# Nighttime features derived from topic models for classification of patients with COPD

Spina, Gabriele ; Casale, Pierluigi ; Albert, Paul S ; Alison, Jennifer ; Garcia-Aymerich, Judith ; Clarenbach, Christian F ; Costello, Richard W ; Hernandes, Nidia A ; Leuppi, Jörg D ; Mesquita, Rafael ; Singh, Sally J ; Smeenk, Frank W J M ; Tal-Singer, Ruth ; Wouters, Emiel F M ; Spruit, Martijn A ; den Brinker, Albertus C

DOI: https://doi.org/10.1016/j.compbiomed.2021.104322



Originally published at:
Spina, Gabriele; Casale, Pierluigi; Albert, Paul S; Alison, Jennifer; Garcia-Aymerich, Judith; Clarenbach, Christian F; Costello, Richard W; Hernandes, Nidia A; Leuppi, Jörg D; Mesquita, Rafael; Singh, Sally J; Smeenk, Frank W J M; Tal-Singer, Ruth; Wouters, Emiel F M; Spruit, Martijn A; den Brinker, Albertus C (2021). Nighttime features derived from topic models for classification of patients with COPD. Computers in Biology and Medicine, 132:104322.
DOI: https://doi.org/10.1016/j.compbiomed.2021.104322

# Nighttime features derived from topic models for classification of patients with COPD

Gabriele Spina [a,*], Pierluigi Casale [b], Paul S. Albert [c], Jennifer Alison [d], Judith Garcia-Aymerich [e], Christian F. Clarenbach [f], Richard W. Costello [g], Nidia A. Hernandes [h], Jorg D. Leuppi¨ [i], Rafael Mesquita [j], Sally J. Singh [k], Frank W.J.M. Smeenk [l], Ruth Tal-Singer [m], Emiel F. M. Wouters [j], Martijn A. Spruit [j,n], Albertus C. den Brinker [o]

[a] *HumanTotalCare, Data Science Department, Utrecht, the Netherlands*
[b] *Jheronimous Academy of Data Science, 'S-Hertogenbosch, the Netherlands*
[c] *School of Ageing and Chronic Disease, University Hospital Aintree, Liverpool, United Kingdom*
[d] *Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia*
[e] *CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain and Universitat Pompeu Fabra (UPF), Barcelona, Spain*
[f] *Department of Pulmonology, University Hospital Zurich, Zurich, Switzerland*
[g] *Department of Respiratory Medicine, Royal College of Surgeons, Dublin, Ireland*
[h] *Laboratory of Research in Respiratory Physiotherapy, Dept. of Physiotherapy, State University of Londrina (UEL), Londrina, Brazil*
[i] *University Department of Medicine, Cantonal Hospital Baselland and University of Basel, Basel, Switzerland*
[j] *Department of Research and Education, CIRO+, Centre of Expertise for Chronic Organ Failure, Horn and MUMC+ Dept. of Respiratory Medicine, Maastricht University Medical Center+, the Netherlands*
[k] *NIHR Leicester Respiratory Biomedical Research Centre, Dept of Respiratory Sciences, University of Leicester, Leicester, UK*
[l] *Dept. of Respiratory Medicine, Catharina Hospital, Eindhoven, the Netherlands*
[m] *GlaxoSmithKline R&D, King of Prussia, PA, USA*
[n] *REVAL - Rehabilitation Research Center, BIOMED - Biomedical Research Institute, Faculty of Medicine and Life Sciences, Hasselt University, Diepenbeek, Belgium, and NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht, the Netherlands*
[o] *Philips Research, Data Science Department, Eindhoven, the Netherlands*

A B S T R A C T

Nighttime symptoms are important indicators of impairment for many diseases and particularly for respiratory diseases such as chronic obstructive pulmonary disease (COPD). The use of wearable sensors to assess sleep in COPD has mainly been limited to the monitoring of limb motions or the duration and continuity of sleep. In this paper we present an approach to concisely describe sleep patterns in subjects with and without COPD. The methodology converts multimodal sleep data into a text representation and uses topic modeling to identify patterns across the dataset composed of more than 6000 assessed nights. This approach enables the discovery of higher level features resembling unique sleep characteristics that are then used to discriminate between healthy subjects and those with COPD and to evaluate patients' disease severity and dyspnea level. Compared to standard features, the discovered latent structures in nighttime data seem to capture important aspects of subjects sleeping behavior related to the effects of COPD and dyspnea.

## 1. Introduction

Chronic obstructive pulmonary disease (COPD) is a respiratory disease characterized by chronic inflammation of the lung airways, and degradation of lung tissue which result in airow limitation [1]. It is caused, among others, by long-term exposure to irritating gases or particulate matter, most often deriving from cigarette smoke. COPD is one of the leading causes of mortality worldwide and imposes a signicant burden on patients' daily lives due to its symptoms and exacerbations that significantly reduce patients' quality of life [2]. COPD is estimated to affect over 380 million people worldwide and in the last decades has become a global health problem because of its increasing incidence and associated socio-economic costs [3,4]. Although detecting the disease at an early stage may help to provide the best symptom control, disease progression and outcomes, COPD remains largely underdiagnosed [5,6] and often misdiagnosed [7].

Artificial intelligence in respiratory medicine, and particularly in patients with COPD, has gained considerable interest over the last years thanks to its potential for improving the quality of medical services and addressing the shortcomings in the diagnosis by utilizing large datasets of symptoms, patient history, and diagnostic tests [8]. Several machine learning (ML) techniques, such as artificial neural networks [9,10], and multiple instance learning [11] have been used for automatic recognition of COPD based on high resolution computed tomography scanning, which enables the direct evaluation of the lungs and airways. Arpaia et al. [12] proposed a method based on particle swarm optimization to predict the condition of a subject with COPD from clinical parameters monitored monthly for one year. Newandee et al. [13] applied cluster analysis to identify the most severe patients with COPD in a mixture of healthy and COPD population for which heart rate, blood pressure, and respiration signals were recorded. Results demonstrated that these two groups could be differentiated with 99.0% accuracy. Furthermore, the four severity levels of COPD were classified with over 88.0% accuracy. Although promising, these methods are restricted in the sense that data were obtained under the expert supervision of a clinician and in controlled conditions often very dissimilar to those characteristic of daily life. Moreover, due to the time and resources required, these methods cannot be performed in large scale, limiting as well their applicability as screening tools.

