A probabilistic multi-aspect learning model for the early detection of COPD patients' symptoms

Şefki Kolozali
sefki.kolozali@essex.ac.uk
Institute for Analytics and Data Science, University of Essex, UK

Lia Chatzidiakou, Roderic Jones
ec571@cam.ac.uk,rlj1001@cam.ac.uk
Department of Chemistry, University of Cambridge, Cambridge, UK

Jennifer K. Quint
j.quint@imperial.ac.uk
Faculty of Medicine, National Heart & Lung Institute, Imperial College London, UK

Frank Kelly, Benjamin Barratt
frank.kelly@kcl.ac.uk,benjamin.barratt@kcl.ac.uk
MRC-PHE Centre for Environment & Health, King's College London, UK

ABSTRACT

Environmental pollutants can play a major factor on the increased cough and phlegm, asthma as well as chronic obstructive pulmonary disease (COPD). According to the British Lung foundation (BLF), lung disease in the UK costs the UK £11 billion a year and it is the 4th most costly disease are to the UK, after mental health conditions, musculoskeletal diseases and heart disease[1]. It costs the UK healthcare system between £810 and £930 million. COPD patients are at risk of sudden and acute worsening of symptoms. Their symptoms reduce the quality of their lives and sometimes they can leads to hospitalisation. Although there is a possibility that environmental pollutants can play a significant role in the triggering process of the symptoms, there has not been a detailed study to identify whether or not we can use the personal exposure to detect COPD patients' daily symptoms.

In this study, we present a cohort study with 106 COPD patients using mobile sensors, peak flow meter, and daily symptom records to monitor patients' activity and personal exposure to investigate the possibility of symptom detection based on personal air pollution exposure and daily peak-flow meter measurements. Our aim is to build a system that can detect COPD patients' symptoms and notify them as early as possible to avoid their personal exposure to the specific compound of gases and severe symptom experiences. One of the challenges in symptom detection is the multi-aspect learning and prediction process. We developed a system that can detect COPD patients' co-existing symptoms as well as supports joint tracking of symptoms over time. In order to achieve this, we applied the Probabilistic Latent Component Analysis (PLCA) model, which is based on a 3-dimensional dictionary of spectral templates/tensors. The model is combined with Linear Dynamic Systems (LDS), which is used to track the patients' symptoms. We present our results and examine the advantages of PLCA-LDS over other traditional multi-class – multi-label techniques, such as Random Forest.

KEYWORDS

Internet of Things, Probabilistic Models, Air Pollution, Personal Exposure, Respiratory illnesses, COPD, Interpretable Machine Learning

ACM Reference Format:
Şefki Kolozali, Lia Chatzidiakou, Roderic Jones, Jennifer K. Quint, and Frank Kelly, Benjamin Barratt. 2019. A probabilistic multi-aspect learning model for the early detection of COPD patients' symptoms. In *KDD '19: A KDD Workshop on Tensor Methods for Emerging Data Science Challenges, August 4 - 8, 2019, Anchorage, Alaska, USA.* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

Environmental pollutants have been an important factor in economy and health for the UK. It has been significantly supported by many studies that air pollutants can play a major factor on increased cough and phlegm, asthma as well as chronic obstructive pulmonary disease (COPD)[2]. For instance, it has been estimated that COPD costs the UK healthcare system between £810 and £930 million. Moreover, it has been estimated that it results in loss of 24 million working days annually. There are many different risk factors that can trigger an exacerbation of symptoms. While smoking is the most important risk factor for COPD, an estimated 25-45% of patients have never smoked [1]. Other risk factors include a history of pulmonary tuberculosis, chronic asthma, childhood respiratory tract infections, occupational exposure to dusts and gases, air pollution and low socioeconomic status. The prevalence of COPD is increased among individuals who live close to traffic [2]. The COPD patients have significant mortality risks associated with particles and temperature changes [3].

[1] https://www.blf.org.uk/what-we-do/our-research/economic-burden

[2] https://www.gov.uk/government/collections/comeap-reports

Obtaining comprehensive information about the patients' daily symptoms and exposure can aid doctors in the accuracy of the diagnosis process. Moreover, it can assist patients to alter their behaviour to prevent worsening of their symptoms. Having greater evidence-based information can also help policymakers to make better decisions to minimise citizens' risk, improving both their health and quality of life. According to the medical literature, daily self-reported symptoms correlate with patients' deterioration. Having a personal digital health system that encapsulates the recording of daily symptoms, personal exposure and activities of patients' can help provide a solution for effective self-monitoring of symptoms and vital signs. Moreover, it could prevent or decrease the hospital admissions caused by exacerbations using such systems to notify patients as early as possible. Such a system needs to learn and adapt to each patient's symptoms and dynamic environment while taking into account the overall pattern of the symptom. Personalised algorithms that take into account personal exposure, patients' living environment, and self-reported symptoms could help to identify the onset of vital signs and notify patients and doctors before exacerbation of patients' symptoms.

Our aim is to build a system that can detect COPD patients' symptoms as early as possible and notify them. To understand patients' personal exposure and well-being, we use miniature sensory devices, which can monitor their personal air pollution exposure and physical activities, and symptom diary cards collected daily, and peak flow measurements recorded daily by patients'. Figure 1 demonstrates the devices used in the measurement of environmental pollutants and lung capacity of patients. It is challenging to detect symptoms from noisy multimodal and multi-channel signals captured in dynamic environments. First, we need a rich set of features to characterise and detect symptoms in sensory environments. Second, traditional multi-class and multi-label pattern recognition techniques cannot model and predict multiple different sets of unknown variables as well as their relationships at the same time in a continuous manner. In this paper, we aim to tackle both of these aspects by using a rich set of predictors in the analysis and applying a dynamic probabilistic model that allows us to formulate our problem in a flexible way to learn multiple different aspects at the same time.

