Preprint UVSQ; June 14, 2021

Tell Me What Air You Breath, I Tell You Where You Are

Hafsa El Hafyani
DAVID Lab
UVSQ - Université Paris-Saclay
Versailles, France
hafsa.el-hafyani@uvsq.fr

Mohammad Abboud
DAVID Lab
UVSQ - Université Paris-Saclay
Versailles, France
mohammad.abboud@uvsq.fr

Jingwei Zuo
DAVID Lab
UVSQ - Université Paris-Saclay
Versailles, France
jingwei.zuo@uvsq.fr

Karine Zeitouni
DAVID Lab
UVSQ - Université Paris-Saclay
Versailles, France
karine.zeitouni@uvsq.fr

Yehia Taher
DAVID Lab
UVSQ - Université Paris-Saclay
Versailles, France
yehia.taher@uvsq.fr

ABSTRACT

Wide spread use of sensors and mobile devices along with the new paradigm of Mobile Crowd-Sensing (MCS), allows monitoring air pollution in urban areas. Several measurements are collected, such as Particulate Matters, Nitrogen dioxide, and others. Mining the context of MCS data in such domains is a key factor for identifying the individuals' exposure to air pollution, but it is challenging due to the lack or the weakness of predictors. We have previously developed a multi-view learning approach which learns the context solely from the sensor measurements. In this demonstration, we propose a visualization tool (**COMIC**) showing the different recognized contexts using an improved version of our algorithm. We also demonstrate the change points detected by a multi-dimensional CPD model. We leverage real data from a MCS campaign, and compare different methods.

KEYWORDS

Activity Recognition, Multivariate Time Series Classification, Multi-view Learning, Mobile Crowd Sensing, Air Quality Monitoring

1 INTRODUCTION

Air quality and exposure to pollution is a central concern for people living in urban areas. As the harmful effects of air pollutants on their health is alarming. The key concern to reduce the risk of these pollutants on individual's health is by understanding the totality of exposure. Air pollution monitoring is getting more interest nowadays, due to the rapid advances of the Internet of things (IoT) along with the emergence of the Mobile Crowd Sensing (MCS) paradigm.

The mentioned technologies coupled with the widespread use of GPS, allows volunteers to contribute their collected data in order to get personalized insights about their exposures to pollution. Polluscope [1] is a French project deployed in Île-de-France (i.e., Paris region), and is a typical use case study based on MCS. In Polluscope, participants are equipped with a sensor kit which can measure different pollutants such as Nitrogen dioxide (NO2), Particulate Matters (PMS), Black Carbon (BC), Temperature, etc… independently form their environment either indoor or outdoor.

Air quality strongly depends on the context of the participant, thus in order to understand and identify participants' exposure to pollution, it is essential to identify the context of the participants. To avoid miss-classification of the exposure wrt the context (micro-environment), the participants need to fill a time-use diary, but in real-life, they rarely thoroughly do this self-reporting task. Therefore, mining the context of participants based on the data collected from the crowd is an attractive solution. But, it is challenging due to the imperfection of the data as shown in [8]. Consequently, learning from individual sensor data fails to identify the micro-environment. This classification might be improved by combining multiple sensor data. However, this is not straightforward due to the complexity and interdependence of such multi-dimensional time series [3].

This task is more or less related to human activity recognition, where there exists a long established research, and a wide range of applications. Different types of data are used ranging from GPS only, to one or many accelerometers, sound or combinations. An attempt to work with environmental data was introduced by Asimina et al. [2] where the authors explore the capability of predicting location from sensors data using an Artificial Neural Network (ANN) model. The authors use low-cost individual sensors and GPS for data collection, and provide a location predictive model based on Artificial Neural Network (ANN) to derive the time-space activity daily profile of the individuals. Although their approach predicts very well indoor locations, however, there is still room for improvement in discriminating between in transit and outdoor locations.

Furthermore, there is a need for a full-fledged implementation that can be used in real-world applications to fill the gap between data collection and context recognition using environmental data. In our case, the input is both GPS and environmental sensor data. We formulate the task of recognition of micro-environment as a multivariate time series classification (MTSC) [4]. In a previous work [1], we've proposed a multi-view stacking generalization approach in order to detect the micro-environments from the air pollution data collected based on different learners.

