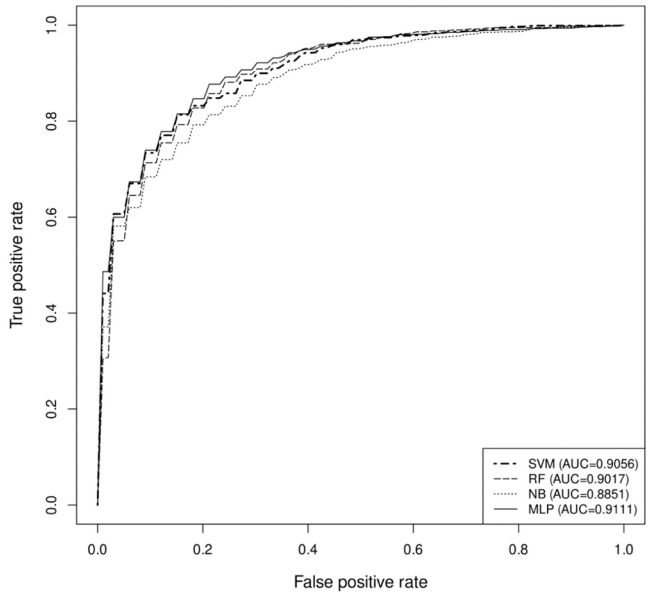

Fig. 2. ROC plot of different models trained with under-sampled datasets. (SVM - Support Vector Machine; RF - Random Forest; NB - Naive Bayes; MLP - Multi-Layer Perceptron; AUC - Area Under Curve).

Table 3
Accuracies of predicting the spirometry reports of obstructive or non-obstructive disease-specific diagnosis in the training dataset.

	Models trained with whole training dataset				Models trained with under-sampled datasets			
	SVM	RF	NB	MLP	SVM	RF	NB	MLP
Prediction of asthma as obstructive	0.77	0.94	0.87	0.96	0.75	0.77	0.79	0.80
Prediction of COPD as obstructive	0.93	0.98	0.98	0.98	0.93	0.93	0.96	0.93
Prediction of DPLD as non-obstructive	0.90	0.72	0.68	0.75	0.88	0.90	0.81	0.89

Abbreviations: COPD - Chronic Obstructive Pulmonary Disease; DPLD - Diffuse Parenchymal Lung Disease; SVM - Support Vector Machine; RF - Random Forest; NB - Naive Bayes; MLP - Multi-Layer Perceptron.

Table 4
Performance of models on predicting the validation dataset.

Training dataset	Model	Accuracy	Sensitivity	Specificity	F1-score	MCC
Whole training dataset	SVM	0.853	0.897	0.765	0.891	0.667
	RF	0.835	0.971	0.561	0.887	0.619
	NB	0.823	0.944	0.580	0.877	0.586
	MLP	0.857	0.986	0.596	0.902	0.677
Under-sampled datasets	SVM	0.854	0.898	0.766	0.892	0.669
	RF	0.862	0.902	0.781	0.897	0.687
	NB	0.855	0.926	0.712	0.895	0.665
	MLP	0.849	0.886	0.774	0.887	0.663

Abbreviations: MCC - Matthews correlation coefficient; SVM - Support Vector Machine; RF - Random Forest; NB - Naive Bayes; MLP - Multi-Layer Perceptron.

4. Using the web application

The PulmoPred homepage is shown in Fig. 3. The form with 12 mandatory fields is given in the homepage for providing the input spirometry values. There are buttons to insert two sets of sample data for quick demonstration. A "Reset" button is present that enables users to clear all input values from the form. Finally, there is a "Submit" button to perform the prediction task based on the given input. The output page consists of three tables as shown in Fig. SF4 in Supplementary material. The first table from the top shows the input values provided by the user. The second table displays the predicted class label and the prediction probability of the six models trained with under-sampled datasets. The third table shows the result of total probability (*P (+)* and *P (-)*) calculations for each class using Eq. (6) and Eq. (7). It also displays the final predicted class (Obstructive/Non-obstructive) by comparing *P (+)* with *P (-),* and the final probability in the last column (lower right of the page). A "*Help*" page was also created that contains more information regarding the usage of the web application.

Table 5
Accuracies of predicting the spirometry reports of obstructive or non-obstructive disease-specific diagnosis in the validation dataset.

