Machine learning for nocturnal diagnosis of chronic obstructive pulmonary disease using digital oximetry biomarkers

Jeremy Levy1,2, Daniel Álvarez3,4,5, Felix del Campo3,4,5 and Joachim A Behar2,*

1 Faculty of Biomedical Engineering, Technion Institute of Technology, Haifa, Israel
2 Faculty of Electrical Engineering, Technion Institute of Technology, Haifa, Israel
3 Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
4 Pneumology Department, Río Hortega University Hospital, Valladolid, Spain
5 Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Valladolid, Spain

* Author to whom any correspondence should be addressed.
E-mail: jbehar@technion.ac.il

Keywords: chronic obstructive pulmonary disease, oximetry digital biomarkers, machine learning, medicine during sleep

Abstract

Objective. Chronic obstructive pulmonary disease (COPD) is a highly prevalent chronic condition. COPD is a major cause of morbidity, mortality and healthcare costs globally. Spirometry is the gold standard test for a definitive diagnosis and severity grading of COPD. However, a large proportion of individuals with COPD are undiagnosed and untreated. Given the high prevalence of COPD and its clinical importance, it is critical to develop new algorithms to identify undiagnosed COPD. This is particularly true in specific disease groups in which the presence of concomitant COPD increases overall morbidity/mortality such as those with sleep-disordered breathing. To our knowledge, no research has looked at the feasibility of automated COPD diagnosis using a data-driven analysis of the nocturnal continuous oximetry time series. We hypothesize that patients with COPD will exert certain patterns and/or dynamics of their overnight oximetry time series that are unique to this condition and that may be captured using a data-driven approach.

Approach. We introduce a novel approach to nocturnal COPD diagnosis using 44 oximetry digital biomarkers and five demographic features and assess its performance in a population sample at risk of sleep-disordered breathing. A total of n=350 unique patients’ polysomnography (PSG) recordings were used. A random forest (RF) classifier was trained using these features and evaluated using nested cross-validation.

Main results. The RF classifier obtained $F_1 = 0.86 \pm 0.02$ and AUROC $= 0.93 \pm 0.02$ on the test sets. A total of 8 COPD individuals out of 70 were misclassified. No severe cases (GOLD 3–4) were misdiagnosed. Including additional non-oximetry derived PSG biomarkers resulted in minimal performance increase.

Significance. We demonstrated for the first time, the feasibility of COPD diagnosis from nocturnal oximetry time series for a population sample at risk of sleep-disordered breathing. We also highlighted what set of digital oximetry biomarkers best reflect how COPD manifests overnight. The results motivate that overnight single channel oximetry can be a valuable modality for COPD diagnosis, in a population sample at risk of sleep-disordered breathing. Further data is needed to validate this approach on other population samples.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a highly prevalent chronic condition with a prevalence of 11.8% (95% confidence interval: 11.2–12.5) as estimated in the general population aged 40 years or older in all 17 regions of Spain (Soriano et al 2020) with similar estimate globally (Adeloye et al 2015). COPD is characterized by persistent airflow limitation that is usually progressive and an enhanced chronic inflammatory
response to noxious particles or gases in the airways and the lungs (Singh et al. 2019). COPD is a major cause of morbidity, mortality and healthcare costs. Etiological factors of COPD include aging, indoor and outdoor air pollution, and history of smoking (Rice and Malhotra 2015). COPD is suspected based on the clinical presentation of symptoms such as dyspnea, chronic cough, or sputum production, reporting history of exposure to risk factors with mainly tobacco (Vogelmeier et al. 2017). Diagnosis is confirmed if the ratio of forced expiratory volume within one second to forced vital capacity \( (\text{FEV}_1/\text{FVC}) \) is less than 0.70 in post-bronchodilator spirometry. Spirometry is the gold standard test for a definitive diagnosis and severity grading of COPD (Singh et al. 2019).

### 1.1. COPD an underdiagnosed condition

A large proportion of individuals with COPD are undiagnosed and untreated (Diab et al. 2018). Gershon et al. (2018) reported 13.7% undiagnosed COPD cases in a Canadian adult (aged \( \geq 40 \) years) random population-based sample \((n = 1,403 \text{ participants})\). This incidence was over 74.7% undiagnosed among the COPD patients in a Spanish adult population \((n = 9,092)\), as reported by Soriano et al. (2020). Çolak et al. (2017) recently found that individuals with undiagnosed, asymptomatic COPD had an increased risk of exacerbations and pneumonia. In addition, undiagnosed subjects showing symptoms had also an increased risk of death. These authors suggest that novel techniques for early COPD diagnosis and treatment are needed. Given the high prevalence of COPD and its clinical importance, it is critical to develop new approaches to identify undiagnosed COPD, especially in specific groups at risk, such as those with sleep disorder breathing.