2 BACKGROUND

Feature extraction

Feature extraction is a crucial step in data analysis, particularly in audio analysis. To date, most of the studies focused on magnitude-based features (i.e. average) to detect worsening of symptoms. However, such basic features are not adequate to capture patients' daily exposure, activities, worsening of symptoms. For instance, a patient sometimes may have a constant low-level peak flow measurement or can be affected by acute change in their environment, which may not be detected by only using magnitude-based features, such as average. This is often the case with patients, who suffer from chronic breathlessness. While in the former case energy spectral density can be more useful, the latter can be better characterised with spectral flux and kurtosis type of features to detect and track the worsening of their symptoms. In our model, we used a variety of statistical and spectral features, such as spectral entropy, kurtosis, spectral flux, spectral energy, skewness, variance, mean, median, minimum, maximum.

Figure 1: The Personal Air Monitor (PAM) being carried on the left, peak flow meter (top-right) and spirometer instruments (bottom-right) (without case).

Classification

The classification problems concerning high-dimensional categorical predictors have become popular in a variety of application areas. For instance, epidemiology studies commonly focus on relating a particular medical disease or genetic information that involves substantial amounts data to various environmental or health factors. The interactions between the factors and the significant amount of data play a crucial role in such big data analysis. However, there is a lack of data analysis methods that can cope with the representation and analysis of such high-dimensional data and obtain a meaningful outcome. The traditional approach is to use popular methods (e.g. Linear Regression or Support Vector Machines) along with feature selection techniques to allow simultaneous selection of correlated sets of features. L1 penalisation [4] and the elastic net [5] that combines L1 and L2 penalties are among the most popular approaches in this regard. On the one hand, insertion of high dimensional data into a logistic model triggers a massive dimensionality problem. On the other hand, utilisation of only selected features can limit experiments to low-order interactions and modest

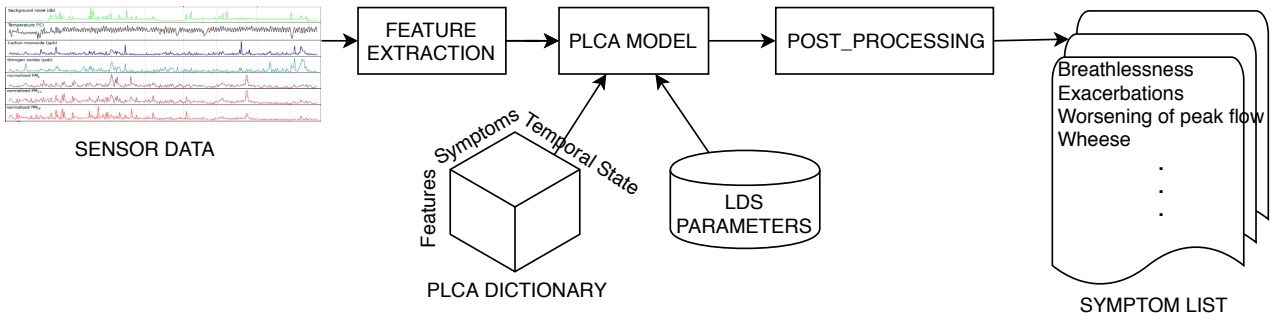


Figure 2: The proposed framework for the detection of COPD patients' co-existing symptoms.

numbers of features. It is possible to see similar problems with other feature selection techniques such as classification and regression trees (CART) [6] and random forests (RFs) [7]. While the former tends to be computationally unstable and leads to low classification performance, the latter provides uninterpretable model despite reducing variance and achieving high accuracy.

A comprehensive comparison has been made between Bayesian Tensor Factorisation and various feature selection techniques and traditional classifiers (i.e. RFs, Lasso, SVM) in [8]. The results showed that Tensor Factorisation techniques have outperformed the traditional techniques when the dimensionality of datasets increased [8]. Another popular approach that can help to learn from multiple datasets at the same time is multi-aspect learning techniques, which are based on coupled matrix and tensor factorisation techniques [9–11].

However, while tensor factorisation techniques can be used to estimate the latent clusters, they are more suitable for relational learning problems, learning subspaces, clustering, and data fusion [12]. The obtained eigen tensors need to be fed into a classifier to carry out a classification task as in [13, 14]. There has been a few studies where hybrid tensor factorisation techniques developed for supervised classification tasks [15, 16]. The shortcoming of these studies is either they cannot deal with multi-label classification or time series prediction problems. While a time series prediction method is proposed in [17], which can smooth and regularise across time, it cannot achieve multi-label classification. The nature of Internet of Things Healthcare and technology supported Epidemiological studies often manifest several different sets of unknowns in dynamic environments. For instance, we often interested in not only a phenomenon/symptom, but also whether or not it is the onset of a symptom, or the seasonality of the symptom. This would normally require either training three different classification models and then predicting outcomes separately or using multi-class and multi-label classifiers, which do not provide information about the probabilistic interaction between each set of labels.