In this demonstration, we include time series segmentation based on change point detection to emphasize the automatic detection of the change in the user's context. Furthermore, we develop a visualization tool **COMIC** (**C**ontext **O**f **M**ob**I**le **C**rowdsensing) showing the different recognized contexts and illustrating the importance of multi-view approach when compared to single view approach and other baseline learners. This visualization interface highlights also the importance of our approach vis-a-vis users' declared contexts.

[1] http://polluscope.uvsq.fr

1.1 System Overview

We propose a context recognition approach based on multi view learning which takes environmental data collected from different sensors and GPS and predicts the user's context.

Our approach is mainly based on our previous work on micro-environment recognition [1] where we propose a holistic approach that promotes the idea of combining air quality plus mobility dimensions to recognize users' whereabouts (i.e., micro-environments). In our previous work, we evaluated different approaches and provided a framework dedicated to the preparation, the application and the comparison of different machine learning algorithms for micro-environment recognition as shown in Figure 1. We proposed a multi-view learning approach based on *2NN-DTW* for the first-level learners and *Random Forest* (*RF* for short) as a meta-learner. Our previous proposed approach was compared to state-of-the-art techniques namely MLSTM-FCN [7] model and kNN with Dynamic Time Warping (DTW) distance metric model which was considered for a long time the state-of-the-art in the time series classification problem [4].

In this work, we propose an improved and more robust multi-view system that we call **COMIC**, for micro-environment recognition in the context of environmental crowdsensing which combines time series segmentation based on change point detection [6] and micro-environment recognition [1]. COMIC consists of a Graphic User Interface (GUI) and four different layers: (1) The declared micro-environments by the participants, (2) the environmental data measurements over time, (3) the detected change points, and (4) the detected context.

- **Interface.** The GUI of COMIC is a front-end interactive interface which has three functionalities. It provides a list of all the participants. The user is invited to choose one participant's data then the interface provides the plotted measurements with the corresponding user annotations (i.e., declared). The user is also allowed to choose the first-level and meta learners among a predefined learning models list. COMIC back-end then applies the classification for this participant's data and shows the predicted context for each single view and for the multi-view.
- **Declared micro-environments.** This information, which is also called self-reporting, characterizes the declared micro-environment by participants. This information is not always accurate as participants do not bother to fill thoroughly this information. Our GUI permits to show to what extent this information is accurate or not compared to our detected contexts and vice versa.
- **Data measurements.** Data measurements consist of the collected data by the participants. These data is used as input for the multi view learning model to detect automatically the context of the participants.
- **Change points.** Change points consist of the exact timestamps of the change in participants' micro-environments. Based on participants measurements, our GUI calls on the back-end the detected change points based on multidimensional change point detection [6], and displays the results.
- **Micro-environment recognition.** Our micro-environment recognition model is based on multi view learning modelling. The user is allowed to choose which model to use for the first-level and the meta learners. The GUI displays then the detected micro-environment for the selected participant.

First-Level Learners						Associated Prediction Probabilities						True Label
l_1	l_2	...	l_i	...	l_n	p_1	p_2	...	p_i	...	p_n	y

Table 1: An example of the new generated dataset D'.

2 MULTI-VIEW LEARNING

Our multi-view model is based on the idea of multi-view stacking from [5] and we adapted it to best fit our needs. As we suggest in [1], to learn the micro-environment of participants from multi-variate time series, we propose a two-stage model based on multi-view learning. Figure 2 shows our multi-view learning architecture. It consists of two steps multivariate time series classification while preserving the properties of each source of data. A first-level learner is used for each view independently, then the results of the first-level classification are used to generate a D' dataset having feature vector which contains the prediction and its corresponding probability for each class as shown in Table 1 to train the meta-learner.

Assuming that Y_{it} is a dimension of the n-dimenstional time series $Y_t = (Y_{1t}, Y_{2t}, ..., Y_{it}, ..., Y_{nt})$, the first-level learner takes as input the time series data coming from each view V_i, where V_i represents a dimension Y_{it} of the multi-variate time series Y_t and $V = (V_1, V_2, ..., V_i, ..., Vn)$ is the set of views. Then, each view will generate its own predicted labels with associated prediction probabilities with the form $[l_i, p_1, p_2, ..., p_j, ..., p_k, y]$, where l_i is the predicted label of the first-level learner i, p_j is the associated prediction probability for each class j of the k possible classes, and y is the true label.

The main advantages of using this approach is separating first-level learners from meta-learner, so we can use any desired learners as first-level and meta learners. Moreover, we are preserving the properties of each data source as we are not aggregating data from different sources.