In the last years unobtrusive sensor technologies have gradually permeated into patients' homes, making health-related data more available than ever before [14]. In this context, wearable devices could effortlessly assist and support caregivers [14] enabling new care services that could allow a paradigm shift from the established centralized healthcare model to a pervasive, user-centred and preventive health management [15]. In Ref. [16] sensor data were acquired by several external devices attached to a jacket worn by patients with COPD and used to classify the disease in different severity levels. Although the reported accuracy is 94% the system developed was rather cumbersome and difficult to use in clinical practice. In Ref. [17] daytime activity routines derived from data recorded by an unobtrusive activity monitor were used to discriminate between subjects with and without COPD with 86% of accuracy. Taking into account that sleeping hours may offer a better trade-off between patients' comfort, sensor unobtrusiveness and signal quality [15], the monitoring and classification of subjects with COPD using nighttime data is an interesting opportunity for future wearable applications. However, when analysing sleep, the features to be extracted are typically based on duration and continuity of sleep or structurally not obvious often relying on domain expertise [18]. One solution to overcome the shortcoming of handcrafted features is to use deep learning methods since they are adept at learning abstract features directly from the raw data [19]. Despite the enormous steps forward which this technique has taken, several studies touch on how the cost of the computational infrastructure required to efficiently learn models and data issues such as low volume, high sparsity and poor quality can limit the use of deep learning in clinical applications [20]. A model-free unsupervised methodology that is gaining popularity in sensor data processing is the implementation of topic modeling to extract meaningful information from daily life data. Its applicability spans from the discovery of activity routines [17,21], human activity recognition [22, 23] to behaviours analysis [24,25]. In this study, multimodal

sensor data recorded continuously by an activity monitor during nighttime are initially transformed into artificial "words" used to create a corpus of text documents in which each document represents the night of a subject. Latent Dirichlet Allocation (LDA) [26] is then used to discover the hidden semantic structures (sleep topics) that characterize the nights of COPD and healthy subjects and to describe each night with a probability distribution over the discovered sleep topics. The sleep topics are then used as data driven features to classify healthy subjects and those with COPD and to evaluate the severity of the disease for each subject. To the best of our knowledge, this is the first study that introduces the use of a topic modeling technique to extract powerful predictors from multisensory sleep data for COPD classification and assessment.

## 2. Latent Dirichlet Allocation

Topic modeling is an unsupervised machine learning technique capable of scanning a large and unstructured collection of documents to automatically cluster similar word patterns and infer topics that best characterize the text data [27]. LDA is an example of a topic model in which data are treated as observations arising from a generative probabilistic process [26]. In the context of text modeling, given a set of topics defined as distributions over words, the generative process populates the documents with words such that the documents have a particular desired thematic structure. Besidesb its generative process, LDA can also be used to calculate the hidden variables that likely generated the collection of documents. One of the ways to achieve this is to use variational inference to approximate the posterior distribution over the hidden variables defined by LDA. In a nutshell, variational inference posits a parametrized family of distributions over the hidden structure, and finds the member of that family that is closest to the posterior according to the Kullback-Leibler divergence. The intuition behind using LDA for sleep monitoring is that each night is a mixture of thematically coherent physiological measures just as a text document is a mixture of thematically coherent words. The graphical model for LDA in relation to the sleep of a cohort of subjects is shown in Fig. 1.

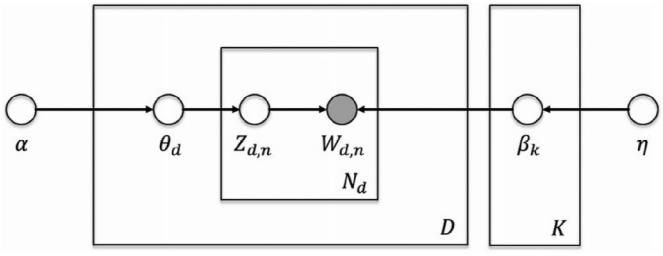


**Fig. 1.** Graphical model for LDA. Each node is a random variable, edges denote possible dependencies. The only observed variables (shaded) are the artificial words (*W*). The distribution of the words in a sleep topic (β) and the distribution of the sleep topics during a night (θ) depend only on the sleep topics hyperparameters η and α that control the mean shape and sparsity of the distributions. *Z* represents the symbol sleep topic assignment.

All the assessed nights ($d_{1:D}$) share the same set of sleep topics ($\beta_{1:K}$) that are defined as Dirichlet distributions over the observed set of artificial words (*W*) which are the terms of a fixed vocabulary. The observed artificial words (input of the model) are composed of multimodal measures recorded by the sensors of an activity monitor. Each assessed night exhibits sleep topics in different proportion providing an explicit finger print θ. In particular, each night is a different distribution ($\vartheta_{1:D}$) over the sleep topics activation probabilities that also follows a Dirichlet distribution. In such a model, the *N* artificial words ($W_{d,n_{(1:N)}}$) that compose the *D* nights are the only random variables observed and depend on the per word sleep topic assignment ($Z_{d,n}$) and all the $\beta_k$. Each sleep topic is composed indirectly of low-level sensor measures that belong, with a certain probability distribution, to different thematic areas. Our hypothesis is that nights related to different group of subjects with and without COPD would have a different distribution over the sleep topics that in turn would be composed of different distributions over artificial words defined combining discretized sensor measurements. The proposed methodology first extracts in a data driven fashion all the $\beta_k$ sleep topics using data from healthy subjects and patients with COPD and then, for each assessed night, it calculates a probabilistic feature vector θ composed of the histogram of activation probabilities of all the sleep topics $\beta_{1:K}$. These probabilistic features are used to classify the severity of the disease and the level of dyspnea in a larger cohort of patients. The dataset and the methodology developed are described in detail in the following sections.

## 3. Methods

### *3.1. Participants*

Data from 1384 patients from ten countries (United Kingdom, Ireland, The Netherlands, Germany, Switzerland, Italy, Spain, The United States of America, Brazil, and Australia) diagnosed with mild to very severe COPD were pooled from previous studies and considered for analysis. Extensive references to the previous studies utilizing these data can be found in Ref. [28]. Participants were included if they had COPD with a post-bronchodilator ratio of forced expiratory volume in the first second ($FEV_1$) to forced vital capacity ($FVC$) < 0.70 and were clinically stable (i.e. stable shortness of breath and sputum production). We used baseline data, i.e., recorded before that any specific intervention was undertaken. Centers from The Netherlands and UK also provided data on 66 healthy control subjects that were matched for age, gender, and BMI with a subgroup of 66 subjects with COPD. On the basis of a 1:1 multivariate matching, the closest possible case-control matches were determined. Subjects matched exactly for age and gender, the median error between BMI values of matching subjects was 0.58 [0.29–1.2] $kg/\ m^2$. Subject group characteristics are presented in Table 1. The data collection was approved by ethics committees at each of the participating centers, according to local regulations. Written informed consent was provided by all participants.