In this study, we used PLCA method [18, 19] that can predict multiple sets of outcomes in high dimensional tensor format at the same time. The problem with tensor and matrix factorisation techniques is that they do not model time-axis and produce continuous output. Therefore, it is possible to get fragmented output rather than a smooth predictive outcome. As a solution, we applied a hybrid approach, that is proposed in [20]. The probabilistic approach combines the PLCA model with Markov chain of latent variables or State Space Models, where each observation conditioned on the state of the corresponding latent variable.

3 PROPOSED METHODOLOGY

Motivation and System Overview

The overall aim of the proposed framework is the creation of a system for the detection of COPD patients' co-existing symptoms which can also support joint tracking of symptoms over time. Figure 2 illustrates the proposed framework. In this paper, we aim to express a symptom as a linear combination of symptoms and a collection of temporal states, onsets, transients and offsets. Thus, the personalised PLCA models are based on a 3-dimensional dictionary of spectral tensors: symptom type, temporal state of the symptoms, and frequency. It should be noted that the proposed PLCA-based model is expressed as a mixture of latent components corresponding to symptoms. Thus, it cannot jointly model multiple concurrent symptoms. The model can, however, infer the presence of concurrent symptoms by calculating the posterior probability of each symptom over all possible symptoms.

Unlike HMMs, which support only one-dimensional discrete latent variable, Linear Dynamic Systems (LDS) support a multidimensional and continuous latent variable space. The integration of PLCA with LDS enabled us to jointly

track multiple coexisting symptoms over time using LDS. The proposed system takes multi-channel sensory observations as input. The specific observations include Nitric Oxide (NO), Carbon Monoxide (CO), Particulate Matter 1 (PM1), Particulate Matter 2.5 (PM2.5), Particulate Matter 10 (PM10), Relative Humidity (RH), background noise, triaxial accelerometer, temperature, spatial coordinates (GPS). Moreover, it also takes daily peak flow measurement of each participant in one of our experiments. The model uses a pre-extracted dictionary of spectral templates in the form of tensors. Non-negative Matrix Factorisation (NMF) approach used in the creation of the dictionary templates. Symptom tracking using LDS can take place within the PLCA inference or can take place as a post-processing step. We used the LDS within the PLCA model. The model output is finally converted into a list of symptoms identified along with the temporal states, such as onset, transient, and offset. Then, a thresholding approach is applied on the output, where we compute sets of metrics, F-measure, for all possible permutations and use the best threshold values for each symptom in the testing phase.

Preprocessing

The window size is a crucial aspect of time series analysis. It is often difficult to determine the window size in a high dimensional space. Autocorrelation is one of the techniques that are used in the determination of lags. The lags determines how often an incident may occur. In our experiments, we computed lags based on peak flow measurements of patients in order to determine the window size. We have individually calculated autocorrelations and lags of patients and then calculated the average and median of the outputs to determine the window size for entire cohort. We have used the small lag as window size both for environmental and health data streams (e.g. 5 days). In order to decompose the environmental data into lower resolution components, we have used Discrete Wavelet Transform (DWT), particularly Daubechies wavelets.

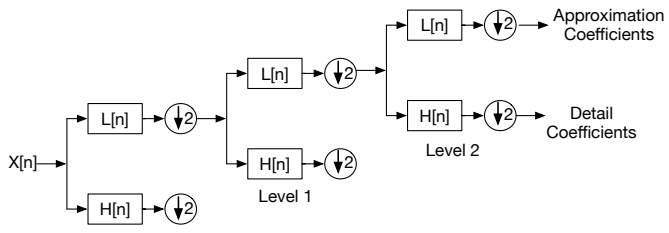


Figure 3: Wavelet decomposition over 2 levels. L[n] is the low-pass filter and H[n] is the high-pass filter. The lower resolution output involves approximation and detail coefficients.

Figure 3 illustrates a decomposition process of DWT, using low pass and high pass filters to calculate the approximation and detail coefficients. Once we have obtained our lower resolution data, we have extracted magnitude, acute and cumulative spectral and statistical features, such as Spectral Flux, Spectral Centroid, Spectral Energy, Average, Median, Kurtosis, Variance, Skewness. The same features have also been extracted for peak flow measurements. Our goal is to detect daily symptoms of the patients given the features extracted from the past few days (i.e. fixed window size). We used a sliding window in the feature extraction phase and the window size was 5 days.

PLCA

The applied PLCA method takes multi-channel sensor readings, $V_{f,t}$ and approximates it as a bivariate probability distribution $P(f,t)$. The model decomposes the approximated spectral features $P(f,t)$ into a dictionary of spectral templates per symptom s, and temporal state of symptom a, as well as probability distributions for symptom activations. The model is formulated as:

$$P(f|t) = P(t) \sum_{s,a} P(f|s,a)P(s|t)P(a|s,t) \tag{1}$$

where s $\in \{1, ..., S\}$ denotes the symptom class, a $\in \{1, ..., A\}$ denotes the temporal state of symptom. $P(t)$ is defined as $\sum_f V_{f,t}$, which is a known quantity. It corresponds to the sum of all spectral features for each time frame t. Each t corresponds to spectral features calculated for the last 5 days. By applying NMF on the extracted features, we obtained a 3-dimensional tensor dictionary $P(f|a,s)$ that contains the spectral features for each patient's symptom s, temporal state a. $P(s|t)$ is the time-varying symptom activation. $P(a|s,t)$ represents the symptom state activation per symptom s, across time t.

