BMJ Open Respiratory Research

Deep learning using multilayer perception improves the diagnostic acumen of spirometry: a single-centre Canadian study

Amanda Mac,[1] Tong Xu,[1] Joyce K Y Wu ,[1,2] Natalia Belousova ,[1,2] Haruna Kitazawa ,[1,2] Nick Vozoris ,[1] Dmitry Rozenberg ,[1,2] Clodagh M Ryan ,[1,2] Shahrokh Valaee ,[3] Chung-Wai Chow [1,2]

To cite: Mac A, Xu T, Wu JKY, *et al*. Deep learning using multilayer perception improves the diagnostic acumen of spirometry: a single-centre Canadian study. *BMJ Open Resp Res* 2022;**9**:e001396. doi:10.1136/bmjresp-2022-001396

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjresp-2022-001396).

AM and TX are joint first authors.

Received 9 August 2022
Accepted 14 December 2022

[1]Medicine, Division of Respirology, University of Toronto, Toronto, Ontario, Canada
[2]Department of Medicine, University Health Network, Toronto, Ontario, Canada
[3]Electrical and Computer Engineeing, University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Chung-Wai Chow;
Chung-Wai.Chow@uhn.ca

ABSTRACT

Rationale Spirometry and plethysmography are the gold standard pulmonary function tests (PFT) for diagnosis and management of lung disease. Due to the inaccessibility of plethysmography, spirometry is often used alone but this leads to missed or misdiagnoses as spirometry cannot identify restrictive disease without plethysmography. We aimed to develop a deep learning model to improve interpretation of spirometry alone.

Methods We built a multilayer perceptron model using full PFTs from 748 patients, interpreted according to international guidelines. Inputs included spirometry (forced vital capacity, forced expiratory volume in 1 s, forced mid-expiratory flow$_{25-75}$), plethysmography (total lung capacity, residual volume) and biometrics (sex, age, height). The model was developed with 2582 PFTs from 477 patients, randomly divided into training (80%), validation (10%) and test (10%) sets, and refined using 1245 previously unseen PFTs from 271 patients, split 50/50 as validation (136 patients) and test (135 patients) sets. Only one test per patient was used for each of 10 experiments conducted for each input combination. The final model was compared with interpretation of 82 spirometry tests by 6 trained pulmonologists and a decision tree.

Results Accuracies from the first 477 patients were similar when inputs included biometrics+spirometry+plethysmography (95%±3%) vs biometrics+spirometry (90%±2%). Model refinement with the next 271 patients improved accuracies with biometrics+pirometry (95%±2%) but no change for biometrics+spirometry+plethysmography (95%±2%). The final model significantly outperformed (94.67%±2.63%, p<0.01 for both) interpretation of 82 spirometry tests by the decision tree (75.61%±0.00%) and pulmonologists (66.67%±14.63%).

Conclusions Deep learning improves the diagnostic acumen of spirometry and classifies lung physiology better than pulmonologists with accuracies comparable to full PFTs.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Spirometry is the most commonly used pulmonary function test for screening and management of lung disease. Without assessment of lung volumes using plethysmography, spirometry misses restrictive defects and can lead to misdiagnoses. Computer-aided tools have been developed to improve classification of lung physiology patterns. However, these tools require the inclusion of plethysmography measurements and/or clinical symptoms. No study has developed machine learning algorithms for classifying the major lung conditions using spirometry only.

WHAT THIS STUDY ADDS

⇒ Deep learning using a multilayer perceptron model with spirometry data provides classification accuracies of lung physiology patterns that are comparable to full pulmonary function testing, which includes both spirometry and plethysmography, and better than trained pulmonologists.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Implementation of the deep learning model (code appended in this paper) in spirometers will facilitate accurate identification of pulmonary physiology patterns to appropriately triage patients for subsequent investigations and/or therapy. This will improve equity in healthcare access; patients who live in regions of the world where spirometry is available, but access to diagnostic laboratories for full pulmonary testing is limited, will receive equitable care when machine learning is applied. The machine learning model will also increase healthcare delivery efficiency and improve patient outcomes by facilitating earlier diagnosis of lung diseases.