### 1.2. COPD diagnosis and management

According to GOLD, post-bronchodilator spirometry is the standard technique to confirm COPD (Singh et al. 2019). Alternative systems for COPD diagnosis and management have been suggested. For example, an AI powered mobile health system using a spirometer has been suggested to support the diagnosis of patients in remote areas (Gurbeta et al. 2018). Phone questionnaires (Bischoff et al. 2012) and software applications using artificial intelligence (Boer et al. 2018) have been proposed to manage exacerbations. Furthermore, automated analysis of capnogram (Mieloszyk et al. 2014), computerized tomography images (Revel et al. 2008, Galbán et al. 2012), and text mining in the electronic health record (Akgün et al. 2020) have been evaluated for phenotyping COPD patients and characterizing severity. In a recent perspective paper (Behar 2019) we motivated to use overnight physiological recordings for the study, diagnosis, and monitoring of non-sleep specific conditions. In the case of COPD, oximetry is of particular interest as it reflects respiratory function, and recording it overnight (versus daytime) is the only option to obtain good quality continuous physiological data.

### 1.3. COPD and sleep-disordered breathing

COPD is associated with other morbid conditions such as obstructive sleep apnea (OSA). Both OSA and COPD are highly prevalent diseases. For patients with OSA, diagnosis of COPD is critical to identify the overlap.

---

**Figure 1.** Example of desaturation characteristics. (a) \( \text{SpO}_2 \) signal of COPD patient. (b) Thermistor airflow of the same COPD patient. (c) \( \text{SpO}_2 \) signal of non-COPD patient with OSA. (d) Thermistor airflow of the same patient. The black horizontal lines show the median oxygen saturation.
syndrome (OVS) that consists of OSA and COPD concomitantly (Flenley 1985). OVS occurs in an estimated 1 in 10 patients having one of the two conditions (Malhotra et al 2018). A recent meta-analysis of 17 articles highlighted that cardiovascular risk was significantly increased in OVS versus COPD or OSA alone (Xu et al 2020). Therefore, an early diagnosis of COPD is essential for effective treatment and a reduction in mortality of OSA patients. Unfortunately, in patients with suspected OSA, existing guidelines do not state the need for systematic respiratory functional assessment. A pulmonary evaluation would be particularly relevant for patients with smoking history, obesity, or those showing major respiratory symptoms, such as dyspnea (Lemarié et al 2010).

1.4. Manifestation of COPD on nocturnal oxygen saturation time series
Nocturnal desaturations are frequent in COPD patients, being more common in the most severe cases and particularly in patients with the chronic bronchitis phenotype. These nocturnal desaturations are a consequence of a disease-related mechanism of hypoventilation during sleep time. During daytime, individuals with COPD can suffer from desaturations related to efforts (walking, exercise) as a consequence of ventilation-perfusion mismatch (Krieger 2005). Overnight oximetry drops predominantly occur during rapid eye movement sleep and commonly show night-to-night variability (Buekers et al 2019). COPD and OSA are characterized by different hypoxemia models: OSA individuals show an intermittent pattern of desaturations during sleep (figure 1). In advanced COPD patients, it is common to observe overnight chronic hypoxemia in individuals with no primary sleep disorders (Budhiraja et al 2015), as shown in figure 1. The airflow signal is also shown, in (b) and (d), to support that the desaturation in (a) is not an apnoeic event. It was reported that up to 70% of COPD patients with daytime saturations in the range of 90%–95% had significant nocturnal hypoxemia (Chaouat et al 1997, Lewis et al 2009) and a lower mean overnight oxygen saturation as compared to controls (Valipour et al 2011). The overall prevalence of nocturnal desaturations in COPD patients, defined as desaturations below 90% for more than 5 min in Fletcher et al (1987), or as more than 30% of the night with a saturation level less than 90% in (Lewis et al 2009), was reported to vary from 27% to 49.2% (Fletcher et al 1987, Lewis et al 2009). In the recent randomized trial by Lacasse et al (2020), 79% of COPD patients showed an oxygen saturation of less than 90% for at least 30% of their overnight recording time. However, these estimations are very limited because relying on human observers to eyespot certain events (Chaouat et al 1997, Valipour et al 2011, Buekers et al 2019) or algorithms detecting simple predefined oximetry patterns (Fletcher et al 1987) (Lacasse et al 2020).

1.5. The knowledge gap
To our knowledge, no research has looked at the feasibility of automated COPD diagnosis using a data-driven analysis of the nocturnal continuous oximetry time series, either alone or when found concomitantly with another breathing disorder such as OSA.

1.6. Hypothesis and objectives
We hypothesize that patients with COPD will exert certain patterns and/or dynamics of their overnight oximetry time series that are unique to this condition and that may be captured using a data-driven approach. We introduce a novel approach to nocturnal COPD diagnosis using a machine learning (ML) model trained
with nocturnal oximetry digital biomarkers (Levy et al 2020) and assess its performance on a population sample of 350 individuals with suspected sleep disorder breathing.

2. Methods

A block diagram describing the steps in elaborating the ML model is shown in figure 2. The model takes the raw data as an input, performs a preprocessing step, extracts the digital oximetry biomarkers thanks to the pobm toolbox developed by Levy et al (2020), and then performs the classification task for a binary output of COPD versus non-COPD. The pobm toolbox is a Python library that enables to engineer oximetry digital biomarkers from continuous oximetry time series.