COMIC leverages the idea of our previous work and improves it. Instead of being restricted to *kNN* as first level learner and *RF* as second level learner, we carried several extensive experiments trying different classifiers (including SVM) and added a new multi-view model with *RF* as first-level learner and meta learner since this shows a solid robustness compared to either our previous model or state-of-the-art models [5]. Plus, even in the case of missing dimensions, COMIC maintains its robustness and detects the context with the existing dimensions. Furthermore, we improved the performance of our model by integrating some a priori rules, such as the unlikelihood of being in some micro environment at some time of day (e.g., being in office at 3 am), or of transitions between some micro-environments (e.g., going from store to home without passing by a movement segment).

We have implemented the Multivariate Long Short Term Memory with Fully Connected Network model (MLSTM-FCN) as well as KNN-DTW classifier for the aggregated data (all views are aggregated together) which are considered as state-of-the-art by the time series classification community [4].

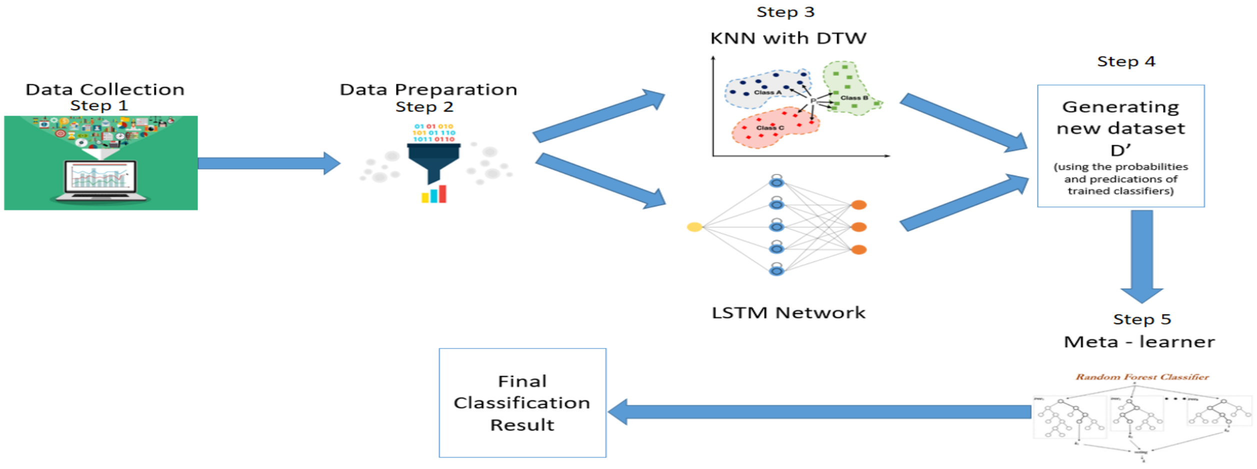


Figure 1: Overview of the Micro-Environment Recognition Process.

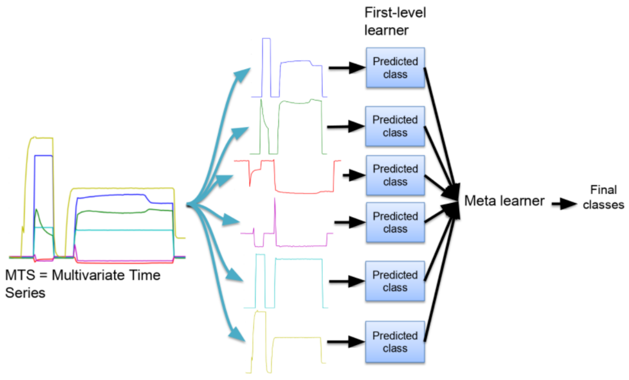


Figure 2: Multi-view Learning Approach

Compared to our previous work, these changes resulted in a significant improvement of the global accuracy (0.95 against 0.83 that we obtained in [1]). The new model also beats MLSTM-FCN which global accuracy is 0.70, and the accuracies of the individual classes are lower in general.

Additionally, COMIC includes another component which consists of segmenting the multivariate time series into coherent segments, each segment represents a micro-environment by resorting to multivariate time series change point detection. Our change point detection approach consists of applying the CUSUM algorithm as first-level learner on each dimension separately. Each dimension generates a set of detected change points. The output is then fed to a second-level learner to learn the weights of every dimension in proportion to the performance of individual learners using a gold set of annotated data as ground truth [6].