INTRODUCTION

The current gold standard pulmonary function test (PFT) consists of both spirometry and plethysmography,[1 2] which have well-established guidelines for conduct and interpretation.[3] However, patient access to plethysmography is often limited due to the need for expensive infrastructure and technical expertise. Furthermore, the plethysmograph does

not readily accommodate patients with physical disabilities and/or claustrophobia. Spirometry alone is the most common PFT modality; it is portable and easily deployed in multiple settings including the bedside, clinic, home or workplace.[4] However, spirometry can miss or misdiagnose lung disease as it has limited ability to identify early obstructive lung disease and restrictive defects in the absence of plethysmography.[5 6]

Many physicians use computer-aided tools to facilitate PFT interpretations, such as the decision-tree algorithm developed in our laboratory (online supplemental file 1). However, few studies have applied machine learning for de novo interpretation of PFTs. One group developed a multiclass support vector machine algorithm to classify normal, obstructive and restrictive patterns based on forced expiratory volume in 1 s (FEV_1), forced vital capacity (FVC) and FEV_1/FVC.[7] While they reported high validation accuracies, determination of the true labels of PFTs did not adhere to international guidelines; obstruction was defined by FEV_1/FVC <75%, rather than the lower limit of normal and restriction by FVC <80% with normal FEV_1/FVC but without plethysmography.[7] Biometrics, key for the derivation of normal reference values,[8] were not included in their model.[7] Topalovic *et al* hypothesised that machine learning could reduce the inter-rater variability commonly observed in PFT interpretation by pulmonologists[9] and compared their interpretations to a decision tree model built with data from 1430 subjects.[10] The gold standard label used for comparison was based on consensus diagnosis made by three clinicians following review of the clinical history, complete PFTs (prebronchodilator and postbronchodilator spirometry, lung volumes, airways resistance and diffusing capacity) and all tests deemed necessary by the responsible physician. Pulmonologists and the machine learning model were given the same data: full PFTs and clinical information (smoking history, cough, sputum and dyspnoea). Physicians correctly classified the respiratory patterns with 74% accuracy (ranging 56%–88%) in contrast to 100% by the machine learning model.[9]

Multilayer perceptron (MLP) is a type of artificial neural network (ANN) that models non-linear input and output relationships by learning the statistics of large general datasets.[11] MLP consists of neurons arranged in an input layer, one or more hidden layers and an output layer.[12] MLP containing more than one hidden layer is called deep MLP (DMLP). The input layer includes multiple attributes and input variables, which the model uses to classify the data into different categories. The hidden layers between the input and output layers include intermediate neurons. Each intermediate neuron performs a weighted summation of its inputs and passes the sum to an activation function to produce a value that represents the neuron's firing intensity.[12] Each layer of neurons activates the sequential layer, eventually generating the output variables in the output layer.[12] The output variables are digital series representing the defined categories that the model aims to classify. DMLP has no restriction on the type or number of input variables; it considers every possible interaction between input variables, enhancing complexity and classification capability.[13]

In developing a DMLP, the model is first given a training dataset of prelabelled samples to learn the classification rules. Learning is achieved by adjusting network hyperparameters to generate the best fit for the dataset without explicit instructions. A subsequent validation dataset is used to estimate how well the model learnt the classification rules and tune the hyperparameter values to optimise classification accuracies. Finally, a test dataset containing unseen samples is applied to the refined model to assess its classification performance. During training, weights are updated layer-by-layer based on discrepancies between the actual and output label of each sample. Since ANNs with more than one hidden layer and non-linear activation functions cannot be expressed using linear equations, trained models provide limited information on the decision-making processes.[13] We can evaluate whether the ANN has been appropriately trained by assessing its performance or classification accuracy, but we cannot identify how the model learnt to make classifications.[13 14] MLP has been shown to yield better classification outcomes compared with statistical methods in practice.[11 12]

We hypothesise that a DMLP model can accurately distinguish normal, obstructive, restrictive and mixed obstructive-restrictive physiology patterns based on spirometric and biometric measurements.

METHODS

Data collection and labelling

We used 3827 full PFTs from 748 adult patients collected as routine care between June 2018 and October 2021. Spirometry and plethysmography were performed in the sitting position using BodyBox (Medisoft, Sorinnes, Belgium), following American Thoracic Society/European Respiratory Society (ATS/ERS) guidelines.[4 15] Tests were labelled with their 'true' physiological pattern (normal, obstructive, restrictive or mixed obstructive-restrictive) based on biometrics (sex, age, height), spirometry and lung volume measurements, in accordance with ATS/ERS guidelines,[4 16] facilitated by a computer-aided algorithm (online supplemental file 1) and confirmed by one of six pulmonologists.

Patient and public involvement

Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this study.

Data processing

Sex was binarily converted into 0 (female) and 1 (male). Age, height, spirometric and plethysmographic absolute values were scaled from 0 to 1, using MinMax Scaling equation, $y = \frac{x - min(X)}{max(X) - min(X)}$, where y is the scaled value,

x is the original value and X is the collection of values for a specific input variable. True labels of each test were digitalised: output classes were converted to 3 for normal, 4 for obstructive, 5 for mixed obstructive-restrictive and 6 for restrictive pattern.