### *3.2. Data recordings*

Study participants wore the SenseWear Armband or Mini Armband activity monitors [29] (BodyMedia Inc., Pittsburgh, PA, USA) on the upper arm both during daytime and nighttime so that continuous, real-life data were collected in a natural environment. These devices included an accelerometer with various physiological sensors: a heat flux sensor, a galvanic skin response (GSR) sensor, a skin temperature (Temp) sensor, and a near-body ambient temperature sensor [30,31]. Data are sampled in 1-min epochs and together with demographic characteristics (such as gender, age, height, and weight) were used to estimate

metabolic equivalent of task (MET) [32] using proprietary algorithms developed by the manufacturer. The use of multisensory data in combination with pattern recognition algorithms ensured that the MET estimation is insensitive to noise and random motion artefacts [33]. For each minute, the device also recorded step count (SC), information about the sleeping status of a patient (0 = awake, 1 = sleeping), and posture (0 = lying down, 1 = not lying down). Nighttime sleep was defined as sleep that occurs between 21:00 and 06:00 of the following day [34]. For this analysis we considered only MET, Temp, GSR, SC, and Sleeping Status (SL) data within this time interval. Conventional sleeping features such as total night sleeping time, number of nocturnal sleeping bouts and duration of nocturnal sleeping bouts were also derived from the sleep status information provided by the sensor. Sleeping bouts were defined as consecutive minutes marked by the sensor as sleeping. Participants who wore the device for at least four nights (two weekdays, one Saturday and one Sunday) were included in the analysis [31]. In total 1059 patients with COPD and 66 healthy controls were included in the analysis. The number of total valid nights was 6446, of which 4335 (67.3%) were during weekdays and 2111 (32.7%) during weekends. The median number of nights analysed per patient was six (four during weekdays, two during weekends).

### *3.3. Topic modeling*

Sensor data from 66 healthy subjects and 66 matched patients with COPD were used to create the vocabulary of words necessary to model the data of a much larger cohort of patients. Our approach in selecting the letters and then the words composing the vocabulary benefits from a methodology that preserves the interpretability of the original signals and that allows the generation of words that actually do not occur in the current documents. The process consists of three steps and it is shown in Fig. 2. In the first step, the recorded 1-min epoch signals are encoded into artificial letters. In the second step, the letters are concatenated generating a bag of words representation of the original 1-min epoch signals. Lastly, LDA is applied on the corpus of documents obtained. The trained LDA model is applied to unseen data sequences without update or retraining. Each step is described in details in the following sections.

#### *3.3.1. Sequence encoding*

Let $D = \{\mathbf{S}_n\}^4_{n=1}$ be a multisensory data sequence where $S_1$, $S_2$, $S_3$, and $S_4$ are sequential measurements of MET, GSR, Temp and SC collected by the activity monitor during a single night of recording. Each $S_n = [s_1, s_2, \ldots, s_t]$ is of length t = 540 corresponding to the number of minutes assessed between 21:00 and 6:00. In the encoding step, we processed each sensor data sequence $S_n$ assigning a symbol to each 1- min data sample. METs data were encoded with the symbols *S*, *VL*, *L* and *MV* using the thresholds and intensity categories (*IC*) proposed by the American College of Sports Medicine [13]: very light intensity (*VL*), < 2.0 METs; light intensity (*L*), 2.0 to 2.9 METs; moderate-to-vigorous intensity (*MV*), ≥ 3.0 METs. Epochs marked by the activity monitor as sleeping and with METs < 2.0 formed a separate category named sleeping (*S*). Step counts were encoded in a binary form depending on whether the participant performed steps in each assessed minute (*SC* = 0: no steps performed, *SC* = 1: steps performed). Fig. 3a shows an example of data stream for a single patient night with METs (blue line) encoded with four ICs and step counts encoded with ON/OFF values (purple line). Temp and GSR data sequences were first cleaned from missing values and outliers (i.e. Temp values outside the range [24-40°C] and GSR values outside the range [0 − 8$\mu S$]) and then, for each subject, were centred across the mean over multiple assessed nights. In order to have sparse topics and symbols that best represent the original signal, it is desirable to have a discretization technique that produces letters with equal probabilities [35] and that minimizes the distortion of the partitioned signal [36]. For this reason the following procedure was used. Firstly, in order to conveniently describe a repertoire of activities in which a person may participate [32], we created eight measurement subspaces (four for Temp measurements and four for GSR measurements) by pooling together measurement samples collected in the same IC. In Fig. 3a, as an example, the Temp measurements collected while sleeping are highlighted in the rectangles. These data, together with the other Temp samples from other subjects collected while sleeping, form the sleeping subspace for the Temp measurements (Fig. 3b). In each subspace we estimated the empirical cumulative distribution function (ECDF) and we derived the three breakpoints ($a_1$, $a_2$, $a_3$) that divided the data into four equiprobable partitions. Separately, we also calculated the breakpoints ($b_1$, $b_2$, $b_3$) which divided the same set of data in four partitions minimizing the mean square distortion of the quantization [36]. The final partition breakpoints ($c_1$, $c_2$, $c_3$) were calculated averaging the corresponding pairs of breakpoints $c_i = \frac{a_i + b_i}{2}$ and used to divide the data into four contiguous, non-overlapping ranges of values. Final partition ranges ($P_1$, $P_2$, $P_3$, $P_4$) were sorted in ascending order such that the first range ($P_1$) represents the partition of data with the smallest values, and the last range ($P_4$) represents data with the highest values Fig. 3c. Each data sample can then be encoded with the symbol of a partition. In the example reported in Fig. 3 the first Temp sample highlited with a red circle is encoded with the symbol $P^{\text{Temp}}_2$ (Fig. 3d). The same process is applied for measurements in other ICs, as well as to convert GSR data.

**Table 1**
Subject group characteristics.

| | All COPD* | Matching healthy** | Matching COPD*** |
|---|---|---|---|
| | *n* = 1059 | *n* = 66 | *n* = 66 |
| Male/Female (*n*) | 689/370 | 30/36 | 30/36 |
| Age (years) | 66 [61–72] | 65 [61–70] | 65 [61–70] |
| BMI ($kg/m^2$) | 25.9 [22.5–29.6] | 25.2 [23–27.3] | 25 [22.5–27.8] |
| $FEV_1$ (% predicted) | 49 [34–64] | 107 [97–117] | 42 [29–63] |
| GOLD 1-2-3-4 (*n*) | 93-419-354-193 | – | 8-16-23-19 |
| MMRC 0-1-2-3-4 (*n*) | 145-279-228-195-67 | 44-1-0-1-0 | 6-15-18-15-5 |
| Assessed nights (*n*) | 6446 | 404 | 411 |
| Weekdays (%) | 67.3 | 69 | 65 |
| Weekends (%) | 32.7 | 31 | 35 |
| Nights per subject | 6 [6-6] | 6 [6-6] | 6 [6-6] |

Data are summarized as absolute frequency (*n*), relative frequency (%) or median and quartiles [Q1-Q3]. *MMRC data for 914 subjects. *MMRC data for 46 subjects. ***MMRC data for 59 subjects.