3 DEMONSTRATION SCENARIO

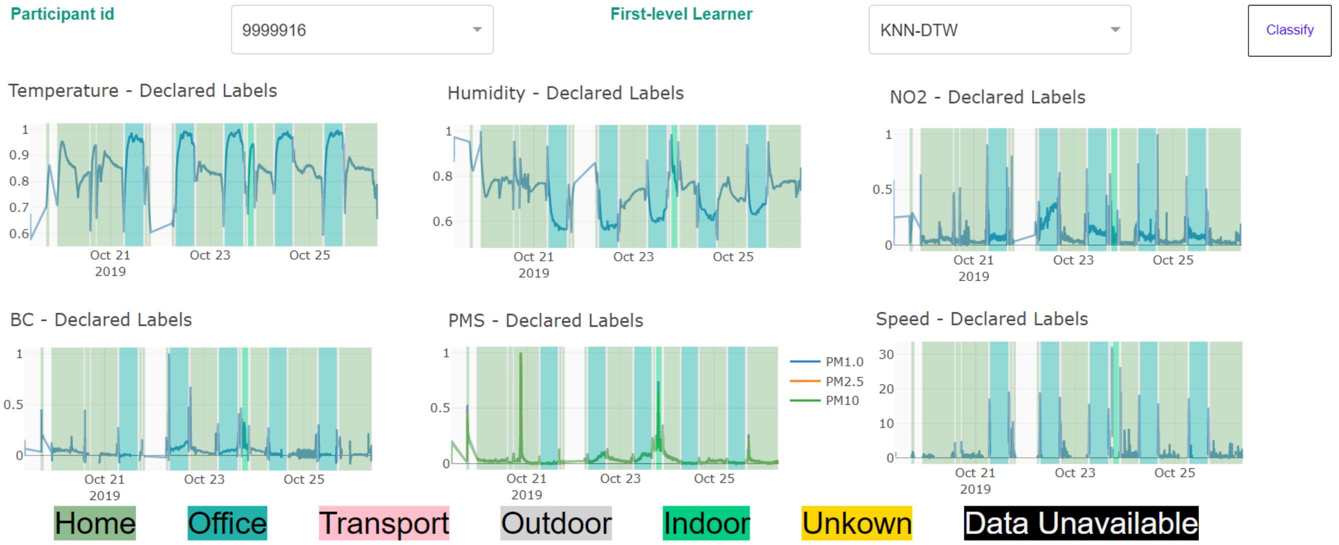
In this section, we illustrates the user interaction with COMIC interface. The experiments are carried out on different environments. The multi-view learning model was implemented in Python 3.6 using scikit-learn 0.23.2 and tslearn [9]. The deep-learning model MLSTM-FCN [7] was trained using Keras 2.2.4. Our GUI (graphical user interface) was implemented using python 3.6, Plotly [2], and Dash [3] framework. A real-world environmental data collected in the context of Polluscope project is used as a benchmark of environmental crowdsensing data. In this context, participant collect air quality measurements (NO2, PM1.0, PM2.5, PM10, BC, Temperature, Humidity) plus GPS locations which are used to derive participants speed. The recruited participants where given a mobile app in order to annotate their micro-environments whenever it changes. Micro-environments are grouped into five categories: Home, Office, Indoor, Outdoor and Transport. Indoor spaces incorporate all closed spaces except home and office, such as restaurants, stores and stations, while outdoor spaces, as its name indicates, consist of open spaces such as park and street. Users can load our interface and start enjoying its appealing functionalities. We emphasize three main scenarios of COMIC:

Data visualization:

User can visualize the collected data during the campaign period. Each dimension is plotted in a separate graph. Along with each plot, the corresponding declared activity at that time is shown. Thus users can easily see visually how much the changes in participant's context and the changes in data are correlated. Figure 3 shows the different dimensions plots with the corresponding declared activities. Users can choose the participant id from the drop down in order to navigate through the data of different participants.

Classification and CPD over all dimensions: Users are able to perform classification and change point detection algorithm over the

[2] https://plotly.com/
[3] https://dash.com/



Participant id: 9999916

First-level Learner: KNN-DTW

Figure 3: COMIC visualization GUI.

data of a specified participant. The users need only to specify the participant id, then they can choose the learner in the first-level learner (i.e., KNN-DTW or Random Forest), and by clicking on the classify button, classification and change point detection (CPD) will be applied on each view.