DMLP model development

The DMLP model was developed using 2582 tests (collected June 2018 to March 2020) from 477 patients. A Random Search was performed to identify optimal DMLP hyperparameters and regularisation values to be used in the model.[17] The model was built with two hidden layers of 180 and 30 neurons each, with drop-out rates of 0.2 and 0.1, respectively. Weights were initialised to a random set of small values from a normal distribution with mean of 0.005 and SD of 0.001667. Adaptive moment estimation optimiser ($\beta1=0.9$, $\beta2=0.900$, $\varepsilon=10^{-8}$) was used to regularise the learning rate. Learning rates from 0.0001 to 0.1 were tested with logarithmic increments. The model was trained in batches of 32 samples for a maximum of 900 epochs. Early stopping was set such that model training would stop if the validation loss had not improved for 100 epochs (online supplemental file 2).

The DMLP model was given four input variable combinations: (1) biometrics with spirometry and plethysmography; (2) biometrics with spirometry; (3) spirometry and plethysmography and (4) spirometry alone. The included values for spirometry were FVC, FEV_1 and forced mid-expiratory flow (FEF_{25-75}); plethysmography were total lung capacity (TLC) and residual volume (RV), and biometrics were sex, age and height. Ten experimental runs were completed for each input combination. For each run, the model randomly selected one test per patient so that each run used 477 tests from 477 unique patients to reduce redundancy in the dataset. The inputted data were randomly partitioned into training (80%), validation (10%) and test (10%) sets of unique patients.[18 19]

The mean accuracy, precision and recall values of each experimental run were calculated as follows: accuracy by dividing the number of correct predictions by the total number of samples in the test set; precision by dividing the number of true positives by the sum of true positives and false positives for each lung pattern predicted by the model on the test set; recall by dividing the number of true positives by the sum of true positives and false negatives for each lung condition predicted by the model on the test set. The F1 score, a machine learning metric of model performance, for each lung pattern was calculated using the formula, $\frac{2(Precision)(Recall)}{Precision+Recall}$.[20]

Model refinement and application to unseen data

The DMLP model was refined using 1245 previously unseen PFTs (collected July 2020–October 2021) from 271 new patients. Here, the data from the first 477 patients used in model development were placed into the training set and those from the 271 new patients were split evenly and randomly into validation (136 patients) and test (135 patients) sets. Performance of the refined model using the four input combinations were repeated, as described above. DMLP design and data partitioning for model development, refinement and application is outlined in online supplemental file 3.

Comparing DMLP to pulmonologists and ATS/ERS decision tree

We evaluated the performances of the DMLP model, ATS/ERS decision tree (online supplemental file 4) and six pulmonologists who were given standard reports (online supplemental file 5) in classifying 82 spirometry tests. Accuracies for the DMLP model, ATS/ERS decision tree and pulmonologists were calculated by comparing their classifications to the 'true' interpretation which were determined based on full PFTs (described above). We used two-sample t-tests to compare the model to the decision tree and pulmonologists.

RESULTS

The PFTs were concordant with their 'true' lung pattern labels (table 1). Normal PFTs had values greater than 80% predicted. Obstructive patterns had FEV_1/FVC ratios below 80%, and FVC and TLC were greater than 80% predicted. Restrictive patterns had FEV_1/FVC ratios greater than 80%, with both FVC and TLC below 70% predicted. Mixed obstructive-restrictive patterns exhibited FEV_1/FVC ratios below 82%, with both FVC and TLC below 70% predicted.

Using data from the first 477 patients to build the DMLP model, we found comparable accuracies when the inputs included biometrics with spirometry and plethysmography versus biometrics with spirometry only (table 2). Next, we validated the DMLP model with previously unseen data. Here, data from the initial 477 patients were placed into the training set; data from the 271 new patients were equally divided into the validation (136 patients, to further refine the hyperparameters) and test (135 patients) sets. Again, we observed high test set classification accuracies when inputs included biometrics with full PFTs versus biometrics with spirometry only (table 2). Biometrics are important as their absence reduced tests accuracies for spirometry and plethysmography, and spirometry only (table 2).