2.1. Database

A total of 350 oximetry recordings were obtained during in-lab polysomnography (PSG). This database is described in the original work of Andrés-Blanco et al (2017) which aimed at assessing the feasibility of automated OSA diagnosis from oximetry in patients with COPD. Briefly, this database consists of a total of \( n = 350 \) unique patients’ polysomnography (PSG) recordings. A total of 70 patients were originally referred to the pneumology outpatient facilities due to symptoms indicative of the COPD disease and then referred to the sleep clinic because of suspected sleep disorder breathing. The remaining 280 patients were referred to the sleep clinic because of suspected for sleep-disordered breathing. The latter group will be assumed to be ‘non-COPD’ as these patients had no history or signs of COPD. All participants showed high-to-moderate clinical suspected of sleep-disordered breathing due to the presence of at least one of the following symptoms: daytime hypersomnolence, loud snoring, nocturnal choking and awakenings, and/or apnoic events, reported by the patient or a bed partner. As per protocol, they were referred for PSG in the sleep unit of the Rio Hortega University Hospital, in Valladolid, Spain. The COPD confirmed individuals were subjects aged 35 years and older, current or ex-smokers with a smoking history of at least 10 packs/years, referred to the pneumology outpatient facilities due to symptoms indicative of COPD disease. Complete pulmonary function assessment (Master screen PFT, Jaeger) was conducted for COPD patients, including pre-and post-bronchodilator spirometry, lung volumes, and lung diffusion capacity. The threshold used to confirm COPD from spirometry was FEV1/FVC < 0.70. Standard in-lab PSG was carried out using a PSG E-series by Compumedics (Compumedics Limited, Victoria, Australia). Among patients with COPD (\( n = 70 \)), different subgroups were defined in terms of airflow limitations according to the global initiative for chronic obstructive lung diseases (GOLD) (Singh et al 2019): GOLD 1 (20.0%, \( n = 14 \)), GOLD 2 (65.7%, \( n = 46 \)), GOLD 3 (12.9%, \( n = 9 \)) and GOLD 4 (1.4%, \( n = 1 \)). GOLD 1 refers to a mild airflow limitation severity, while GOLD 4 means very severe airflow limitation (Vogelmeier et al 2017).

For OSA diagnosis, the apnea hypopnea index (AHI) was determined manually by a sleep expert following the AASM 2012 guidelines (Thornton et al 2012). OSA severity was defined as mild (\( 5 \leq AHI < 15 \)), moderate (\( 15 \leq AHI < 30 \)) or severe (\( AHI \geq 30 \)).
2.2. Preprocessing

For the SpO₂ time series, the Delta Filter (Taha et al 1997, Levy et al 2020) was applied: all samples with values larger than 100 or smaller than 50 were considered non-physiological and excluded. Then a median filter of length 9 s was applied to remove sharp changes (Deviene et al 2019, Levy et al 2020). An example of preprocessing is shown in figure S1 (available online at stacks.iop.org/PMEA/42/054001/mmedia).

2.3. Feature engineering

A total of 58 different features were computed (table S1). These include 5 demographic features, 9 common, non-oximetry, PSG-derived features, and 44 oximetry digital biomarkers engineered from the continuous SpO₂ time series (Levy et al 2020). The body mass index (BMI) was omitted, as it is redundant given weight and height are available as individual features. Besides, desaturation biomarkers were computed in two different ways namely, with a relative threshold for short-term drop detection (denoted ‘relative’) and with a hard threshold for maintained hypoxemia or long-term drop detection (denoted ‘hard’). The relative threshold desaturation detector corresponds to the one used to compute the oxygen desaturation index in sleep medicine. A hard threshold means that a desaturation is detected when the oximetry signal falls below a defined and constant threshold value—here taken at 90%. The intuition behind the hard threshold detector is that it may enable the model to detect the longer hypoxic events that are characteristic of COPD while the relative desaturation detector enables the identification of the shorter and more frequent desaturations observed in OSA patients. In the case of OVS, short desaturations may be embedded within those events, as can be seen in figure 3 where there are four desaturations detected by the relative threshold (in red), whereas the hard threshold detected one longer desaturation (in green). In the case of a relative threshold, the maximum length of desaturation was set at 120 s. In the case of the hard threshold, there was no constraint on the desaturation length. Oximetry biomarkers are computed over two hours long windows (denoted ‘window’) as well as over the full recording length (denoted ‘overall’). For individual window, this process leads to a total of 118 oximetry biomarkers, which combined with the demographic and PSG features results in 132 features overall.

2.4. Machine learning

2.4.1. Database split

The database was separated into a training-validation set (80%) and test set (20%) using stratification with respect to the class COPD and non-COPD. Because of the imbalanced database, data augmentation was performed: each SpO₂ time series in the training set was decomposed into windows of two hours. For COPD patients, an overlap of one hour between consecutive windows was used. For non-COPD individuals, non-overlapping windows were used. The data augmentation procedure was used on the training set only. In the first step, individual windows of two hours were classified as COPD or non-COPD. As COPD is a chronic condition, a majority vote was then performed over the predicted labels of all the windows for a given recording in order to classify the patient as COPD or non-COPD.