To be regarded as probabilities, spectral features $P(f|s,a)$ are normalised with respect to f as to sum to one. Moreover, $P(s|t)$, and $P(a|s,t)$ are similarly normalised with respect to s and a, respectively. On the contrary, $P(f|t)$ and $P(t)$ are not normalised since they carry information on the energy of the spectral features. Nonetheless, since $P(t)$ and $P(f,t)$ are cancelled out through the partition functions, this doesn't affect inference. The unknown model parameters, $P(s|t)$ and $P(a|s,t)$, were estimated using iterative update rules, Expectation-Maximisation (EM) algorithm. For the E-step, the following posterior is computed:

$$P(s,a|f,t) = \frac{P(f|s,a)P(s|t)P(a|s,t)}{\sum_{s,a} P(f|s,a)P(s|t)P(a|s,t)} \tag{2}$$

For the $M-step$, $P(s|t)$, $P(p|s,t)$, $P(a|s,t)$ are updated using the posterior of Eq 2:

$$P(s|t) = \frac{(\sum_{a,f} P(s,a|f,t)V_{f,t})^{\kappa}}{\sum_{s}(\sum_{a,f} P(s,a|f,t)V_{f,t})^{\kappa}} \quad (3)$$

$$P(a|s,t) = \frac{(\sum_{f} P(s,a|f,t)V_{f,t})^{\lambda}}{\sum_{a}(\sum_{f} P(s,a,|f,t)V_{f,t})^{\lambda}} \quad (4)$$

In order to lower the entropy in $P(s|t)$ and $P(a|s,t)$ and to promote sparsity, we set κ and $\lambda > 1$ (i.e. typical values are between 1.1-1.5). We set κ to 1.4 and λ to 1.2. Due to the fact that $P(f|s,a)$ was pre-extracted and considered as fixed variable, we have not used any update rule for the symptom feature templates. We initialised the unknown parameters $P(s|t)$ and $P(a|s,t)$ in the EM updates with random values between 0 and 1. We iterated Eq. 3 and Eq. 4 until convergence. In our experiments, we found 40 iterations to be sufficient. The obtained output of the PLCA model is a 2-dimensional non-binary representation of symptom activations over time, given by $P(s,t) = P(t)P(s|t)$ with dimensions of $S \times T$. Essentially, the output created by calculating the posterior probability of each symptom over all possible symptoms (i.e. $P(s=1|t), P(s=2|t), ..., P(s=S|t)$) weighted by energy of the spectral features.

Linear Dynamic Systems

Linear Dynamic Systems (LDS) is a special case of State Space Models (SSM), where the latent and observed variables are multivariate Gaussian distributions, and their means are linear functions of their parent states. LDS estimates the state $x \in \Re^n$ of a discrete-time controlled process that is governed by the linear stochastic difference equation given below:

$$X_k = Ax_{k-1} + Bu_k + w_{k-1} \quad (5)$$

where x_{k-1} is the hidden state, A represents the transition model, B is the driving force, u is the l-dimensional driving force and w is the process noise. In addition, the variable k represents the time step in the tracking process. The observation that is made are m-dimensional with a measurement $z \in \Re^m$ that is:

$$z_k = Hx_k + v_k \quad (6)$$

where z_k is the observation, H is the observation model. The random variables w_k the process noise and v_k represent the measurement noise. They are assumed to be independent of each other with normal probability distribution as given in the following equation:

$$p(w) \sim N(0, Q), \quad (7)$$

$$p(v) \sim N(0, \mathcal{R}) \quad (8)$$

where Q is the process noise covariance and $\mathcal{R}$ is measurement noise covariance which change at each time step.

The PLCA model output $P(s,t)$ contains the non-binary activation of overlapping symptoms s over time t. However the model of Eq. 1 does not support any temporal constraints. Thus, it can lead to temporally fragmented output. Here, we used LDS to perform symptom tracking. In the computation, we assume that symptom activation $P(s,t)$ is a 'noisy' observation x_t in an LDS and latent states z_t correspond to our desired output. By initialising $P(s|t)$ in the EM updates with a binary mask that corresponds to the ground truth annotations, the resulting output only has nonzero activations in the time instants and classes corresponding to ground truth symptoms. We used standard binary transition matrix, A, and binary observation model, H, in our experiments. We set Q and R as 0.01 and 0.01, respectively.

4 EVALUATION

Description of dataset

Overall, 106 participants were recruited over a period of two years. Figure 4 shows the patient monitoring coverage during our study. Out of 106 participants, only 50 participants carried our sensors for minimum 60% of the monitoring time. The number of participants can decrease depending on the preferred personal coverage. For instance, only 40 participants carried our sensors if we seek to have 70% coverage for patients.

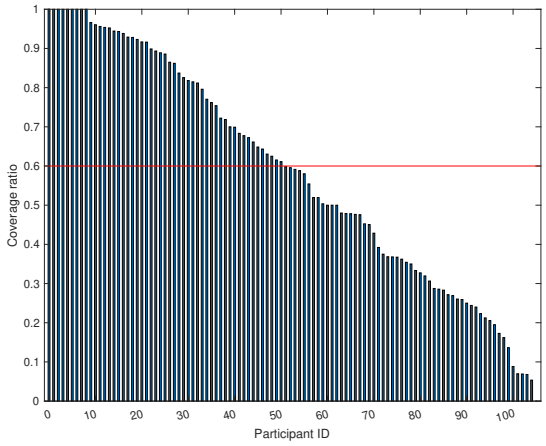


Figure 4: The ratio of personal exposure coverage. The red line indicates the 60% (50 participants) exposure coverage limit.