The test set precision improved from 74%–100% for the input combination of biometrics, spirometry and plethysmography to 88%–100% for the combination of biometrics and spirometry (tables 3 and 4). Test set recall values were similarly high between the input combinations of biometric, spirometry and plethysmography (88%–100%) and biometrics with spirometry (86%–100%) (tables 3 and 4). Both the precision and recall values drastically decreased when biometrics were omitted (tables 3 and 4). Larger datasets improve the accuracy of machine learning as illustrated by higher F1,

Table 1 Mean spirometry and plethysmography measurements for the lung pattern labels

	First 477 patients (2582 tests) collected June 2018–March 2020				Later 271 patients (1245 tests) collected July 2020–October 2021			
Lung conditions	Normal	Obstructive	Restrictive	Obstructive and restrictive	Normal	Obstructive	Restrictive	Obstructive and restrictive
No of tests (%)	780 (30)	335 (13)	1130 (44)	337 (13)	276 (22)	147 (12)	643 (52)	179 (14)
FEV_1 (L) (%FEV_1)	3.06 (92.79)	1.95 (63.51)	2.24 (68.77)	1.56 (48.12)	2.97 (100.76)	1.90 (58.69)	2.20 (68.88)	1.50 (43.59)
FVC (L) (%FVC)	3.66 (86.90)	3.27 (80.63)	2.58 (63.77)	2.51 (59.73)	3.55 (93.93)	3.39 (80.29)	2.58 (62.21)	2.66 (59.02)
FEV_1/FVC (%FEV_1/FVC)	84.61 (107.76)	59.16 (78.59)	86.79 (107.73)	61.96 (80.93)	84.45 (108.02)	55.94 (71.52)	86.18 (110.16)	56.19 (72.32)
FEF_{25-75} (L/s)(%FEF_{25-75})	3.85 (142.22)	1.24 (52.53)	3.39 (121.02)	1.00 (39.90)	3.45 (144.23)	1.03 (38.82)	2.96 (113.17)	0.72 (25.67)
TLC (L) (%TLC)	5.79 (92.05)	5.56 (89.99)	4.26 (69.48)	4.30 (69.22)	5.43 (91.82)	5.85 (91.60)	4.11(65.19)	4.67 (68.76)
RV (L) (%RV)	2.08 (105.25)	2.22 (111.76)	1.66 (85.33)	1.75 (92.03)	1.85(91.78)	2.31 (112.43)	1.52 (75.93)	1.96 (89.73)
RV/TLC (%RV/TLC)	36.33 (102.12)	40.50 (112.30)	39.50 (109.56)	41.11 (118.08)	34.45 (89.61)	39.92 (108.5)	37.50 (103.79)	42.23 (113.99)

FEV_1, FVC, TLC and RV are in litres, FEF_{25-75} in litres per second.
FEF, forced mid-expiratory flow; FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity; RV, residual volume; TLC, total lung capacity.

precision and recall values in the larger (table 4) versus the smaller datasets (table 3).

Lastly, we compared the DMLP classification with ATS/ERS decision tree[3] and interpretations by six board-certified pulmonologists using 82 spirometry tests (table 5). The DMLP significantly outperformed the decision tree and pulmonologists ($p<0.0001$ and 0.0051, respectively) with no significant difference between pulmonologists and the decision tree ($p=0.5958$). The confusion matrix (online supplemental file 6) indicated that the mixed obstructive-restrictive pattern was the most difficult to classify, correlating with the lower F1 score for this category when the DMLP was inputted with biometrics and spirometry (table 4).

DISCUSSION

We specifically focused on improving the diagnostic accuracies of spirometry as it is the most common PFT modality used for initial assessment of patients with suspected lung disease. Many patients, particularly those in underserviced, rural and remote areas, have limited access to the gold standard full PFT. While diagnosis and management of patients with lung diseases, particularly restrictive lung disease, require clinical evaluation and full PFT that includes spirometry, plethysmography and diffusion capacity, maximising the utility of readily available diagnostic modalities to improve diagnostic acumen will alleviate some of the inequities of healthcare access. Thus, development of machine learning that improves the diagnostic yield of spirometry will improve equity and healthcare delivery for everyone regardless of access to plethysmography.

Our DMLP model was developed with readily available, clinically relevant spirometry variables (FVC, FEV_1 and FEF_{25-75}). We compared its classification accuracies to true labels determined by full PFTs following international interpretation standards.[4] The model using biometrics and spirometry classified the major physiological patterns with 95% accuracy and was comparable to the model with full PFTs and biometrics. In other words, interpretation of spirometry inclusive of biometrics using DMLP classifies respiratory patterns accurately without the need for plethysmography. The DMLP model also out-performed the ATS/ERS decision tree and trained pulmonologists. For both the DMLP model and physicians, the mixed obstructive-restrictive defect was the most difficult to classify as indicated by the low F1 score (91.53%±11.83%) and confusion matrix (online supplemental file 6). This pattern also had the highest interphysician discrepancies (online supplemental file 7) and suggests these patients should be triaged early for further investigations (full PFT, imaging) to better characterise the disorder.