2.4.2. Models

Four ML models were evaluated (table 1): model 1 uses the demographic features only, model 2 uses the SpO₂ biomarkers extracted by the pobm toolbox, model 3 uses the SpO₂ biomarkers and the demographics features. Finally, model 4 uses all the features i.e. including other PSG features, and is implemented in order to evaluate if there is value in using other standard PSG features versus oximetry alone.

| Model | Demographics | PSG | All SpO₂ Biomarkers | Number of features | Selected features |
|-------|--------------|-----|---------------------|-------------------|------------------|
| Model 1 | X | | | 5 | 5 |
| Model 2 | | X | | 118 | 38 |
| Model 3 | X | | | 123 | 35 |
| Model 4 | X | X | X | 132 | 35 |

2.4.3. Feature selection

Given the limited number of examples (n = 350) and the high number of features (up to 132 for model 4), it is important to reduce the dimensionality of the classification problem and see if this enables better performances to be reached. Since model 1 has a low number of features, no feature selection step was applied. For models 2, 3, and 4 feature selection was performed using minimum redundancy and maximum relevance (mRMR) (Peng et al 2005). This algorithm aims to maximize the following operator:
\[
\phi(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c) - \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j),
\]

where \(S\) is a subset of features, \(I(x_i, c)\) is the information of the feature \(x_i\) relative to the class \(c\), \(I(x_i, x_j)\) is the mutual information of features \(x_i\) and \(x_j\). This operator combines the Max-Relevance (first term), and the Min-Redundancy (second term). The set of features with the highest \(\phi\) will be the set of features selected.

### 2.5. Classifiers and cross-validation procedure

For each model, two classifiers were trained; Logistic regression (LR) as a baseline linear model and Random Forests (RF) to evaluate the benefit of nonlinear classification. The Python library scikit-learn was used. Hyperparameters were optimized using 5-fold cross-validation. A large random grid of hyperparameters was searched (see supplementary note 2). For each iteration of the cross-fold, training examples were divided into train and validation set with stratification by patient. Because of the low number of patients in a single test set (20% of the overall database i.e. 70 patients), a nested cross-fold validation approach was taken. This means that 5-fold cross-validation was performed 5 times, each time by rotating the train-test split. This was done to report the median and variance performance of the models on the test sets.

#### 2.5.1. Performance measures

The \(F_1\) measure, sensitivity (Se), specificity (Sp), negative predictive value (NPV), positive predictive value (PPV) and Kappa statistic are reported:

\[
F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN},
\]

\[
Se = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},
\]

\[
NPV = \frac{TN}{TN + FN}, \quad PPV = \frac{TP}{TP + FP},
\]

\[
Kappa = \frac{p_0 - p_e}{1 - p_e},
\]

where TP is the number of true positives, FP the number of false positives (FP), FN the number of false negatives, \(p_0\) is the relative observed agreement among raters and \(p_e\) is the hypothetical probability of chance agreement.
2.6. Statistical analysis
To evaluate whether an individual feature was discriminative between the COPD and non-COPD groups, the Wilcoxon rank-sum test was applied and a \( p \)-value cut-off at 0.05 was used. For demographics, the Chi2 test was applied for categorical variables. Violin plots are produced for the most discriminative features.

3. Results
A total of 407 patients were included in this study. Data recorded from 57 patients were excluded due to invalid PSG. Hence, 350 subjects (age: 55.9 \( \pm \) 6.7 years (SD), 83 females), with a BMI of 29.2 \( \pm \) 2.8 kg m\(^{-2}\) and AHI of 39.9 \( \pm \) 15.0 per hour of sleep were considered for this study.

### Table 2. Median (MED) and interquartile range (1st Quartile, 3rd Quartile) descriptive statistics of the population sample studied.

| Name              | COPD (\( n = 70 \)) | non-COPD (\( n = 280 \)) | \( p \)-value |
|-------------------|---------------------|--------------------------|---------------|
| Gender            | Male: 61 (87\%)     | Male: 206 (74\%)         | \( p > 0.05 \) |
|                   | Female: 9 (13\%)    | Female: 74 (26\%)        |               |
| Age               | 64.5 (59.3, 70.8)   | 54.0 (45.0, 64.0)        | \( p < 10^{-5} \) |
| Weight            | 84.0 (76.3, 94.0)   | 82.0 (73.0, 93.0)        | \( p < 0.05 \) |
| Height            | 167.5 (163.0, 176.0)| 170.0 (163.0, 176.0)     | \( p > 0.05 \) |
| Smoking status    | No smoker: 0 (0\%)  | No smoker: 83 (24\%)     | \( p < 10^{-5} \) |
|                   | Smoker: 25 (36\%)   | Smoker: 139 (40\%)       |               |
|                   | Ex-smoker: 45 (64\%)| Ex-smoker: 58 (16\%)     |               |

#### Figure 5. Violin plots for six discriminative (\( p < 0.05 \)) features for the COPD and non-COPD groups. (a) MED (window) (%); (b) AV (overall) (%); (c) AV (window)%; (d) \( DS_\text{h} \) (hard, window) (%/sec\(^2\)); (e) \( DS_\text{s} \) (hard, window) (%/sec); (f) Age (years).

2.6. Statistical analysis
To evaluate whether an individual feature was discriminative between the COPD and non-COPD groups, the Wilcoxon rank-sum test was applied and a \( p \)-value cut-off at 0.05 was used. For demographics, the Chi2 test was applied for categorical variables. Violin plots are produced for the most discriminative features.

3. Results
A total of 407 patients were included in this study. Data recorded from 57 patients were excluded due to invalid PSG. Hence, 350 subjects (age: 55.9 \( \pm \) 6.7 years (SD), 83 females), with a BMI of 29.2 \( \pm \) 2.8 kg m\(^{-2}\) and AHI of 39.9 \( \pm \) 15.0 per hour of sleep were considered for this study.