	Models trained with whole training dataset				Models trained with under-sampled datasets			
	SVM	RF	NB	MLP	SVM	RF	NB	MLP
Prediction of asthma as obstructive	0.86	0.94	0.90	0.97	0.86	0.86	0.87	0.86
Prediction of COPD as obstructive	0.93	0.99	0.98	1.0	0.93	0.94	0.97	0.91
Prediction of DPLD as non-obstructive	0.76	0.56	0.54	0.59	0.75	0.77	0.70	0.77

Abbreviations: COPD - Chronic Obstructive Pulmonary Disease; DPLD - Diffuse Parenchymal Lung Disease; SVM - Support Vector Machine; RF - Random Forest; NB - Naive Bayes; MLP - Multi-Layer Perceptron.

5. Discussion

Two sets of models were trained – one with the whole training dataset; and another with the under-sampled datasets. The six under-sampled datasets were created because the whole training dataset was imbalanced (*P:N:: 1:6*). The under-sampling approach used in this work allowed the models to be trained with balanced datasets without discarding any data sample. In the models trained with the whole training dataset, the specificity and MCC values were low. The models trained with under-sampled datasets performed better which was evident from higher specificity and MCC values. While analyzing the ability of the models to perform disease-specific diagnosis, it was observed that spirometry reports of COPD and DPLD patients were predicted with high accuracy (~90%) as compared to prediction of asthma in the training dataset (Table 3). However, in the validation phase with blind dataset, the accuracy of prediction of asthma was increased considerably (Table 5).

The models performed well with accuracies of ~85% in both training and validation phases. Any performance deterioration was caused by lack of rich enough training data which can be attributed to common physiological and structural causes or symptoms of the two classes. Also, there are evidences in the literature which showed that obstructive and non-obstructive diseases are co-existent in a small fraction of patients [23,24]. This may lead to wrong diagnostic annotations of patients for supervised learning. The availability of spirometry reports of patients with non-obstructive diseases was low. Also, the majority of these patients were diagnosed with DPLD. The classification of other non-obstructive diseases can be performed in future with the availability of data.

6. Conclusion

In this work, we created supervised ML models to classify obstructive and non-obstructive pulmonary diseases in patients of eastern India

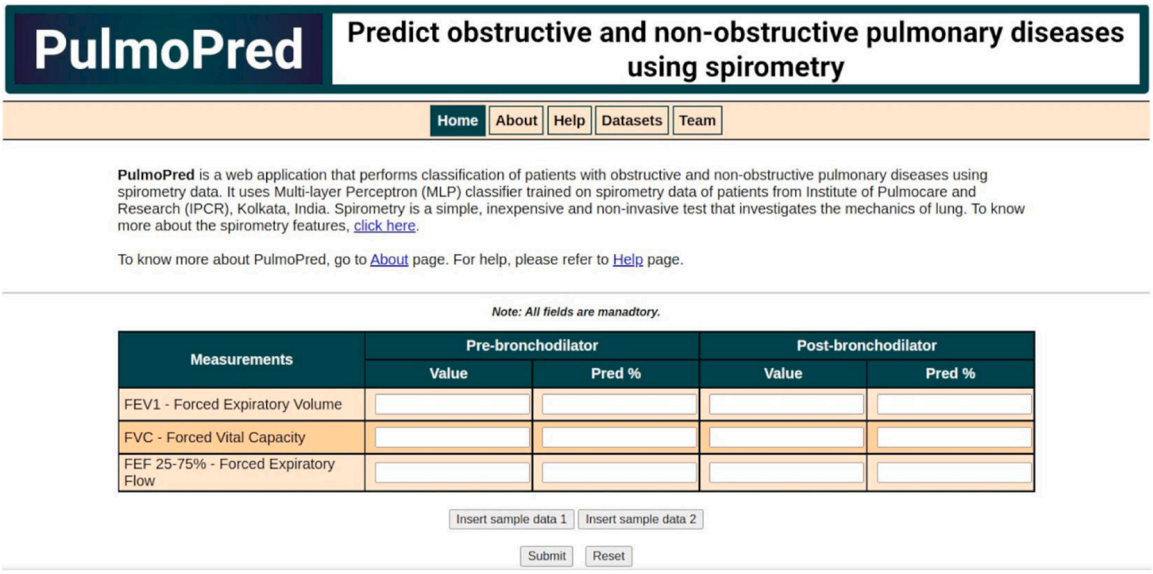

Fig. 3. Homepage of the web application.

using only 12 attributes of spirometry which is a simple and fairly inexpensive investigation. Models were created using support vector machine, random forest, Naive Bayes and multi-layer perceptron algorithms. A web application was developed to serve as an easy-to-use interface for naive users performing the ML-based prediction.

Funding

This work is supported by Indian Council of Medical Research [Project ID: *2019–0075*].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials

The web application, PulmoPred, is available at *http://dibresources.jcbose.ac.in/ssaha4/pulmopred*. Source code and data are available at *https://github.com/ttsudipto/PulmoPred*.