Figure 5(a) depicts the number of symptom onsets and Figure 5(b) depicts the number of symptom transients (i.e. healing period) for participants who carried our sensors for more than 60% of the time during our study. The number of symptom onsets showed a large variance depending on the symptom. The highest variation was seen for worsening of peak flow measurements, on average 10 onsets, where as the

numbers were much lower for the rest of the symptoms. It is worth to point out that only exacerbations and peak flow measurements reflects the biomarker measurements since the rest of them are based on the patients' personal diary symptom records. While the exacerbations of symptoms were decided by a GP, peak flow measurements shows a quantitative measurement of patients' daily lung capacity.

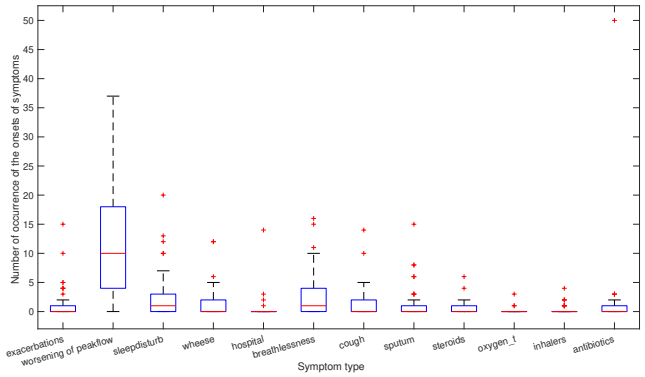
Most of the symptoms have fairly low number of onset occurrence, such as being hospitalised, using inhalers, streroids, experiencing sputum, taking oxygen. However, this is not the case for peak flow measurements where the number of occurrence and variance in worsening of peak flow measurements is very high. On the contrary, it can be seen in Figure 5(b) that there was high numbers of symptom transients in the cohort study. In other words, although the patients may not experience their symptoms very often, once they experienced it, it may take a long time to recover from their symptoms. This poses a particular data analysis challenge for us. It means that we will have fewer symptom onsets in our analysis and our dataset is mostly dominated by transients.

Training and Testing

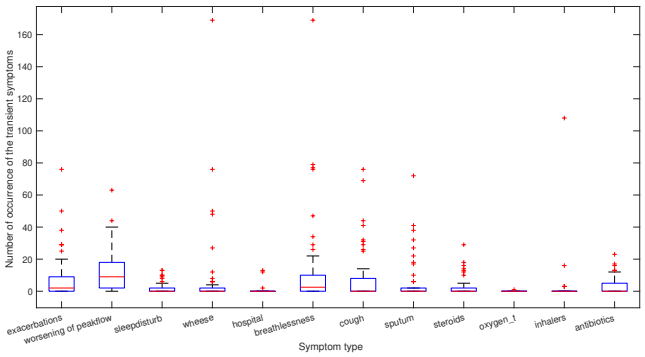
In our experiments, we only used the participants who carried our sensors for more than 60% of the observation period. Then, we split the dataset into the first and second half of each month and trained our model on the first half of each month and aimed to predict the symptoms occurred in the second half of each month per participant. The aim was to obtain a personalised symptom prediction. We conducted two different experiments. In the first experiment (i.e. referred to as Exp. 1), we trained our model with all sensor observations. This includes NO, CO, PM1, PM 2.5, PM 10, Relative Humidity, triaxial accelerometer, temperature, audio from microphones, GPS data, and peak flow measurements of participants. In the second experiment, which is referred to as Exp. 2, we used only air pollutants (i.e. NO, CO, PM1, PM 2.5, PM 10, Relative Humidity) to train our model. In order to have a balanced training and testing, we used only symptoms that have minimum one complete symptom (i.e. a symptom that involves all three temporal states, namely onset, transient, offset) during the first half and second half of each month. Although using such an approach dramatically reduced our training and testing data set, we believe that it allowed us to capture different temporal states of symptom occurrences in a better and more detailed way. We compared our model against Random Forest classifier. It is a multi-class and multi-label classifier and proven to be a powerful pattern recognition technique [21].

Results

Table 1 and Table 2 show the results obtained in our experiments. Table 1 shows results for each symptom and its



(a) Number of symptom onsets for the participants with minimum 60% exposure coverage ratio.



(b) Number of symptom transients for the participants with minimum 60% exposure coverage ratio.

Figure 5: Figure 5(a) and Figure 5(b) show the number of occurrence of the symptom onsets and transients, respectively.

temporal state. The results show that PLCA performed better in the prediction of sputum, inhalers and antibiotics usage, sleep disturbance of COPD patients, whereas the Random Forest model performed better in the prediction of breathlessness, wheese, coughs, and exacerbations of COPD patients' symptoms. However, it is worth to point out that the Random Forest model is more susceptible to having a higher variance in the prediction rate compare to the PLCA model. It seems like some of the symptoms were better captured and detected when we used all sensor observations. This can be explained by the fact that patients' daily physical activities can be better characterised when we use accelerometer, temperature, microphone, distance, and GPS sensor observations in addition to air pollutants, i.e. NO, CO, PM1, PM10, PM2.5, RH.