A key strength of our study is the adherence to international guidelines for conduct and interpretation of PFTs. The 'true' labels of PFTs used to evaluate the performance of the DMLP model were determined using spirometry, plethysmography and calculated lower and

Table 2 Classification accuracies for each combination of input variables

Combination of input variables	Training set (%)	Validation set (%)	Test set (%)
DMLP development using 477 patients divided into training (80%), validation (10%) and test (10%) sets			
Biometrics with spirometry and plethysmography	97.38±1.03	95.53±2.55	94.79±2.56
Biometrics with spirometry	89.58±1.14	87.87±2.47	89.81±2.06
Spirometry and plethysmography	62.50±7.58	57.41±7.43	59.87±9.17
Spirometry	61.28±4.19	57.42±8.76	59.39±7.30
DMLP refinement and application to unseen data: first 477 patients used for training, later 271 patients for validation (136) and test (135) sets			
Combination of input variables	**Training set (%)**	**Validation set (%)**	**Test set (%)**
Biometrics with spirometry and plethysmography	98.24±0.85	95.42±1.88	95.04±1.71
Biometrics with spirometry	98.39±1.07	94.83±2.7	95.19±2.35
Spirometry and plethysmography	67.51±7.29	65.00±5.71	63.33±6.64
Spirometry	58.01±7.41	56.69±8.24	56.37±8.91

Accuracies (mean±SD) for each combination of input variables after 10 experimental runs are shown. For each experimental run, the model randomly selected one test per patient. Biometric variables are sex, age, height. Spirometry included FVC, FEV_1, FEF_{25-75} and plethysmography metrics are RV and TLC.
DMLP, deep multilayer perceptron; FEF, forced mid-expiratory flow; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; RV, residual volume; TLC, total lung capacity.

Table 3 Test set precision, recall and F1 scores for each input variable combination when using data from the first 477 patients

Combination of input variables	Lung pattern	Precision for test set (%)	Recall for test set (%)	F1 score for test set (%)
Biometrics with spirometry and plethysmography	Normal	91.60±2.99	88.30±4.35	89.87±3.10
	Obstructive	73.50±6.85	100.00	84.56±4.56
	Restrictive	100.00	91.90±2.85	95.76±1.49
	Obstructive and restrictive	100.00	96.00±8.43	97.78±4.68
Biometrics with spirometry	Normal	90.70±6.34	92.60±4.90	91.36±2.22
	Obstructive	89.60±11.67	85.90±17.21	86.28±10.87
	Restrictive	89.90±6.37	88.00±7.39	88.67±4.51
	Obstructive and restrictive	88.00±17.76	88.20±15.92	87.42±14.86
Spirometry and plethysmography	Normal	59.10±10.73	65.00±11.45	61.63±10.30
	Obstructive	50.80±42.04	35.70±29.29	Undefined
	Restrictive	62.50±11.03	59.20±12.21	60.45±10.64
	Obstructive and restrictive	38.20±29.72	59.60±44.28	Undefined
Spirometry	Normal	60.70±10.13	70.30±10.99	64.59±7.92
	Obstructive	51.50±34.16	49.00±36.74	Undefined
	Restrictive	59.80±11.67	49.50±17.89	53.03±14.47
	Obstructive and restrictive	44.00±24.76	59.00±41.55	Undefined

Mean±SD for the test sets of 10 experimental runs are shown. In these experiments, data from the initial 477 patients were used for training (80%), validation (10%) and test (10%) sets. For each run, the model randomly selected one test per patient. Biometric- variables are sex, age, height. Spirometry metrics included are FVC, FEV_1, FEF_{25-75} and plethysmography metrics are RV and TLC. Undefined F1 score indicates that either precision, recall or both for that particular class was equal to zero for at least 1 of the 10 experimental runs.
FEF, forced mid-expiratory flow; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; RV, residual volume; TLC, total lung capacity.

Table 4 Test set precision, recall and F1 scores for each input variable combination when using previously unseen data obtained from the next 271 patients