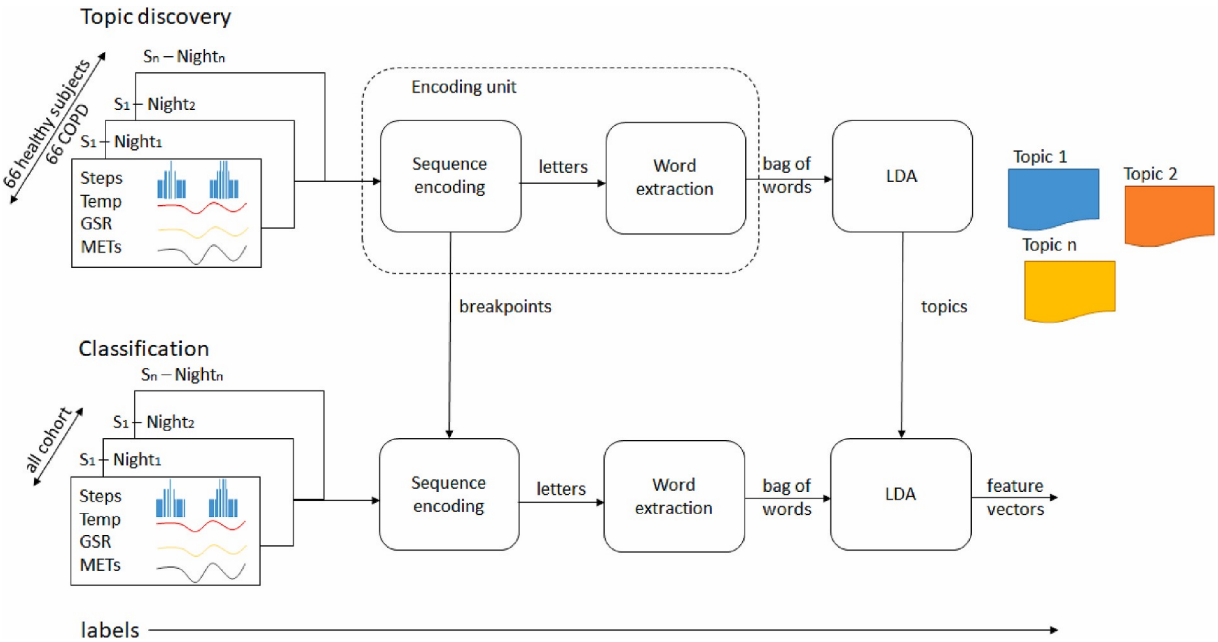


**Fig. 2.** The 1-min epochs signals enter into the encoding unit that first converts them into a discrete alphabet of letters and then combines them into artificial words composing text documents. LDA is applied to these documents to find the sleep topics.

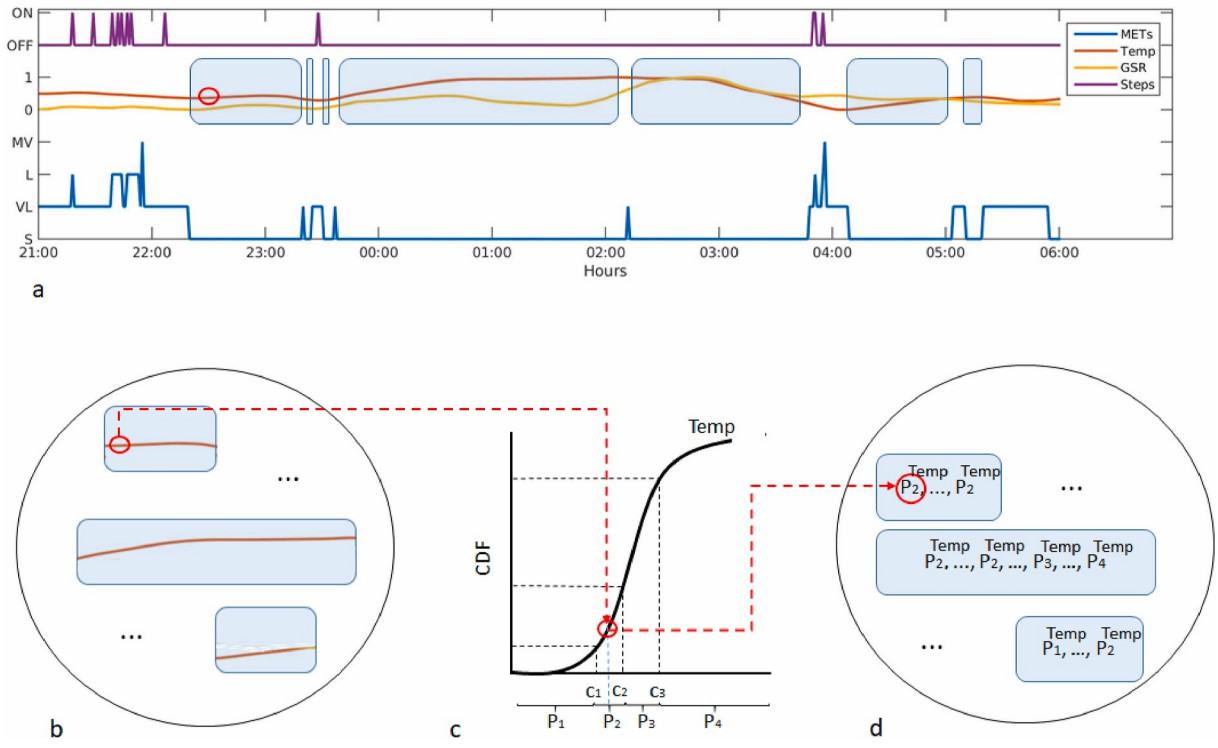


**Fig. 3.** a) Nighttime data from 21:00 to 06:00 of a participant in the study. Steps (in purple) are encoded in a binary form (ON = at least 1 step performed, OFF = no steps performed). GSR (in yellow) and temperature (in red) data are scaled in the interval 0–1 for visualization. METs data (in blue), defined as the energy cost of physical activities as a multiple of the resting metabolic rate, are encoded in four intensity category (S, VL, L, and MV). The red circle on the Temp signal indicates a 1-min measurement sample of the temperature signal. b) All the Temp data collected while sleeping are pooled together in order to create a measurement subspace in which c) three breakpoints are calculated and d) used to encode the data into partitions. Each 1-min sample, as the one in the red circle, is encoded with a symbol reflecting a partition, e.g. $P^{Temp}{}_2$. Concatenating all the symbols related to the same 1-min epoch a word is created. As an example the word [*S*, $P^{Temp}{}_2$, $P^{GSR}{}_1$, *SC* = 0] indicates a minute in which the subject was asleep (S), had a temperature value in the second range ($P^{Temp}{}_2$) and a galvanic skin response value in the lowest range ($P^{GSR}{}_1$). Obviously, the subject did not perform any step in that minute (*SC* = 0).

*3.3.2. Vocabulary*

The vocabulary of artificial words was built by allowing all the possible combinations of ICs, Temp and GSR partitions ranges, and binary values of steps.

For the sleeping category, for example, the 32 words $w^S_i$ of the vocabulary describing the sleeping intensity category can be represented by:

$$w_1^S : \left[S, P_1^{\text{Temp}}, P_1^{\text{GSR}}, \text{SC} = 0\right],$$

$$w_2^S : \left[S, P_1^{\text{Temp}}, P_1^{\text{GSR}}, \text{SC} = 1\right],$$

$$w_3^S : \left[S, P_1^{\text{Temp}}, P_2^{\text{GSR}}, \text{SC} = 0\right],$$

$$\vdots$$

$$w_{32}^S : \left[S, P_4^{\text{Temp}}, P_4^{\text{GSR}}, \text{SC} = 1\right].$$

A vocabulary of 128 words is created by iterating this process for each category level.