3.1. Preprocessing
The average number of non-physiological samples discarded by the preprocessing step was 63 per recording (0.2% of the total recording time) while the maximum was 2502 samples (10% of the total recording time). The median and interquartile range of valid total recording time was 7.36 [7.12–7.60]. The median of the AHI among COPD patients in the database was 34.7, while among non-COPD patients it was 35.2. Figure 4 presents the repartition of the GOLD levels (1–4) with respect to the OSA severity levels (mild, moderate and severe) (Thornton et al 2012). The median age for patients with COPD was higher than for the non-COPD individuals (64.5 versus 54 years old, \( p < 0.05 \)) and more were male (87% versus 74%, \( p < 0.05 \)).
3.2. Statistical analysis of the features

The Wilcoxon rank-sum test performed for individual features for the groups COPD and non-COPD, rejected the null hypothesis for 115 out of 132 features. Tables S2–S5 present summary statistics for individual feature for patients divided by GOLD levels. In particular, **AV** and **MD** window (**$M_ED$**) yielded the lowest p-values. For 17 features, including height (**$D_I$**) or **DI**, the null hypothesis could not be rejected. The ranking of the 20 features with the lowest p-value can be seen in figure S2. Summary statistics for demographic features are shown in table 2, along with the p-value from the Wilcoxon and Chi2 tests. Additionally, a heatmap of features correlation is shown in figure S3. This statistical analysis provides some insights about what features might be most discriminative between COPD and non-COPD patients. Figure 5 shows violin plots for the six individual features with the smallest p-values to be interpreted as the individual most discriminative features of COPD versus non-COPD. Table 3 defines the acronyms of these features.

### Table 3. Definition of the most individual discriminative features (figure 5) and set of most predictive features determined using the RF classifier (figure 8).

| Features  | Definition                                                                 | Unit |
|-----------|----------------------------------------------------------------------------|------|
| AV        | Blood oxygen saturation (**$SpO_2$**) mean                                   | %    |
| MED       | **$SpO_2$** median                                                         | %    |
| P x       | xth percentile **$SpO_2$** value.                                          | %    |
| ZC x      | Number of zero-crossing points at the x% **$SpO_2$** level.                 | nu   |
| LZ        | Lempel-Ziv complexity.                                                     | nu   |
| PRSAdx    | Phase-rectified signal averaging (PRSA) capacity, with d the fragment duration. | %    |
| AC        | Autocorrelation                                                           | %    |
| PSDtotal  | The integral of the power spectral density (PSD) function.                 | %    |
| ODlx      | The oxygen desaturation index.                                             | event/h |
| DLx       | Mean of desaturations length.                                              | sec  |
| Dlx       | Standard deviation of desaturations length.                               | sec  |
| DDmaxlx   | Standard deviation of desaturations depth.                                 | %    |
| DSx       | Mean of the desaturation slope.                                            | %/sec|
| DSx       | Standard deviation of the desaturation slope.                              | (%/sec)$^2$ |
| DA100g    | Desaturation area: mean of desaturation area under the 100% **$SpO_2$** level as baseline. | %/sec |
| DA100b    | Standard deviation of desaturation area under the 100% **$SpO_2$** level as baseline. | (%/sec)$^2$ |
| PODlx     | Time of oxygen desaturation event, normalized by the total recording time. | sec  |
| CTx       | Cumulative time below the x% oxygen saturation level.                      | %    |
| Gender    | Gender                                                                     | n.u. |
| Age       | Age                                                                        | years|
| Smoking status | Whether smoker, ex-smoker, or non-smoker.                        | n.u. |

**Figure 6.** Test sets ROC for the four models. Results are presented for the per patient classification. Each ROC is summarized by its median value (solid curve) and an envelope representing the minimal and maximal value obtained within the outer loop of the nested cross validation.
3.3. Feature selection

Using mRMR, a total of 38 features were selected for model 2 and 35 for models 3 and 4. The ranking of the selected features for models 2–4 is shown in figure S4. The feature with the highest score is the most-relevant least-redundant feature.

Table 4. Per patient classification results for the outer-loop i.e. over the test sets for the RF and LR classifiers. The median and standard deviation of each performance measure over the five outer loops is presented. Underlined data represent the best performance for each performance measure.

|        | AUROC   | F1      | Kappa   | AUROC   | F1      | Kappa   |
|--------|---------|---------|---------|---------|---------|---------|
| Model 1| 0.69 ± 0.13 | 0.60 ± 0.06 | 0.39 ± 0.12 | 0.64 ± 0.06 | 0.56 ± 0.10 | 0.35 ± 0.17 |
| Model 2| 0.80 ± 0.09 | 0.72 ± 0.10 | 0.62 ± 0.05 | 0.74 ± 0.08 | 0.61 ± 0.13 | 0.51 ± 0.08 |
| Model 3| 0.93 ± 0.02 | 0.86 ± 0.02 | 0.81 ± 0.06 | 0.82 ± 0.11 | 0.80 ± 0.08 | 0.64 ± 0.05 |
| Model 4| 0.94 ± 0.01 | 0.86 ± 0.02 | 0.83 ± 0.06 | 0.84 ± 0.09 | 0.81 ± 0.10 | 0.65 ± 0.14 |