Combination of input variable	Lung pattern	Precision for test set (%)	Recall for test set (%)	F1 score for test set (%)
Biometrics with spirometry and plethysmography	Normal	99.20±1.69	90.40±8.98	94.37±5.32
	Obstructive	96.60±7.17	88.20±9.39	91.94±7.11
	Restrictive	94.60±1.43	100.00	97.22±0.76
	Obstructive and restrictive	90.40±6.13	88.00±6.53	88.94±4.02
Biometrics with spirometry	Normal	99.50±1.58	92.10±8.31	95.51±5.29
	Obstructive	96.20±8.07	92.20±5.98	94.15±6.99
	Restrictive	94.60±0.97	100.00	97.22±0.52
	Obstructive and restrictive	91.80±5.43	94.60±31.45	91.53±11.83
Spirometry and plethysmography	Normal	62.90±21.53	44.00±11.73	48.99±10.32
	Obstructive	28.30±21.18	13.80±13.48	Undefined
	Restrictive	79.50±6.57	79.40±14.09	78.80±8.22
	Obstructive and restrictive	48.40±8.57	94.70±5.08	63.63±7.34
Spirometry	Normal	44.00±11.12	67.10±13.53	51.76±8.17
	Obstructive	42.10±14.43	53.40±34.77	43.90±20.78
	Restrictive	75.20±8.20	56.50±20.12	62.87±17.40
	Obstructive and restrictive	39.40±27.83	42.70±36.57	Undefined

Mean±SD for the test sets of 10 experimental runs are shown. Data from the initial 477 patients were used for training and data from the next 271 patients were randomly assigned to the validation (136 patients) and test (135 patients) sets. For each experimental run, the model randomly selected one test per patient. Biometric variables are sex, age, height. Spirometry metrics included are FVC, FEV_1, $FEF_{25–75}$ and plethysmography metrics are RV and TLC. Undefined F1 score indicates that either precision, recall or both for that particular class was equal to zero for at least 1 of the 10 experimental runs.
FEF, forced mid-expiratory flow; FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity; RV, residual volume; TLC, total lung capacity.

Table 5 Performance comparisons of the DMLP model, pulmonologists and ATS/ERS decision tree

	Accuracy (%)	P value (compared with DMLP model)
DMLP model	94.67±2.63	–
Pulmonologists	66.67±14.63	0.0051
ATS/ERS decision tree	75.61	<0.0001

Accuracies (mean±SD) of the DMLP model (n=10), pulmonologists (n=6) and the ATS/ERS decision tree for classifying 82 spirometry tests as normal, obstructive, restrictive and mixed obstructive-restrictive disease tests when compared with the 'true' labels based on the ATS/ERS interpretations using full PFTs with spirometry and plethysmography. Comparison of the groups were conducted using two-sample t-tests. The lowest accuracy was found in the pulmonologists and the highest in the DMLP. No differences were observed in the accuracies of the pulmonologists and the ATS/ERS decision trees (p=0.5958).
ATS/ERS, American Thoracic Society/European Respiratory Society; DMLP, deep multilayer perceptron; PFTs, pulmonary function tests.

upper limits of normal.[4 21] To our knowledge, no study has investigated DMLP with this approach to interpret spirometry. With a few exceptions,[9 22 23] previous studies did not use clear criteria for PFT collection and interpretations nor articulate the criteria used to diagnose the underlying lung disease.[24 25]

Ioachimescu and Stoller developed ANNs with two hidden layers and 15,308 PFTs to classify the four respiratory patterns.[22] PFTs were labelled following ATS/ERS guidelines, with 43% being obstructive, 16.5% restrictive and 4.5% mixed physiological patterns. ANNs using four input parameters (area under the expiratory flow-volume curve and z-scores for FEV_1, FVC, FEV_1/FVC) yielded the highest accuracies (91% and 92% in the validations and test sets, respectively). The area under the expiratory flow-volume curves contributed significantly to the accuracies, when compared with ANN models that only included the FEV_1, FVC, FEV_1/FVC z-scores.[22] This was particularly true for classifying mixed defects.[22] This non-traditional metric is retrievable but not readily available. Unlike our study, biometrics were not included in the ANN but were implied in the z-scores.

Others have compared physicians' interpretation of PFTs and clinical diagnoses to a decision tree model built using data from 1420 patients, MATLAB 8.3, Statistics and Machine Learning Toolbox, with 10-fold internal cross-validation.[9] Inputs included full PFTs (absolute, percent predicted and z-scores for prebronchodilator and postbronchodilator spirometry, plethysmography for lung volumes and airway resistance, diffusing capacity), age, sex, body mass index, smoking pack-years, presence of cough, sputum and dyspnoea.[9 10] Given 50 cases, pulmonologists interpreted lung function patterns with 74.4%±5.9% accuracy, with lower rates for restrictive patterns. Conversely, the machine learning model had 100% classification accuracy. When asked to categorise the cases into specific diagnostic categories (eg, asthma, chronic obstructive pulmonary disease (COPD), neuromuscular, interstitial lung disease), machine learning achieved accuracies of only 82%, but still higher than the clinicians at 44.6%.[9]