Next, we pruned the vocabulary adding wildcard characters to remove nonsense words and to increase the frequency of words related to the *VL* and *MV* intensity levels. This helps in finding sparse topics [26]. In particular, we used a wildcard character to replace the letters related to the situation in which steps are performed during sleeping *IC* (i.e. 16 words removed). Considering the neutral wildcard characters the original 32 words representing the sleeping category can be represented by 16 words as:

$$w_1^S : \left[S, P_1^{\text{Temp}}, P_1^{\text{GSR}}\right],$$

$$w_2^S : \left[S, P_1^{\text{Temp}}, P_2^{\text{GSR}}\right],$$

$$\vdots$$

$$w_{16}^S : \left[S, P_4^{\text{Temp}}, P_4^{\text{GSR}}\right].$$

Two wildcard characters replaced the temperature and galvanic skin response letters during *L* and *MV* intensities (i.e. 60 words removed) because values of these physiological parameters are affected by the high intensity of the activity performed. Considering the neutral wildcard characters the original 32 words representing the L category can be represented by two words as:

$$w_1^L : [L, SC = 0],$$

$$w_2^L : [L, SC = 1].$$

Similarly, the words representing the MV category are:

$$w_1^{MV} : [MV, SC = 0],$$

$$w_2^{MV} : [MV, SC = 1],$$

The VL category kept all the 32 original words. The number of letters per sensor modality describing each IC are shown in Table 2.

In view of sparsity we also weighted the informativeness of the remaining words in the vocabulary based on their inverse document frequency (IDF) score. Those words that have a high IDF are considered more informative, because they rarely occur in the collection. In particular, we set a threshold on the IDF score equal to the one removing the words occurring in at least 90% of the documents since, occurring so frequently, they are more likely to obscure than facilitate a meaningful decomposition of the collection of documents [17]. However, the only removed word by setting this threshold is [L, SC = 1]. The term frequency (TF), usually used in combination with IDF to form the TF-IDF score [27], was not considered since it would penalize words that rarely appear within a document such as words related to light or moderate to vigorous activity. Exactly these words are important for the identification of sleep topics correlated with the disease because they represent sleep fragmentation episodes with relatively high physical activity intensity. The IDF score of the words, and, subsequently, the set of removed words, are related to the wildcarding procedure previously described. If more letters are used, the words created will be more specific with the consequence of a higher IDF score average for the words in the vocabulary. On the other hand, a higher threshold on the IDF score (i.e. IDF equal to the one of the words present in 70% of documents) could cause the removal of all the terms useful to identify specific patients' subtypes.

*3.3.3. Topic discovery*

For topic discovery, we used the LDA implementation presented in Ref. [26], and we considered each night of assessment as a separate document. Following the process described in the previous sections the continuous, multivariate signals were mapped into a set of discrete symbols which can be handled by LDA. In particular each 4-elements vector *V* =[*IC,Temp,GSR,SC*], containing the raw 1-min epochs sensor measurements, was mapped with an instance of the vocabulary by associating the selected values in *V* with their partitions. Once that a term of the vocabulary was assigned to each minute, documents were created by constructing for each night a histogram of terms. We computed the results varying the number of desired topics from 3 to 20, and the hyperparameter α was set equal to 0.01 as in Ref. [21]. Hyperparameters are optimized with a variational expectation maximization algorithm initialized by randomly choosing a small number of "seed" documents [27]. We selected 18 seeds (nine from healthy subjects and nine from COPD patients). Topics did not change in their overall composition with different seed sets. Once the topics were calculated and the data of the all cohort of subjects encoded, we inferred the documents representing the nights of the subjects. The output of the algorithm is a vector that contains the coverage (or activation) of every topic for the document being modeled. We use these vectors as engineered features in three different classification problems. It is important to note that labels were never shown to the system up to the moment of training the classifiers.

**Table 2**
Number of words per sensor modality.

| Intensity Category | Sensor Modality | | | Total Number of words |
|---|---|---|---|---|
| | Temp | GSR | SC | |
| Sleeping (S) | 4 | 4 | 1 | 16 |
| Very Light (VL) | 4 | 4 | 2 | 32 |
| Light (L) | 1 | 1 | 2 | 2 |
| Moderate to Vigorous (MV) | 1 | 1 | 2 | 2 |

If the number is 1, it means this category equals a wildcard.

### *3.4. Classification*

To assess the discriminative power of the proposed feature extraction methodology we considered three classification problems: the classification of healthy and COPD subjects, the classification of COPD patients into four disease severity groups (GOLD 1, GOLD 2, GOLD 3, GOLD 4), and the estimation of patients' dyspnea score (MMRC 0, MMRC 1, MMRC 2, MMRC 3, MMRC 4). In each classification problem, the data set was divided into training (70% of the total data set) and test (30% of the total data set) sets. Grid-search was used on training data to choose the best set of model parameters using a 10-fold cross validation approach. Due to the repeated measures for each subject in the dataset (i.e. multiple assessed nights for each participant), the data for each participant was grouped within the folds to avoid introducing bias in the performance predictions that would have occurred if the nighttime data from a given participant were included in both the training and testing data. From each subject in the training set of each fold we randomly selected one night, represented by its characteristic vector of activation probabilities $\vartheta = [\vartheta_1,\ldots,\vartheta_k]$ over the topics $\theta_{1:K}$. We used these distributions to compute a square dissimilarity matrix $A$ between pairs of nights according to Kullback-Leibler divergence as in Ref. [37], in which $A(i,j)$ denotes the dissimilarity between the $i^{th}$ and $j^{th}$ randomly selected nights

$$A(i,j) = \sum_k \theta_k^i log \frac{\theta_k^i}{\theta_k^j}$$

with $\vartheta^i{}_k$ and $\vartheta^j{}_k$ the activation probabilities of the sleep topic $\theta_K$ for the nights represented by $\vartheta^i$ and $\vartheta^j$. The choice of the dissimilarity measure is critical and must fit the nature of the features in question, which in this case are discrete probability functions. Secondly, we calculated the eigenvectors and eigenvalues of A so that $AV = DV$, where D is a diagonal matrix of eigenvalues and V a matrix whose columns are the corresponding right eigenvectors and there are as many eigenvectors and eigenvalues as there are rows in the initial matrix. Eigenvalues were ranked from the greatest to the least. Using the transformation $V_T D_T^{-1}$, with $V_T$ the truncated eigenvector matrix of the first n eigenvectors of $V$ and $D_T$ the associated and truncated eigenvalue matrix, we summarized and attempted to represent inter-nights dissimilarities in a lower dimensional space. We iteratively projected into $V_T D_T^{-1}$ all the nights of each subject in the training set and then the ones of the subjects in the validation set such that the between-object dissimilarities are preserved as well as possible. In particular, given a vector of sleep topic activation probabilities θ representing one night, we calculated the vector $x$ of pairwise dissimilarities between θ and the nights used to compute $V_T D_T^{-1}$. Then we assigned to θ a location in a low-dimensional space projecting $x$ into the learned space $V_T D_T^{-1}$ according to $v = xV_T D_T^{-1}$.