As shown in Figure 4a, each dimension is plotted with its declared micro-environments versus the ones detected by the first level learner on this dimension. Moreover, three plots appear showing the aggregated view using the KNN-DTW, the MLSTM-FCN algorithm, and another one showing the results of the multi-view learner as shown in Figure 2. We can also notice that in the absence of some dimensions, KNN-DT and MLSTM-FCN will fail to detect the micro-environment, while the multi-view learner keeps detecting the micro-environments. Furthermore, the detected changes by the CPD algorithm are plotted also as red dash lines. This interface allows users to see to what extent the results of COMIC are accurate vis-a-vis the declared micro-environments and vice versa, by comparing the declared micro-environments in Figure 4a (e.g., BC - Declared labels) and the predicted micro-environments in Figure 2 (i.e., Multi-view - Predictions).

Classification and CPD over a specified dimension: Another functionality of COMIC allows users to focus on one dimension and plot the classification result of the first-level learner, which in most the cases is not accurate. Users only need to specify the participant id, the dimension, and the first level learner from the drop-down lists, then they are invited to click on the classify button. The output of this functionality shows three plots: (1) a plot of the specified dimension with the declared micro-environment, (2) another plot showing the results of the chosen first level learner on the specified dimension, and (3) a third plot showing the results of the multi-view learner. All the plots include the Change points detected as a vertical dashed red line.

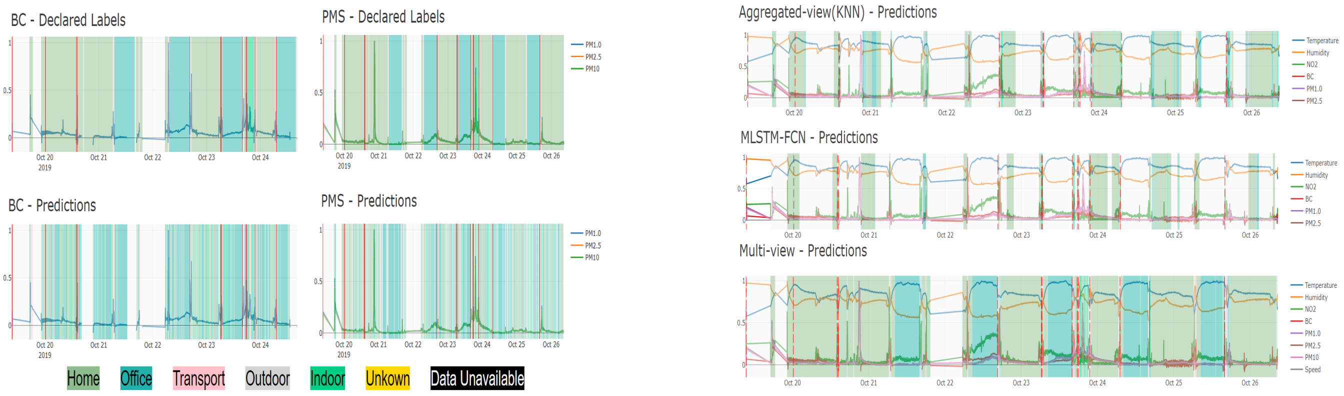
Figure 5 shows the comparison between the declared micro-environments, the ones predicted from the single dimension in question, and the ones detected by the COMIC model. On the one hand, the first graph indicates that the participant stayed outdoor for two consecutive days, meaning that this participant did not thoroughly annotated their data. Yet, our multi-view model (third plot) can detect successfully the micro-environment of this participant during this two days. On the other hand, the second plot shows the results of the first level learner. While this learner fails to detect all the micro-environments correctly, we recognize that the multi-view approach does a good job in detecting the participant's micro-environments. It is worth mentioning that when no data is collected whatsoever, our multi-view model can not detect the micro-environment.

Note that all the plots of the different scenarios are interactive plots, thus users have flexibility to select some areas or to zoom in/out over the plot. Moreover, users can download these plots as a PNG images.

4 CONCLUSION & PERSPECTIVES

In this work, we promote the idea of multi-view learning with stacking to detect the user context from environmental data collected from several sensors plus mobility. We cover also time series segmentation based on change point detection to demonstrate the change points in participants' contexts detected automatically by a multi-dimensional CPD model. Furthermore, we develop COMIC, a visualization GUI to show the different recognized context and highlight the importance of multi-view learning compared to single view learning and other baseline approaches.