Regarding the exacerbations and worsening of peak flow, while the Random Forest model performed better when we used all sensory observations as input in experiment 1, the PLCA model performed better than Random Forest when

Symptom	Experiment No.	Temporal State	Random Forest			PLCA		
			Precision	Recall	F_1 score	Precision	Recall	F_1 score
Breathlessness	1	Symptom	0.27 (0.12)	0.29 (0.17)	**0.22 (0.09)**	0.17 (0.05)	0.27 (0.09)	0.18 (0.04)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.06 (0.01)	0.09 (0.02)	**0.05 (0.01)**
		Transient	0.11 (0.06)	0.15 (0.10)	**0.12 (0.07)**	0.18 (0.06)	0.05 (0.00)	0.05 (0.00)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.07 (0.02)	0.15 (0.09)	**0.08 (0.02)**
	2	Symptom	0.27 (0.12)	0.28 (0.15)	**0.21 (0.09)**	0.15 (0.04)	0.28 (0.12)	0.15 (0.04)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.06 (0.01)	0.10 (0.03)	**0.06 (0.01)**
		Transient	0.10 (0.05)	0.13 (0.08)	0.11 (0.06)	0.26 (0.08)	0.13 (0.02)	**0.14 (0.02)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.01)	0.04 (0.01)	**0.04 (0.01)**
Wheese	1	Symptom	0.26 (0.11)	0.31 (0.16)	**0.28 (0.13)**	0.30 (0.07)	0.23 (0.04)	0.26 (0.05)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.10 (0.04)	0.03 (0.00)	**0.05 (0.01)**
		Transient	0.34 (0.19)	0.21 (0.11)	**0.22 (0.09)**	0.36 (0.12)	0.06 (0.00)	0.10 (0.01)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.17 (0.01)	0.30 (0.13)	**0.20 (0.03)**
	2	Symptom	0.34 (0.09)	0.33 (0.14)	**0.31 (0.11)**	0.21 (0.03)	0.30 (0.11)	0.20 (0.02)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.00)	0.02 (0.00)	**0.03 (0.00)**
		Transient	0.33 (0.18)	0.15 (0.08)	**0.16 (0.07)**	0.18 (0.08)	0.04 (0.00)	0.06 (0.01)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14 (0.04)	0.11 (0.02)	**0.12 (0.03)**
Sputum	1	Symptom	0.39 (0.18)	0.14 (0.03)	0.20 (0.05)	0.37 (0.03)	0.27 (0.02)	**0.29 (0.02)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.00)	0.01 (0.00)	**0.01 (0.00)**
		Transient	0.18 (0.12)	0.04 (0.01)	0.06 (0.01)	0.23 (0.11)	0.08 (0.01)	**0.07 (0.00)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.06 (0.01)	**0.05 (0.00)**
	2	Symptom	0.28 (0.12)	0.15 (0.04)	0.19 (0.06)	0.45 (0.07)	0.29 (0.02)	**0.34 (0.03)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		Transient	0.18 (0.12)	0.04 (0.01)	0.05 (0.01)	0.13 (0.06)	0.04 (0.00)	**0.05 (0.01)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.06 (0.01)	0.20 (0.12)	**0.09 (0.02)**
Worsening of Peak Flow	1	Symptom	0.49 (0.11)	0.40 (0.12)	**0.39 (0.09)**	0.29 (0.03)	0.48 (0.11)	0.33 (0.04)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.02)	0.18 (0.04)	**0.12 (0.01)**
		Transient	0.22 (0.14)	0.13 (0.07)	0.15 (0.08)	0.26 (0.04)	0.22 (0.04)	**0.21 (0.03)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.01)	0.14 (0.02)	**0.10 (0.01)**
	2	Symptom	0.45 (0.09)	0.30 (0.09)	0.30 (0.06)	0.31 (0.07)	0.44 (0.14)	**0.31 (0.06)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.01)	0.15 (0.02)	**0.12 (0.01)**
		Transient	0.17 (0.11)	0.07 (0.03)	0.09 (0.04)	0.24 (0.05)	0.16 (0.02)	**0.17 (0.02)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.13 (0.02)	0.15 (0.02)	**0.12 (0.01)**
Inhalers	1	Symptom	0.00 (-)	0.00 (-)	0.00 (-)	0.25 (-)	1.00 (-)	**0.40 (-)**
		Onset	0.00 (-)	0.00 (-)	0.00 (-)	0.08 (-)	0.50 (-)	**0.14 (-)**
		Transient	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)
		Offset	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)
	2	Symptom	0.00 (-)	0.00 (-)	0.00 (-)	0.19 (-)	1.00 (-)	**0.32 (-)**
		Onset	0.00 (-)	0.00 (-)	0.00 (-)	0.08 (-)	0.50 (-)	**0.14 (-)**
		Transient	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)
		Offset	0.00 (-)	0.00 (-)	0.00 (-)	0.09 (-)	1.00 (-)	**0.17 (-)**
Cough	1	Symptom	0.41 (0.18)	0.30 (0.13)	**0.32 (0.12)**	0.22 (0.07)	0.30 (0.10)	0.21 (0.06)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.08)	0.18 (0.14)	**0.11 (0.08)**
		Transient	0.24 (0.13)	0.21 (0.13)	**0.20 (0.10)**	0.20 (0.05)	0.12 (0.02)	0.12 (0.01)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.25 (0.14)	**0.08 (0.01)**
	2	Symptom	0.46 (0.15)	0.30 (0.14)	**0.31 (0.11)**	0.13 (0.06)	0.16 (0.06)	0.13 (0.05)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.14 (0.08)	**0.06 (0.01)**
		Transient	0.19 (0.09)	0.21 (0.14)	**0.19 (0.10)**	0.33 (0.12)	0.13 (0.02)	0.16 (0.03)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.00)	0.05 (0.02)	**0.02 (0.00)**
Antibiotics	1	Symptom	0.12 (0.05)	0.13 (0.05)	0.12 (0.05)	0.16 (0.06)	0.25 (0.16)	**0.19 (0.08)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.00)	0.03 (0.01)	**0.02 (0.00)**
		Transient	0.14 (0.11)	0.16 (0.11)	0.15 (0.11)	0.24 (0.09)	0.13 (0.03)	**0.16 (0.03)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	2	Symptom	0.08 (0.02)	0.12 (0.04)	0.10 (0.03)	0.29 (0.04)	0.29 (0.10)	**0.25 (0.04)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.01)	0.06 (0.03)	**0.05 (0.01)**
		Transient	0.10 (0.05)	0.18 (0.12)	**0.13 (0.07)**	0.10 (0.03)	0.07 (0.02)	0.07 (0.02)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Sleep Disturbances	1	Symptom	0.16 (0.09)	0.12 (0.07)	0.13 (0.08)	0.42 (0.02)	0.55 (0.08)	**0.42 (0.02)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.03)	0.10 (0.02)	**0.08 (0.02)**
		Transient	0.12 (0.09)	0.05 (0.01)	0.07 (0.03)	0.10 (0.03)	0.07 (0.01)	**0.08 (0.02)**
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.16 (0.03)	0.21 (0.02)	**0.15 (0.01)**
	2	Symptom	0.15 (0.08)	0.12 (0.06)	0.14 (0.07)	0.14 (0.05)	0.14 (0.07)	**0.14 (0.06)**
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.22 (0.03)	0.30 (0.04)	**0.21 (0.01)**
		Transient	0.10 (0.07)	0.05 (0.01)	**0.07 (0.03)**	0.05 (0.01)	0.03 (0.00)	0.03 (0.00)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.11 (0.02)	0.13 (0.02)	**0.09 (0.01)**
Exacerbations	1	Symptom	0.42 (0.12)	0.25 (0.05)	**0.31 (0.06)**	0.07 (0.01)	0.19 (0.10)	0.10 (0.03)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.00)	0.12 (0.10)	**0.03 (0.00)**
		Transient	0.35 (0.18)	0.18 (0.05)	**0.22 (0.07)**	0.27 (0.06)	0.06 (0.00)	0.09 (0.01)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.01)	0.04 (0.01)	**0.04 (0.01)**
	2	Symptom	0.36 (0.13)	0.21 (0.06)	**0.26 (0.08)**	0.18 (0.03)	0.33 (0.10)	0.19 (0.02)
		Onset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.00)	0.03 (0.01)	**0.02 (0.00)**
		Transient	0.34 (0.15)	0.15 (0.05)	**0.18 (0.07)**	0.24 (0.12)	0.06 (0.01)	0.09 (0.03)
		Offset	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.00)	0.05 (0.01)	**0.04 (0.01)**