A recent study compared a fully convoluted neural network (CNN), random forest model and traditional spirometry for classifying the COPD phenotypes of predominant airway versus predominant emphysema. Data came from the COPDGene study: 3926 participants had no airflow obstruction, 3901 had Global Initiative for Lung Diseases stages 1–4 and 1066 had preserved ratio impaired spirometry.[23] The COPD phenotypes were labelled according to computer aided quantitative analysis of CT chest imaging. The CNN and random forest model were trained using all the datapoints in the expiratory flow-volume curve. Participants were split 80% for training and 20% for validation. The CNN significantly outperformed the random forest classifier and traditional spirometry (FEV_1/FVC and %predicted FEV_1).[23] A strength of this study, like the study by Ioachimescu,[22] is that the models learnt from all the datapoints in the expiratory flow-volume curve.

Limitations

While our dataset included PFTs from the full spectrum of respiratory defects and a wide range of abnormal findings, it is a limitation as the data came from a single centre. As a tertiary referral centre and the major lung transplant centre in Canada, our data were collected mostly from lung transplant recipients who had higher tests-to-patient ratios and a higher prevalence of restrictive defects compared with other patient cohorts. The imbalance between the number of tests among the four lung physiology patterns may have skewed the variability of our dataset.

While our samples were labelled using the current clinical gold standard (spirometry and plethysmography), this method is imperfect. The FEV_1 is limited in detecting small airway function and early airflow obstruction.[26] The contribution from the small airways to total airway resistance is low unless advanced or severe small airway obstruction is present.[27] Conversely, flow-volume loops demonstrate a complete representation of flow in regions where the small airways are not as distended as they are in the first second of forced expiration.[28] These are important limitations in the current conventional labelling system. Inclusion of data from the entire flow-volume loop may improve the detection of small airway or early-stage abnormalities and will be included in future deep learning models.

Lastly, our laboratory uses the Canadian reference equations to calculate percent predicted values.[21] The application of different reference equations can alter the 'true' label of the PFT from normal to abnormal; this is another limitation. The use of reference equations that are most appropriate for the specific patient population should be considered, and the MLP model retrained.

CONCLUSION

We developed a DMLP model to classify lung function patterns using biometrics and spirometry with comparable accuracies to full PFTs inclusive of plethysmography and biometrics. Hand-held spirometers are affordable and widely used as stand-alone diagnostic tools in primary care and outpatient settings. Implementation of the DMLP model into the software of spirometers can facilitate screening of patients with suspected lung disease. Implementation of the model will improve access to high calibre healthcare for patients who cannot perform or access diagnostic laboratories for full PFT with plethysmography. It will particularly benefit patients who live in regions of the world where only spirometry is available, and thus improve healthcare equity. It is also anticipated to improve patient outcomes by focusing subsequent investigations, such as full PFTs for patients identified by the DMLP model to have restrictive or mixed obstructive-restrictive defects, to facilitate earlier diagnoses, leading to reduced healthcare expenditure.

Contributors AM and TX conducted the research, performed the data analysis and drafted the manuscript. JKYW maintained the research ethics protocol, developed the standard operating procedures and ensured quality control of pulmonary function data. NB, HK, NV and DR contributed to data collection and edited the manuscript. CMR refined the research protocol, ensured quality control of pulmonary function data and edited the manuscript. SV developed the research plan, performed, oversaw the data analysis and drafted the manuscript. C-WC developed the concept, study protocol and oversaw all aspects of the project. C-WC is the guarantor and accepts full responsibility for the conduct of the study; she had access to the data, and controlled the decision to publish.

Funding The study is supported by a grant-in-aid from the Lung Health Foundation, the Pettit Block Term Grants, the CIHR/NSERC Collaborative Health Research Program (grant # 415013) and the Ajmera Foundation Multi-Organ Transplant Innovation Fund. AM was supported by an Amgen Scholarship. HK was supported by the scholarship funded by the Nakayama Foundation for Human Science. DR receives research support from the Sandra Faire and Ivan Fecan Professorship in Rehabilitation Medicine. We thank the Registered Cardio-Pulmonary Technologists at Toronto General Hospital for helping to conduct the study, members of C-WC's laboratory for collecting and maintaining the research data, and the 2019 University of Toronto PFT Committee for their work on developing the University of Toronto Guidelines for PFT Interpretation, eighth Edition

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the University Health Network (protocols 17-5652 and 17-5373) and University of Toronto Research Ethics Board

(protocol 376870). Written informed consent was obtained from all participants. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data sharing will be available in response to a written request following review and approval by the responsible institutions to ensure appropriate guidelines are met.