Table 5. Per patient classification results for the RF classifier and over the test sets. The median and standard deviation of each performance measure over the five outer loops are presented. Underlined data represent the best performance for each performance measure.

|        | Se GOLD1 | Se GOLD2 | Se GOLD3 | Se GOLD4 | Sp     | NPV    | PPV    |
|--------|----------|----------|----------|----------|--------|--------|--------|
| Model 1| 0.29 ± 0.08 | 0.64 ± 0.06 | 0.67 ± 0.15 | 1.00 ± 0.00 | 0.85 ± 0.02 | 0.80 ± 0.05 | 0.42 ± 0.15 |
| Model 2| 0.71 ± 0.11 | 0.76 ± 0.06 | 0.78 ± 0.02 | 1.00 ± 0.00 | 0.79 ± 0.06 | 0.75 ± 0.10 | 0.70 ± 0.12 |
| Model 3| 0.80 ± 0.03 | 0.96 ± 0.03 | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.90 ± 0.03 | 0.95 ± 0.02 | 0.81 ± 0.07 |
| Model 4| 0.80 ± 0.03 | 0.96 ± 0.01 | 1.00 ± 0.00 | 1.00 ± 0.00 | 0.91 ± 0.05 | 0.95 ± 0.02 | 0.85 ± 0.03 |

Figure 7. Test sets confusion matrix for model 3. The OSA and GOLD levels are specified.

Table 6. Mean of most important features for the TP, FN, TN and FP for the per window classification using model 3.

| Feature            | TP    | FN    | TN    | FP    |
|--------------------|-------|-------|-------|-------|
| Age                | 65.4  | 57.4  | 53.1  | 59.6  |
| Smoking status     | 2.0   | 1.5   | 1.0   | 1.2   |
| CT_{mean} (overall)| 46.8  | 42.0  | 14.2  | 18    |
| MED (window)       | 89    | 92    | 94    | 92    |
| DS_{rel} (relative, window) | 0.08  | 0.06  | 0.07  | 0.09  |
3.3.1. Classification
The confusion matrix for the per window classification on the test sets is provided in table S6. The mean and standard deviation of the models’ optimized hyperparameters are provided in table S7. Table 4 presents the results on the test sets of the nested cross fold validation procedure, for models 1–4, for both RF and LR classifiers, and for the per patient classification. Table 5 presents more performance measures for the RF classifier. The receiver operating characteristic (ROC) curves are shown in figure 6. Table S8 presents the results on the test sets for the per window classification. Figure 7 presents the confusion matrix over the test sets, per patient. The per window mean of the five most important features for the TN, TP, FP, and FN are summarized in table 6.

3.4. Features importance
Figure 8 presents the feature importance ranking of the RF classifier for model 3 that is using oximetry biomarkers and demographics. Table 3 defines all the features shown on figure 8. The two most important features were demographic features, namely age and smoking status. The two most important oximetry biomarkers were \( \text{MED overall} \) and \( \text{CT overall} \).

4. Discussion
4.1. Models performance
Our experiments showed that using the RF based model performed significantly better than the LR based model (table 4). This demonstrates that nonlinear relationships between features exist and that using a nonlinear classifier improves the model performance. Model 3 and 4 performed best with \( F_1 = 0.86 \pm 0.02 \), AUROC = 0.93 ± 0.02 and \( F_1 = 0.86 \pm 0.02 \), AUROC = 0.94 ± 0.01 respectively. The performance of model 3 was thus very close to model 4 which suggests that the diagnosis of COPD using single channel oximetry competes with a diagnosis that would use additional PSG biomarkers. Comparing features selected by all the different models is not possible since they do not contain the same features. Yet, considering model 2, 3 and 4 which contain the oximetry biomarkers we observe that MED (overall), AC(overall) and PSD\text{\_\text{total}}(\text{window}) are selected by these three models.
4.2. Interpretability of features importance

It is known that smoking is a risk factor that is highly important in COPD. For age, we observed that our COPD population was older than the non-COPD population (table 6). The third most important feature is MED (overall) and reflects that COPD patients will have a lower SpO2 median value than non-COPD. The fourth most important feature is CT90(overall), the cumulative time under the 90% baseline. This feature captures the long hypoxic events in the signal. Previous research such as the one of Lewis et al (2009) had also reported a high CT90 in COPD patients which is coherent with our findings. In addition, a number of desaturations features ranked high (13 in the top-30 features), both when using the relative (DSs, DA100s) and hard (DA100s, DD maxs) desaturation thresholds. This reflects that the model relies on the desaturations slope and area in order to make the prediction. Figure 1 shows an example of desaturations characteristic of a COPD patient and a non-COPD patient with OSA. For the latest, the slope and area of the desaturations are close to each other. Indeed, many short consecutive desaturations can be observed. In this case, the features DSs, DA100s, and DD maxs will have low values. For the COPD patient, a single desaturation with greater area can be seen. In the case of COPD, the desaturations have a more variable length which will increase the standard deviation of the slope and the area of the desaturations which will lead to higher values of DSs, DA100s, and DD maxs. Finally, overall features such as AC(overall), MED(overall) and AV(overall) had a high feature importance which reflects that the classifier harnesses contextual information from the overall recording. Figure S5 presents the ranking of feature importance for the four models.