Table 1: Overall average and variance of the predictions made for the Random Forest and PLCA models are trained and tested by using sensory observations obtained from patients with minimum 60% coverage of their daily lives and personal exposure. The values outside the parentheses are average and inside the parentheses are variances of the predictions.

	Random Forest			PLCA		
Exp. No.	Precision	Recall	F_1 score	Precision	Recall	F_1 score
1	0.14 (0.09)	0.10 (0.06)	0.10 (0.05)	0.16 (0.04)	0.19 (0.07)	**0.14 (0.03)**
2	0.12 (0.07)	0.08 (0.04)	0.08 (0.04)	0.14 (0.04)	0.17 (0.06)	**0.13 (0.03)**

Table 2: Overall summary of the results provided in Table 1 for the Random Forest and PLCA models. It illustrates the average and variance of the predictions made for all symptoms and symptom states in experiment one and two.

we only used air pollutants in experiment 2. When we look at the results for the detection of exacerbations of patients' symptoms and worsening of peak flow in experiment 2, the Random Forest model performed 0.26 on average with 0.08 variance, whereas the PLCA model performed 0.19 on average with only 0.02 variance. Considering the fact that there is high variation in the prediction rate of the Random Forest model, it wouldn't be wrong to interpret that both of these prediction rates are fairly similar to each other. In other words, the PLCA model is more robust, but it performs worse on average regarding the exacerbations of symptoms in experiment 1. One particularly interesting result is worsening of peak flow measurements of patients. Although the difference is very small (i.e. RF on average 0.30 prediction rate with 0.06 variance and PLCA on average 0.31 prediction rate with 0.06 variance), the PLCA model performed better on the prediction of change of the worsening of peak flow, compared to the Random Forest model in experiment 2. Both of these cases require further investigation, and it will be more clear once we conducted our experiments at the population level.