Iterating this operation for each night, the positions of points relative to each other did not change but the coordinate systems changed resulting in a rotation of the data. In a nutshell we created a transformed feature set where rows represent a night of a patient and columns the projection of the pairwise dissimilarities into the space of the first n eigenvectors and eigenvalues learned using one single night per subject. We evaluated the performance for different combination of sleep topics $\theta_k$ (with $k$ varying from 3 to 20) and number of eigenvectors $n$ (with $n$ varying from 1 to 20). Lastly we fitted a new model on the whole training dataset with the parameters that yielded the best cross-validation performance in terms of average F1 score and presented the results obtained testing the model on the unseen test set. As classifier, we used the default MATLAB implementation of Random Forest (RF). RF hyperparameters were not optimized and a relative low number of trees (50 trees) was selected in view of the computationally intense cross validation process. We report the results for the classification of the subjects as results of a majority- vote mechanism in which the predicted class of each night represents a vote. For comparison we also evaluated the classification performance of models trained with conventional features based on duration and continuity of sleep such as total time spent sleeping, number of nocturnal sleeping bouts and duration of sleeping bouts.

## 4. Results

### *4.1. Healthy vs COPD*

A total of 571 nights from 93 subjects (46 healthy subjects and 47 COPD patients) were classified using cross validation. The average F1 score in classifying each night as healthy or COPD-type varying the number of eigenvectors and number of topic is shown in Fig. 4a. The mean F1 score for the single night classification over all the predictions was 0.87 (*SD*=0.01). The maximum F1 score (0.90, *SD* = 0.05) was achieved setting the number of topics to 13 and using the first 9 eigenvectors when projecting the topic dissimilarities in a lower dimensional space. Using these setting to train the final model we achieved a F1 score of 0.69 for the classification of the 20 healthy subjects and 19 COPD patients composing the test set. In addition, in order to mimic the COPD prevalence observed in real world, we conducted an additional experiment in which we created five imbalanced test groups. Four groups contained four patients with COPD and 20 healthy subjects and one group contained three patients with COPD and 20 healthy subjects. The mean F1-score over the five test sets was 0.77 (SD = 0.08). To compare the time spent in each of the 13 sleep topics in the 1059 COPD patients and the subgroup of 66 COPD patients used to extract the sleep topics, we constructed a linear mixed-effect model (LMM) for each sleep topic, with GOLD and MMRC as ordinal explanatory variables; subset group (i.e. all COPD vs matched COPD), smoking status, country of origin, gender and day group (i.e. weekday vs. weekend day) as categorical explanatory variables; age and BMI as continuous explanatory variables. Least Squares means (LS-means) and differences of LS- means of the fixed effects were used to compare the two subset groups. To account for repeated measurements, we used random effects on two levels. On the highest level, we included a random intercept per patient. The second level, within patients, had a random intercept for each day group (weekdays vs. weekends). The residuals then accounted for the differences between days within

the same day group. The model accounts for by-subject and by-day group variability. Degrees of freedom and p-values for significant differences (significant if $p < 0.05$) were computed using Satterthwaite's approximation [38]. To construct the models we used the *lmer* function of the package *lme*4 in *R* [39].

Comparison between the time spent in each of the 13 sleep topics in the 1059 COPD patients and the subgroup used to extract the sleep topics (Fig. 4b) shows that there are no statistical differences ($p > 0.1$ for all) in the time spent in each sleep topic between the two groups. This indicates that the sleep topics, created using a subset of patients, are able to generalize across many COPD patients.

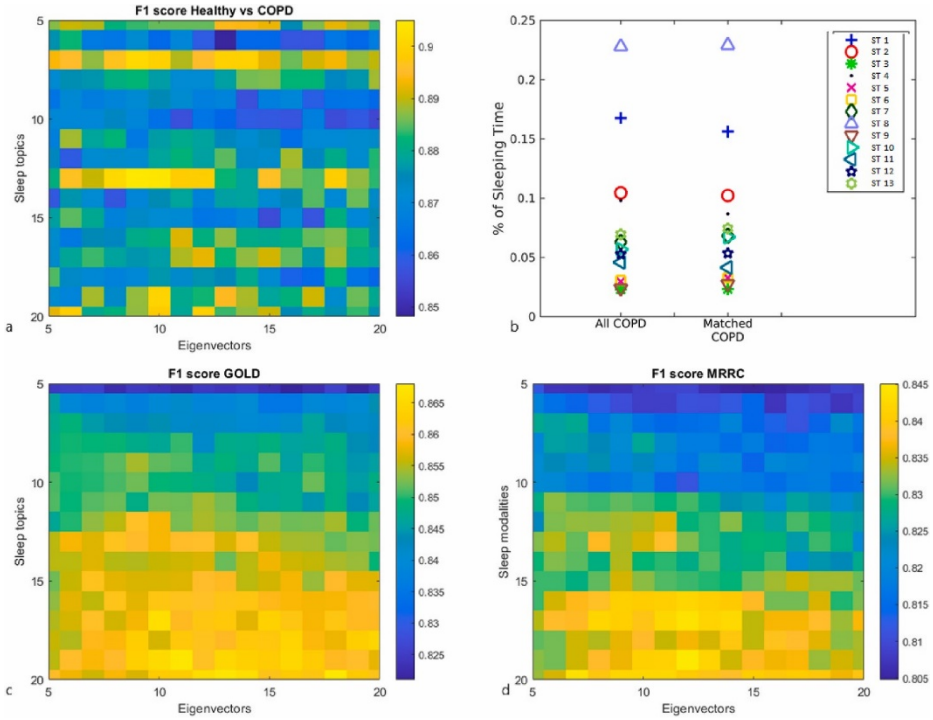


**Fig. 4.** Heatmaps visualizing the cross validation performance in classifying (a) nights of patients with COPD vs. nights of healthy subjects, (c) the patients' level of disease severity and (d) the patients' level of the dyspnea given each assessed night. Each colored square represents the average F1 score for a particular combination number of topics and eigenvalues used. Yellow squares represent higher F1 scores, while blue squares represent lower F1 scores. b) LS-means for time spent (as percentage of the nighttime assessed) in each sleep topic (ST) for all the COPD patients (n = 1059), and the COPD subset used to generate the sleep topics (n = 66).

### *4.2. Disease severity*

The average F1 score across all ten cross validation trials in classifying each night as belonging to one of the four COPD severity groups is shown in Fig. 4c. A total of 4509 nights were classified for a total of 742 subjects. The mean F1 score over the settings for the single night classification was 0.85 (*SD*=0.01). The maximum accuracy (0.87, *SD* = 0.02) was achieved setting the number of sleep topics to 20 and using the first 18 eigenvectors. For this setting, in the test set, we achieved a F1 score of 0.78 for the classification in the four GOLD grades of the 327 subjects in the test set.

### *4.3. MMRC score*

The average F1 score across all ten cross validation trials in classifying each night as belonging to one of the five MMRC classes is shown in Fig. 4d. A total of 3907 nights were classified for a total of 640 subjects. The mean F1 score over the settings for the single night classification was 0.82 (*SD*=0.01). The maximum F1 score (0.84, *SD* = 0.03) was achieved setting the number of sleep topics to 19 and using the first 12 eigenvectors. For this setting we achieved a F1 score of 0.79 for the prediction of the five MMRC scores of the 274 subjects in the test set.