4.3. Error analysis

As it can be seen in figure 7, all FP had mild to severe OSA which highlights that the classifier may be confused by the effect of repetitive desaturations caused by OSA. Secondly, all the FN belong to the GOLD level 1 and 2 that is no severe COPD cases (GOLD 3–4) were missed by the classifier. We noted that all individuals in the COPD group without OSA were correctly classified (5/5). Furthermore, according to table 6, the model misclassified COPD windows (FN) especially for young patients (57.4 years against 65.4 years) and with a low CT90 (overall) (42.0% against 46.8%) reflecting a lower number of hypoxic events in FN. The FP were relatively older than the non-COPD patients in the database. Thus, age may be biasing the classification for some examples.

4.4. Limitations

The non-COPD group might actually contain some individuals with COPD, although there was no previous history of COPD in these patients medical record, i.e., neither symptoms nor exposure to risk factors, which are needed to suspect COPD and refer for spirometry according to the guideline (Vogelmeier et al 2017). Yet, this represents the main limitation of our work and motivates furthering this research by recording a new cohort where all the population sample undergoes a spirometry test. A second important limitation is the relatively low number of patients included in the study (n = 300), in particular, the ones having COPD (70 patients). Finally, it is important to extend this research to other risk groups than the one with suspected sleep disorder breathing.

5. Conclusion

Our research makes a number of novel scientific contributions. First, we demonstrated, for the first time, the feasibility of COPD diagnosis from nocturnal oximetry time series in a population sample at risk of sleep-disordered-breathing. We highlighted what digital oximetry biomarkers best reflect how COPD manifests overnight. In particular, MED(overall), CT90(overall), AC(window) and DSs(relative, window) were found to be the most discriminative. Finally, we show that including additional PSG biomarkers only slightly improves the classifier performance (AUROC of 0.94 ± 0.01 versus 0.93 ± 0.02 without). This motivates single channel oximetry is a valuable option for overnight COPD diagnosis.

Acknowledgments

We acknowledge the financial support of the Technion Machine Learning & Intelligent Systems Center (MLS). This work has been partially supported by ‘CIBER en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)’ through ‘Instituto de Salud Carlos III’ co-funded with FEDER funds. D Álvarez is supported by a ‘Ramón y Cajal’ grant (RYC2019-028566-I) from the ‘Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación’ co-funded by the European Social Fund.
Competing interests

J B holds shares in SmartCare Analytics Ltd. The remaining authors declare no competing interests.

ORCID iDs

Jeremy Levy  https://orcid.org/0000-0001-7498-6413
Daniel Álvarez  https://orcid.org/0000-0003-1027-2395
Felix del Campo  https://orcid.org/0000-0002-4554-2167
Joachim A Behar  https://orcid.org/0000-0001-5956-7034