We intend to test other ML algorithms besides *RF* and *kNN* at different levels. Indeed, the global performance of the model depends on the learners used for the first and second levels. Thus, we aim to explore different learners such as *SuperLearner*, and confirm their ability to provide better performance. The models with best results are to be considered and included in the visualization interface.



(a) First Level Classification versus Declared Classes.

(b) Multi-view Classification versus Declared Classes.

Figure 4: Classification and CPD Dashboard

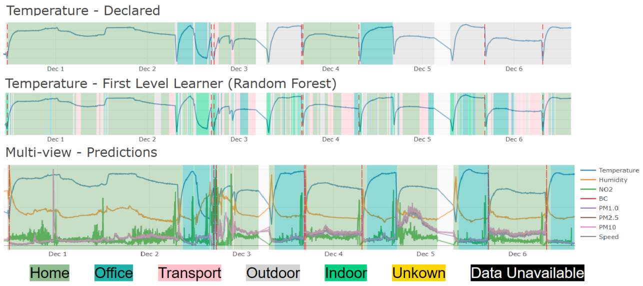


Figure 5: Comparison between single-view and multi-view models.

ACKNOWLEDGMENTS

This work has supported by the French National Research Agency (ANR) project Polluscope, funded under the grant agreement ANR-15-CE22-0018, by the H2020 EU GO GREEN ROUTES funded under the research and innovation programme H2020- EU.3.5.2 grant agreement No 869764, and by the DATAIA convergence institute project StreamOps, as part of the Programme d' Investissement d'Avenir, ANR-17-CONV-0003. Part of the equipment was funded by iDEX Paris-Saclay, in the framework of the IRS project ACE-ICSEN, and by the Communauté d'agglomération Versailles Grand Parc. We are thankful to all the members of the Polluscope consortia who contributed in one way or another to this work: Salim Srairi and Jean-Marc Naude (CEREMA) who conducted the campaign; Boris Dessimond and Isabella Annesi-Maesano (Sorbonne University) for their contribution to the campaign; Valerie Gros, Baptiste Languille and Nicolas Bonnaire (LSCE), and Anne Kauffmann and Christophe Debert (Airparif) for their contribution in the periodic qualification of the sensors and their feedback. Finally, we would like to thank the participants for their involvement.

REFERENCES

[1] Mohammad Abboud, Hafsa El Hafyani, Jingwei Zuo, Karine Zeitouni, and Yehia Taher. 2021. Micro-environment Recognition in the context of Environmental Crowdsensing. *Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference* 2841 (2021).

[2] Stamatelopoulou Asimina, D. Chapizanis, S. Karakitsios, P. Kontoroupis, D. Asimakopoulos, T. Maggos, and D. Sarigiannis. 2018. Assessing and enhancing the utility of low-cost activity and location sensors for exposure studies. *Environmental Monitoring and Assessment* 190 (2018), 1–12.

[3] Hafsa El Hafyani. 2020. Big Data Series Analytics in the Context of Environmental Crowd Sensing. In *2020 21st IEEE International Conference on Mobile Data Management (MDM).* IEEE Computer Society, Los Alamitos, CA, USA, 246–247. https://doi.org/10.1109/MDM48529.2020.00056

[4] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33, 4 (2019), 917–963.

[5] Enrique Garcia-Ceja, Carlos E. Galván-Tejada, and Ramon Brena. 2018. Multi-view stacking for activity recognition with sound and accelerometer data. *Information Fusion* 40 (March 2018), 45–56. https://doi.org/10.1016/j.inffus.2017.06.004

[6] Hafsa El Hafyani, Karine Zeitouni, Yehia Taher, and Mohammad Abboud. 2020. Leveraging Change Point Detection for Activity Transition Mining in the Context of Environmental Crowdsensing. *Actes de la conférence BDA 2020* 1 (2020), 64.

[7] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116 (2019), 237–245.

[8] Baptiste Languille, Valérie Gros, Nicolas Bonnaire, Clément Pommier, Cécile Honoré, Christophe Debert, Laurent Gauvin, Salim Srairi, Isabella Annesi-Maesano, Basile Chaix, et al. 2020. A methodology for the characterization of portable sensors for air quality measure with the goal of deployment in citizen science. *Science of the Total Environment* 708 (2020), 134698.

[9] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research* 21, 118 (2020), 1–6. http://jmlr.org/papers/v21/20-091.html