### *4.4. Classification using conventional features*

For comparison, all the classification tasks were also executed using conventional sleep features such as total night sleeping time, number of nocturnal sleeping bouts and duration of sleeping bouts. The results on the test set are given in Table 3 together with the previously mentioned results obtained using the features derived from topic modeling. The mean F1-score over the five unbalanced test sets of COPD and healthy subjects as in 4.1 was 0.73 (SD = 0.09).

## 5. Discussion and conclusion

Early diagnosis of a disease is probably the most valuable asset in order to prevent damages or stall its progression by effective interventions. For some diseases, a reduction in misdiagnosis could typically be obtained by collecting longer observations during daily life which, however, should cause the smallest burden possible for the patients. For these reasons such observations should ideally be done by unobtrusive means preferably in a familiar environment. Although it is already possible to unobtrusively collect large amounts of data from a single person during sleeping hours, it is still difficult to merge the data into a set of features enabling valid decision support systems for patients suffering from COPD. The use of wearable sensors to assess sleep in COPD has been mainly limited to coarse-

grained methods such as actigraphy, for which limb motions are logged providing some insights in patient's sleep [40]. More recently, activity monitors have become popular to objectively assess the sleep–wake cycle providing minimally invasive measures of the duration, continuity and hence quality of sleep [28].

**Table 3**
Results of each classification task using conventional nighttime features and features derived from topic modeling.

| Classification task | F1 score | |
|---|---|---|
| | Conventional | Topic model |
| Healthy vs COPD | | |
| Healthy | 0.77 | 0.78 |
| COPD | 0.53 | 0.60 |
| **macro avg** | 0.65 | 0.69 |
| Disease severity | | |
| GOLD 1 | 0.44 | 0.74 |
| GOLD 2 | 0.72 | 0.82 |
| GOLD 3 | 0.65 | 0.84 |
| GOLD 4 | 0.44 | 0.73 |
| **macro avg** | 0.56 | 0.78 |
| MMRC score | | |
| MMRC 0 | 0.65 | 0.83 |
| MMRC 1 | 0.71 | 0.77 |
| MMRC 2 | 0.70 | 0.83 |
| MMRC 3 | 0.53 | 0.73 |
| MMRC 4 | 0.64 | 0.80 |
| **macro avg** | 0.65 | 0.79 |

In this paper we presented an approach that converts multimodal sleep data into a text representation and uses topic modeling to identify patterns across the dataset. This approach enables the discovery of higher level features resembling unique sleep characteristics that were then used both to discriminate between healthy subjects and those with COPD and to evaluate for each subject the severity of the disease and dyspnea level. The analysis included a real-world COPD patient cohort of more than 1000 patients and a subset of 66 healthy controls for a total of more than 6000 nights assessed. The conducted experiments have shown that it is possible to classify healthy subjects and subjects with COPD with a F1 score of 0.69. Conventional features such as total night sleeping time, number of nocturnal sleeping bouts and duration of sleeping bouts were able to differentiate between healthy subjects and patients with COPD with similar performances, but performed poorly in discriminating between different disease severity stages and dyspnea grades. Discovered latent structures in nighttime data, instead, were sufficiently sensitive to pick up differences existing between the four groups of COPD subjects (F1 score = 0.78) and five groups of dyspnoeic patients (F1 score = 0.79) with the advantage of being entirely data driven features and not relying on domain expert knowledge or on the detection of crucial sleep events. Since the extracted LDA based features are good predictors, (i.e., the Random Forest models performed well) the topic models would seem to capture important aspects of subjects sleeping behavior related to the effects of COPD and dyspnea. Based on the target outcome, the settings in the latent model should be adapted. In particular, a lower number of latent structures is required to get the best classification performance for a two class problem compared to the classification of more classes. For the four classes and five classes problems, a higher number of latent structures, in turn more specific, led to better classification results. Respectively, models with 13, 20 and 19 topics yielded the best F1 scores in 10 fold cross validation experiments.

While this paper focuses on the application of LDA as processing step to extract the requisite discriminative features for a conventional machine learning technique, it does not compare different methodologies. In particular, alternative advanced document representations for topic modeling able of expanding and enhancing the document representation in terms of syntactic and semantic information [41] should be evaluated in future research. Another powerful approach capable to solve complex classification problems worth exploring in the future is deep learning, that it has been proven to be successful with multi-sensor data [42] and offers the advantage of performing end-to-end training in which latent patterns can be automatically learned without the feature engineering process.

We believe our contribution represents a step towards a better support to the diagnosis of a complicated disease that will hopefully lead to a better patient care by early interventions. With the aim of learning different disease subspaces in mind, an open question for follow-up work is whether it is possible to use (fully or partially) known clinical relevant features provided by clinicians instead of hidden structures extracted from the data.

## Declaration of competing interest

On behalf of all the authors I declare that there are no conflicts of interest.


## Acknowledgment

This work was supported by the iCare4COPD project of Agentschap NL under contract number PNE101005. The authors gladly acknowledge Prof. Michael Polkey for supplying the data collected by his group. At the time of analysis Gabriele Spina was affiliated with the Data Science Department of Philips Research, Eindhoven.

## References


[1] P.M. Calverley, Copd: early detection and intervention, Chest 117 (5) (2000) 365S–371S.

[2] M. Miravitlles, A. Ribera, Understanding the impact of symptoms on the burden of copd, Respir. Res. 18 (1) (2017) 67.

[3] C.F. Vogelmeier, G.J. Criner, F.J. Martinez, A. Anzueto, P.J. Barnes, J. Bourbeau, B. R. Celli, R. Chen, M. Decramer, L.M. Fabbri, et al., Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. gold executive summary, Am. J. Respir. Crit. Care Med. 195 (5) (2017) 557–582.

[4] I. Iheanacho, S. Zhang, D. King, M. Rizzo, A.S. Ismaila, Economic burden of chronic obstructive pulmonary disease (copd): a systematic literature review, Int. J. Chronic Obstr. Pulm. Dis. 15 (2020) 439.

[5] J.B. Soriano, F. Polverino, B.G. Cosio, What is early copd and why is it important? Eur. Respir. J. 52 (6) (2018).

[6] B. Lamprecht, J.B. Soriano, M. Studnicka, B. Kaiser, L.E. Vanfleteren, L. Gnatiuc, P. Burney, M. Miravitlles, F. García-Rio, K. Akbari, et al., Determinants of underdiagnosis of copd in national and international surveys, Chest 148 (4) (2015) 971–985.

[7] D. Bellamy, J. Smith, Role of primary care in early diagnosis and effective management of copd, Int. J. Clin. Pract. 61 (8) (2007) 1380–1389.

[8] E. Mekov, M. Miravitlles, R. Petkov, Artificial intelligence and machine learning in respiratory medicine, Expet Rev. Respir. Med. 14 (6) (2020) 559–564.