References

Adeloye D et al 2015 Global and regional estimates of COPD prevalence: systematic review and meta-analysis J. Glob. Health 5 020415
Akgün K M, Sigel K, Cheung K H, Kidwai-Khan F, Bryant A K, Brandt C, Justice A and Crothers K 2020 Extracting lung function measurements to enhance phenotyping of chronic obstructive pulmonary disease (COPD) in an electronic health record using automated tools PLoS One 15 e0227730
Andrés-Blanco A M, Álvarez D, Crespo A, Arroyo C, Cerezo-Hernández A, Gutiérrez-Tobal G C, Hornero R and Del Campo F 2017 Assessment of automated analysis of portable oximetry as a screening test for moderate-to-severe sleep apnea in patients with chronic obstructive pulmonary disease PLoS One 12 1–21
Behar J 2019 From sleep medicine to medicine during sleep: a new paradigm Sleep 43 zs279
Bischoff E WM A, Boer L M, Molema J, Akkermans R, Van Weel C, Vercoulen J H and Schermer T R J 2012 Validity of an automated telephonic system to assess COPD exacerbation rates Eur. Respir. J. 39 1090–6
Boer L M, van der Heijden M, van Kuijk N M E, Lucas P J F, Verroust J H, Assendelft W J J, Bischoff E W and Schermer T R 2018 Validation of ACCESS: an automated tool to support self-management of COPD exacerbations Int. J. COPD 39 3255–67
Budhiraja R, Siddiqi T A and Quan S F 2015 Sleep disorders in chronic obstructive pulmonary disease: etiology, impact, and management J. Clin. Sleep Med. 11 259–70
Buekers J, Theunis J, De Boever P, Vaes A W, Koopman M, Janssen E V M, Wouters E F M, Spruit M A and Aerts J M 2019 Wearable telephonic system to assess COPD exacerbation rates PLoS One 14 1–12
Çolak Y, Afzal S, Nordestgaard B G, Vestbo J and Lange P 2017 Prognosis of asymptomatic and symptomatic, undiagnosed COPD in the general population in Denmark: a prospective cohort study Lancet Respir. Med. 5 426–34
Deviaene M, Testelmans D, Buyse B, Borzée P, Van Huffel S and Varon C 2019 Automatic screening of sleep apnea patients based on the SpO2 signal IEEE J. Biomed. Heal. Inform. 23 607–17
Diab N, Gershon AS, Sin D D, Tan W C, Bourbeau J, Boulet L-P and Aaron S D 2018 Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease Am. J. Respir. Crit. Care Med. 198 1130–9
Flenley D C 1985 Sleep in chronic obstructive lung disease Sleep 6 565–8
Fletcher E C, Miller J, Divine G W, Fletcher J G and Miller T 1987 Nocturnal oxyhemoglobin desaturation in COPD patients with arterial hypoxaemia Eur. Respir. J. 10 1730–5
Çolak Y, Afzal S, Nordestgaard B G, Vestbo J and Lange P 2017 Prognosis of asymptomatic and symptomatic, undiagnosed COPD in the general population in Denmark: a prospective cohort study Lancet Respir. Med. 5 426–34
Deviaene M, Testelmans D, Buyse B, Borzée P, Van Huffel S and Varon C 2019 Automatic screening of sleep apnea patients based on the SpO2 signal IEEE J. Biomed. Heal. Inform. 23 607–17
Diab N, Gershon AS, Sin D D, Tan W C, Bourbeau J, Boulet L-P and Aaron S D 2018 Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease Am. J. Respir. Crit. Care Med. 198 1130–9
Flenley D C 1985 Sleep in chronic obstructive lung disease Clin. Chest Med. 6 651–61
Fletcher E C, Miller J, Divine G W, Fletcher J G and Miller T 1987 Nocturnal oxyhemoglobin desaturation in COPD patients with arterial oxygen tensions above 60 mm Hg Chest 92 604–8
Gallbl C J et al 2012 Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression Nat. Med. 18 1711–5
Gershon A S, Thiruchelvam D, Chapman K R, Aaron S D, Stanbrook M B, Bourbeau J, Tan W and To T 2018 Health services burden of undiagnosed and overdiagnosed COPD Chest 153 1336–46
Gurbeti A, Badnjievic A, Maksimovic M, Ormanovic-Mikićanin E and Sejdic E 2018 A telehealth system for automated diagnosis of asthma and chronic obstructive pulmonary disease J. Am. Med. Inform. Assoc. 25 1213–7
Krieger A C 2005 Perturbação respiratória durante o sono em doença pulmonar obstrutiva crônica J. Bras. Pneumol. 31 2
Lacasse Y et al 2020 Randomized trial of nocturnal oxygen in chronic obstructive pulmonary disease New Engl. J. Med. 383 1129–38
Lemarié E, Valeyre D and Housset B 2010 De nouvelles recommandations pour la prise en charge de la BPCO Eur. Radiol. 20 3010–23
Lewis C A, Fergusson W, Eaton T, Zeng I and Kolbe J 2009 Isolated nocturnal desaturation in COPD: prevalence and impact on quality of life and sleep Thorax 64 133–8
Malheiro A et al 2018 Research priorities in pathophysiology for sleep-disordered breathing in patients with chronic obstructive pulmonary disease: an official American Thoracic society research statement Am. J. Respir. Crit. Care Med. 197 289–99
Mielszynski R J, Verghese G C, Deitch K, Cooney B, Khalid A, Mirre-Gonzalez M A, Heldt T and Krauss B S 2014 Automated quantitative analysis of capnogram shape for COPD-normal and COPD-CHF classification IEEE Trans. Biomed. Eng. 61 2882–90
Peng H, Long F and Ding C 2005 Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy IEEE Trans. Pattern Anal. Mach. Intell. 27 1226–38
Revel M P, Faivre J, Remy-Jardin M, Dekken V, Duhamel A, Marquette C H, Tacelli N, Bakai A M and Remy J 2008 Automated lobar quantification of emphysema in patients with severe COPD Eur. Radiol. 18 2723–30
Rice M B and Malheiro A 2015 The air we breathe and Lung disease J. Thorac. Dis. 7 E245
Ruoff C and Rye D 2016 The ICSD-3 and DSM-5 guidelines for diagnosing narcolepsy: clinical relevance and practicality Curr. Med. Res. Opin. 32 1611–22
Singh D et al 2019 Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019 Eur. Respir. J. 53 1900164

Soriano J B et al 2020 Prevalence and determinants of COPD in Spain: EPISCAN II Arch. Bronconeumol. 57 61–9
Taha B H, Dempsey J A, Weber S M, Badr M S, Skatrud J B, Young T B, Jacques A J and Seow K C 1997 Automated detection and classification of sleep-disordered breathing from conventional polysomnography data Sleep 20 991–1001
Thornton A T, Singh P, Ruehland W R and Rochford P D 2012 AASM criteria for scoring respiratory events: interaction between apnea sensor and hypopnea definition Sleep 35 425–32
Valipour A, Lavie P, Lothaller H, Mikulic I and Burghuber O C 2011 Sleep profile and symptoms of sleep disorders in patients with stable mild to moderate chronic obstructive pulmonary disease Sleep Med. 12 367–72
Vogelmeier C F et al 2017 Global strategy for the diagnosis, management and prevention of chronic obstructive lung disease 2017 report: GOLD executive summary Respiratology 22 575–601
Xu J, Wei Z, Wang X, Li X and Wang W 2020 The risk of cardiovascular and cerebrovascular disease in overlap syndrome: a meta-analysis J. Clin. Sleep Med. 16 1199–207