ORCID iDs
Joyce K Y Wu http://orcid.org/0000-0002-4113-2531
Natalia Belousova http://orcid.org/0000-0002-2237-4172
Haruna Kitazawa http://orcid.org/0000-0001-7234-1275
Nick Vozoris http://orcid.org/0000-0003-1670-1592
Dmitry Rozenberg http://orcid.org/0000-0001-8786-9152
Clodagh M Ryan http://orcid.org/0000-0002-2372-965X
Shahrokh Valaee http://orcid.org/0000-0001-6254-1660
Chung-Wai Chow http://orcid.org/0000-0001-9344-8522

REFERENCES

1 Tang Y, Zhang M, Feng Y, *et al*. 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.
2 Korten I, Zacharasiewicz A, Bittkowski N, *et al*. Asthma control in children: body plethysmography in addition to spirometry. *Pediatr Pulmonol* 2019;54:1141–8.
3 Stanojevic S, Kaminsky DA, Miller MR, *et al*. ERS/ATS technical standard on interpretive strategies for routine lung function tests. *Eur Respir J* 2022;60. doi:10.1183/13993003.01499-2021. [Epub ahead of print: 13 07 2022].
4 Graham BL, Steenbruggen I, Miller MR, *et al*. Standardization of spirometry 2019 update. An official American Thoracic Society and European Respiratory Society technical statement. *Am J Respir Crit Care Med* 2019;200:e70–88.
5 Johns DP, Walters JAE, Walters EH. Diagnosis and early detection of COPD using spirometry. *J Thorac Dis* 2014;6:1557–69.
6 Aaron SD, Boulet LP, Reddel HK, *et al*. Underdiagnosis and overdiagnosis of asthma. *Am J Respir Crit Care Med* 2018;198:1012–20.
7 Sahin D, Übeyli ED, Ilbay G, *et al*. Diagnosis of airway obstruction or restrictive spirometric patterns by multiclass support vector machines. *J Med Syst* 2010;34:967–73.
8 Quanjer PH, Stanojevic S, Cole TJ, *et al*. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012;40:1324–43.
9 Topalovic M, Das N, Burgel P-R, *et al*. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J* 2019;53:1801660.
10 Topalovic M, Laval S, Aerts J-M, *et al*. Automated interpretation of pulmonary function tests in adults with respiratory complaints. *Respiration* 2017;93:170–8.
11 Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000;43:3–31.
12 Jat DS, Dhaka P, Limbo A. Applications of statistical techniques and artificial neural networks: a review. *Journal of Statistics and Management Systems* 2018;21:639–45.
13 Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;15:20170387.
14 Harradon M, Druce J, Ruttenberg B. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv* 2018; 1802;00541.
15 Miller MR*et al*. Standardisation of spirometry. *Eur Respir J* 2005;26:319–38.
16 Day L, Faughnan M, Furlott H. *Guidelines for PFT interpretation*. 8th edn, 2019.
17 Bergstra J, Bengio Y. Random search for hyperparameter optimization. *J Mach Learn Res* 2012;13:281–305.
18 Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test* 2018;2:249–62.
19 Nguyen QH, Ly H-B, Ho LS, *et al*. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math Probl Eng* 2021;2021:1–15.
20 Seo S, Kim Y, Han H-J, *et al*. Predicting successes and failures of clinical trials with outer product-based convolutional neural network. *Front Pharmacol* 2021;12:670670.
21 Gutierrez C, Ghezzo RH, Abboud RT, *et al*. Reference values of pulmonary function tests for Canadian Caucasians. *Can Respir J* 2004;11:414–24.
22 Ioachimescu OC, Stoller JK. An alternative spirometric measurement. Area under the expiratory flow-volume curve. *Ann Am Thorac Soc* 2020;17:582–8.
23 Bodduluri S, Nakhmani A, Reinhardt JM, *et al*. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. *JCI Insight* 2020;5:e132781.
24 Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics J* 2019;25:811–27.
25 Das DK, Chakraborty C, Bhattacharya PS. Automated screening methodology for asthma diagnosis that ensembles clinical and spirometric information. *J Med Biol Eng* 2016;36:420–9.
26 Johns DP, Das A, Toelle BG, *et al*. Improved spirometric detection of small airway narrowing: concavity in the expiratory flow-volume curve in people aged over 40 years. *Int J Chron Obstruct Pulmon Dis* 2017;12:3567–77.
27 Kakavas S, Kotsiou OS, Perlikos F, *et al*. Pulmonary function testing in COPD: looking beyond the curtain of FEV1. *NPJ Prim Care Respir Med* 2021;31:23.
28 Bhatt SP, Bhakta NR, Wilson CG, *et al*. New spirometry indices for detecting mild airflow obstruction. *Sci Rep* 2018;8:17484.