[9] T.K. Liang, T. Tanaka, H. Nakamura, A. Ishizaka, A neural network based computer-aided diagnosis of emphysema using ct lung images, in: SICE Annual Conference 2007, IEEE, 2007, pp. 703–709.

[10] G. Gonzalez, S.Y. Ash, G. Vegas-S´ anchez-Ferrero, J. Onieva Onieva, F.N. Rahaghi, ´ J.C. Ross, A. Díaz, R. San Jose Est´ ´epar, G.R. Washko, Disease staging and prognosis in smokers using deep learning in chest computed tomography, Am. J. Respir. Crit. Care Med. 197 (2) (2018) 193–203.

[11] V. Cheplygina, L. Sørensen, D.M. Tax, J.H. Pedersen, M. Loog, M. de Bruijne, Classification of copd with multiple instance learning, in: 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 1508–1513.

[12] P. Arpaia, C. Manna, G. Montenero, G. D'Addio, In-time prognosis based on swarm intelligence for home-care monitoring: a case study on pulmonary disease, IEEE Sensor. J. 12 (3) (2011) 692–698.

[13] D. Newandee, S. Reisman, A. Bartels, R. De Meersman, Copd severity classification using principal component and cluster analysis on hrv parameters, in: 2003 IEEE 29th Annual Proceedings of Bioengineering Conference, IEEE, 2003, pp. 134–135.

[14] H. Ding, F. Fatehi, A. Maiorana, N. Bashi, W. Hu, I. Edwards, Digital health for copd care: the current state of play, J. Thorac. Dis. 11 (Suppl 17) (2019) S2210.

[15] B. Arnrich, O. Mayora, J. Bardram, G. Troster, Pervasive healthcare, Methods Inf. ¨ Med. 49 (1) (2010) 67–73.

[16] C.C. Bellos, A. Papadopoulos, R. Rosso, D.I. Fotiadis, "Identification of copd patients' health status using an intelligent system in the chronious wearable platform, IEEE journal of biomedical and health informatics 18 (3) (2013) 731–738.

[17] G. Spina, P. Casale, P.S. Albert, J. Alison, J. Garcia-Aymerich, R.W. Costello, N. A. Hernandes, A.J. van Gestel, J.D. Leuppi, R. Mesquita, et al., Identifying physical activity profiles in copd patients using topic models, IEEE journal of biomedical and health informatics 19 (5) (2015) 1567–1576.

[18] F. Mendonça, S.S. Mostafa, F. Morgado-Dias, A.G. Ravelo-Garcia, T. Penzel, A review of approaches for sleep quality analysis, Ieee Access 7 (2019) 24 527–624, 546.

[19] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, S. Taheri, Sleep quality prediction from wearable data using deep learning, JMIR mHealth and uHealth 4 (4) (2016) e125.

[20] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inf. Assoc. 25 (10) (2018) 1419–1428.

[21] T. Huynh, M. Fritz, B. Schiele, Discovery of activity patterns using topic models, in: Proceedings of the 10th International Conference on Ubiquitous Computing, 2008, pp. 10–19.

[22] N. Hu, G. Englebienne, Z. Lou, B. Krose, Learning latent structure for activity ¨ recognition,, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 1048–1053.

[23] L. Peng, L. Chen, X. Wu, H. Guo, G. Chen, Hierarchical complex activity representation and recognition using topic model and classifier level fusion, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (6) (2016) 1369–1379.

[24] K. Rieping, G. Englebienne, B. Krose, Behavior analysis of elderly using topic ¨ models, Pervasive Mob. Comput. 15 (2014) 181–199.

[25] E.J. McLaurin, J.D. Lee, A.D. McDonald, N. Aksan, J. Dawson, J. Tippin, M. Rizzo, Using topic modeling to develop multi-level descriptions of naturalistic driving data from drivers with and without sleep apnea, Transport. Res. F Traffic Psychol. Behav. 58 (2018) 25–38.

[26] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[27] D.M. Blei, J.D. Lafferty, Topic models," Text mining: classification, clustering, and applications 10 (71) (2009) 34.

[28] G. Spina, M.A. Spruit, J. Alison, R.P. Benzo, P.M. Calverley, C.F. Clarenbach, R. W. Costello, D. Donaire-Gonzalez, S. Dürr, J. Garcia-Aymerich, et al., Analysis of nocturnal actigraphic sleep measures in patients with copd and their association with daytime physical activity, Thorax 72 (8) (2017) 694–701.

[29] D. Andre, R. Pelletier, J. Farringdon, S. Safier, W. Talbott, R. Stone, N. Vyas, J. Trimble, D. Wolf, S. Vishnubhatla, et al., The Development of the Sensewear® Armband, a Revolutionary Energy Assessment Device to Assess Physical Activity and Lifestyle, " *BodyMedia Inc*, 2006.

[30] M.M. Sharif, A.S. BaHammam, "Sleep estimation using bodymedia's sensewear™ armband in patients with obstructive sleep apnea, Ann. Thorac. Med. 8 (1) (2013) 53.

[31] H. Watz, B. Waschki, T. Meyer, H. Magnussen, Physical activity in patients with copd, Eur. Respir. J. 33 (2) (2009) 262–272.

[32] M. Jett´e, K. Sidney, G. Blümchen, Metabolic equivalents (mets) in exercise testing, exercise prescription, and evaluation of functional capacity, Clin. Cardiol. 13 (8) (1990) 555–565.

[33] T. Troosters, F. Sciurba, S. Battaglia, D. Langer, S.R. Valluri, L. Martino, R. Benzo, D. Andre, I. Weisman, M. Decramer, Physical inactivity in patients with copd, a controlled multi-center pilot-study, Respir. Med. 104 (7) (2010) 1005–1011. [34] H. Al Otair, M. Al-shamiri, M. Bahobail, M.M. Sharif, A.S. BaHammam, "Assessment of sleep patterns, energy expenditure and circadian rhythms of skin temperature in patients with acute coronary syndrome," *Medical science monitor*, international medical journal of experimental and clinical research 17 (7) (2011) CR397.

[35] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003, pp. 2–11.

[36] S. Lloyd, Least squares quantization in pcm, IEEE Trans. Inf. Theor. 28 (2) (1982) 129–137.

[37] Y. Rubner, C. Tomasi, L.J. Guibas, "The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.

[38] J.H. Goodnight, Tests of hypotheses in fixed effects linear models, Commun. Stat. Theor. Methods 9 (2) (1980) 167–180.

[39] D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-7, ", 2014.

[40] D.M. Nunes, V.M. de Bruin, F.M. Louzada, C.A. Peixoto, A.G. Cavalcante, C. Castro- Silva, P.F. de Bruin, Actigraphic assessment of sleep in chronic obstructive pulmonary disease, Sleep Breath. 17 (1) (2013) 125–132.

[41] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, M. A. Gonçalves, Cluwords: exploiting semantic word clustering representation for enhanced topic modeling, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 753–761.

[42] J. Yang, M.N. Nguyen, P.P. San, X. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Ijcai, vol. 15, Buenos Aires, Argentina, 2015, pp. 3995–4001.