Acknowledgments

The authors thank Bose Institute, Kolkata, India for providing the server to host the prediction server. The authors thank Institute of Pulmocare and Research, Kolkata for providing the patient spirometry data. SB thanks ICMR for the fellowship.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jocs.2022.101768.

References

[1] F. of I.R. Societies, The Global Impact of Respiratory Disease, Second Edi, European Respiratory Society, Sheffield, 2017. ⟨https://www.firsnet.org/⟩ (accessed April 27, 2020).
[2] A.S. Buist, Similarities and differences between asthma and chronic obstructive pulmonary disease: treatment and early outcomes, Eur. Respir. J. Suppl. 39 (2003) 30s–35s, https://doi.org/10.1183/09031936.03.00404903.
[3] R. Gilbert, J.H. Auchincloss, What is a "restrictive" defect? Arch. Intern. Med. 146 (1986) 1779–1781, https://doi.org/10.1001/archinte.1986.00360210165023.
[4] R.O. Crapo, Pulmonary-function testing, N. Engl. J. Med. 331 (1994) 25–30, https://doi.org/10.1056/NEJM199407073310107.
[5] I. Pino Peña, V. Cheplygina, S. Paschaloudi, M. Vuust, J. Carl, U.M. Weinreich, L. R. Østergaard, M. de Bruijne, Automatic emphysema detection using weakly labeled HRCT lung images, PLoS One 13 (2018), e0205397, https://doi.org/10.1371/journal.pone.0205397.
[6] D. Bermejo-Peláez, S.Y. Ash, G.R. Washko, R. San José Estépar, M.J. Ledesma-Carbayo, Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks, Sci. Rep. 10 (2020) 338, https://doi.org/10.1038/s41598-019-56989-5.
[7] N.R.S. Parveen, M.M. Sathik, Detection of pneumonia in chest X-ray images, J. Xray. Sci. Technol. 19 (2011) 423–428, https://doi.org/10.3233/XST-2011-0304.
[8] P. Lakhani, B. Sundaram, Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2017) 574–582, https://doi.org/10.1148/radiol.2017162326.
[9] P.-H. Feng, Y.-T. Lin, C.-M. Lo, A machine learning texture model for classifying lung cancer subtypes using preliminary bronchoscopic findings, Med. Phys. 45 (2018) 5509–5514, https://doi.org/10.1002/mp.13241.
[10] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R. N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inform. 108 (2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.09.013.
[11] S. Swaminathan, K. Qirko, T. Smith, E. Corcoran, N.G. Wysham, G. Bazaz, G. Kappel, A.N. Gerber, A machine learning approach to triaging patients with chronic obstructive pulmonary disease, e0188532–e0188532, PLoS One 12 (2017), https://doi.org/10.1371/journal.pone.0188532.
[12] S. Aneja, S. Lal, Effective asthma disease prediction using naive Bayes — Neural network fusion technique, in: 2014 Int. Conf. Parallel, Distrib. Grid Comput., IEEE, Solan, India, 2014: pp. 137–140. https://doi.org/10.1109/PDGC.2014.7030730.
[13] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from EMR data using machine learning, AMIA. Annu. Symp. Proceedings. AMIA Symp. 2012 (2012) 606–615. https://pubmed.ncbi.nlm.nih.gov/23304333.
[14] L. Hussain, A. Ahmed, S. Saeed, S. Rathore, I.A. Awan, S.A. Shah, A. Majid, A. Idris, A.A. Awan, Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies, Cancer Biomark. 21 (2018) 393–413, https://doi.org/10.3233/CBM-170643.
[15] H. Ranu, M. Wilde, B. Madden, Pulmonary function tests, Ulst. Med. J. 80 (2011) 84–90.
[16] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.
[17] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[18] H. Zhang, The Optimality of Naïve Bayes, in: Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. (FLAIRS 2004), AAAI Press, Florida, USA, 2004, pp. 562–567.
[19] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536, https://doi.org/10.1038/323533a0.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[21] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.
[22] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLoS One 12 (2017), e0177678, https://doi.org/10.1371/journal.pone.0177678.
[23] E. Diaz-Guzman, K. McCarthy, A. Siu, J.K. Stoller, Frequency and causes of combined obstruction and restriction identified in pulmonary function tests in adults, Respir. Care. 55 (2010), 310 LP – 316, ⟨http://rc.rcjournal.com/content/55/3/310.abstract⟩.
[24] Z.S. Gardner, G.L. Ruppel, D.A. Kaminsky, Grading the severity of obstruction in mixed obstructive-restrictive lung disease, Chest 140 (2011) 598–603, https://doi.org/10.1378/chest.10-2860.