Table 2 illustrates the overall performance for each classification model. From the overall results, it is clear that the results are not very high in the personalised symptom detection. Nonetheless, it is possible to observe that PLCA model outperformed Random Forest in the prediction of symptoms. Moreover, it is interesting to see that there is not only a higher average performance, but there is also a smaller variance in the overall performance of PLCA. This again shows that Random Forest is more susceptible to changes in the prediction of symptoms among COPD patients. The small difference between experiment 1 (i.e. with all sensor observations) and experiment 2 (i.e. with only air pollution observations) shows that air pollutants are as much effective as using entire sensory observations. While the overall results are not always accurate, it is possible to see that air pollutants are associated and can be used in the prediction of COPD patients' symptoms at a reasonable level. However, there is a few challenges in the identification of patterns linking personal exposure to health effects. First, it is very challenging to completely capture personal exposure of patients'. Second, apart from worsening of peak flow and exacerbation symptoms, the majority of the symptoms were subjective symptoms reported by participants. Future work should focus on exploring how best to obtain continuous biomarker data as well as using wearable devices to better capture the daily activities of COPD patients'.

5 CONCLUSIONS

In this paper, we presented our results on the prediction of the COPD patients' daily symptoms by using sensory observations. In general, reasonable results were obtained for both of the classifiers. Although the predictions are not always accurate, it is evident that the utilisation of only air pollutants is as much effective as using rich sensory observations. It should be noted that it is difficult to monitor personal exposure of patients as they volunteered in our study, and we have no control over their daily lives. In the majority of the cases, patients' did not carry the sensory devices for their reasons. There may also be some other possible factors affected our results. For instance, we are aware of the fact that the personalised predictions are highly likely to be affected by overtraining, i.e. over-fitting, since they are trained with a small size of data set of per participant. Nonetheless, it is possible that once we conducted our population experiments, the outcomes may have an opposite effect, they might be undertrained. Future work will include analysis of the symptoms of participants', who had less than 60% coverage. It will help us to observe whether or not retained patients are substantially different to those omitted. Our future experiments will also use many different configurations at both personal and population level. Moreover, we will investigate utilisation of different window size in our experiments.

ACKNOWLEDGMENTS

The COPE: Characterisation of COPD Exacerbations using Environmental Exposure Modelling is funded by Medical Research Council, MR/L019744/1.

REFERENCES

[1] Nicholas S. Hopkinson and Michael Polkey. Chronic obstructive pulmonary disease in non-smokers. *The Lancet*, 374(9706):1964, 2009.

[2] Yusheng Qiu, Jie Zhu, Venkata Bandi, Robert L. Atmar, Keith Hattotuwa, Kay K. Guntupalli, and Peter K. Jeffery. Biopsy neutrophilia, neutrophil chemokine and receptor gene expression in severe exacerbations of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 168(8), 2003.

[3] Guixiang Song, Guohai Chen, Lili Jiang, Yunhui Zhang, Naiqing Zhao, Bingheng Chen, and Haidong Kan. Diurnal temperature range as a novel risk factor for copd death. *Respirology*, 13(7):1066–1069, 2008.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society*, pages 273–282, 1996.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society*, pages 301–320, 2005.

[6] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984.

[7] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] Yun Yang and DB Dunson. Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association*, 1459:656–669, 2013.

[9] Evrim Acar, Anders J. Lawaetz, Morten A. Rasmussen, and Rasmus Bro. Structure-revealing data fusion model with applications in metabolomics. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2(1):6023–6026, 2013.

[10] Evrim Acar, Rasmus Bro, and Age K. Smilde. Data fusion in metabolomics using coupled matrix and tensor factorizations. *Proceedings of the IEEE*, 103(9):1602–1620, 2015.

[11] Evangelos E. Papalexakis, Tom M. Mitchell, Nicholas D. Sidiropoulos, Christos Faloutsos, Partha Pratim Talukdar, and Brian Murphy. Turbo-smt : Parallel coupled sparse matrix-tensor factorizations and applications turbo-smt : Parallel coupled sparse matrix-tensor factorizations and applications. *Statistical Analysis and Data Mining*, 9:269–290, 2016.

[12] Tamara G. Kolda and David Hong. Stochastic gradients for large-scale tensor decomposition. *arXiv.org*, pages 1–30, 2019.

[13] Koji Maruhashi, Masaru Todoriki, Takuya Ohwa, Keisuke Goto, Yu Hasegawa, Hiroya Inakoshi, and Hirokazu Anai. Learning multi-way relations via tensor decomposition with neural networks. *The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3770–3777, 2018.

[14] Evrim Acar, Gözde Gürdeniz, Francesco Savorani, Louise Hansen, Anja Olsen, Anne Tjønneland, Lars Ove Dragsted, and Rasmus Bro. Forecasting chronic diseases using data fusion. *Journal of Proteome Research*, 16(7):2435–2444, 2017.

[15] Fei Wu, Xu Tan, Yi Yang, Dacheng Tao, Siliang Tang, and Yueting Zhuang. Supervised nonnegative tensor factorization with maximum-margin constraint. *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 962–968, 2013.

[16] Emmanouil Benetos, Constantine Kotropoulos, and Senior Member. Non-negative tensor factorization applied to music genre classification. *IEEE Trans. on Audio, Speech and Language Processing*, 18(8), 2010.

[17] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. *Proceedings of NIPS*, (Nips):847–855, 2016.

[18] Paris Smaragdis and B Raj. Shift-invariant probabilistic latent component analysis. *Journal of Machine Learning Research*, (5), 2007.

[19] Emmanouil Benetos and Tillman Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. *16th International Society for Music Information Retrieval Conference*, pages 701 – 707, 2015.

[20] Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Polyphonic sound event tracking using linear dynamical systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–12, 2017.

[21] Manuel Fern and Eva Cernadas. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.