Indoor occupancy prediction based on recorded data from existing smart thermostats in North American houses

Daphne Bard[1], Camilla Losada[1], Imad Rida[1], Plinio Morita[2], Dan Istrate[1] and Vincent Zalc[1]
[1] Université de technologie de Compiègne, CNRS, Biomechanics and Bioengineering, Centre de recherche Royallieu - CS 60 319 - 60 203 Compiègne Cedex
[2]School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, CA
imad.rida@utc.fr

Abstract

In the present paper, we proposed a tool based on feature extraction and machine learning techniques in order to predict human occupancy from the recorded data in existing smart thermostats. Two introduced feature descriptors namely AAS and APAS were combined with conventional classification including random forest, k nearest neighbor and support vector machine. The preliminary performed experiments on Ecobee dataset showed very promising results.

Keywords: occupancy prediction, feature extraction, machine learning.

I. INTRODUCTION

Nowadays, connected sensors are used everywhere. Indeed, we can find a multitude of various sensor types including, position, pressure, movement, temperature, load, etc. When it comes to the application, innovative sensor technologies have been used in a large variety of applications such as lifestyle, healthcare, fitness, manufacturing, aerospace and agriculture [1]. Among the previously mentioned sensor types, we can particularly notice motion detection. The latter is of extreme importance since it is usually used for different applications including energy management. Certainly, the global climate changing and warming up, is posing increasingly severe risks for ecosystems, human health and the economy. One of the frequent measures to tackle the problem of climate change is energy management through motion detection sensors for a better and reasonable energy consumption [2] [3].

A motion detector is an electrical device capable of detecting a physical movement in a particular area, it may detect the movement of any item or human person. The fact that the sensors are connected allows a multiple use of the recorded and detected information. In the current work, we are interested in Ecobee manufactures smart thermostats. They aim to reduce energy costs by improving the regulation of the temperature according to the need. The need calculation is based on the presence detection in the room thanks to the wireless remote motion sensors present in the different rooms connected to the thermostat [4] [5] [6].

In the case of Ecobee connected thermostat, the movement infrared sensors are used not only to manage heat but also to monitor user behavior. Indeed, the recorded data can be used to extract useful information about people's daily habits as well as the impact of the different situations on their habits, physical and mental health [6]. Moreover, the recent advances in remote sensor technology and the availability of large volumes of data have led to significant improvements in data analysis, particularly when combined with machine learning (ML) algorithms.

In the present paper, based on a large recorded dataset from Ecobee thermostats installed in North America houses, we developed a prediction tool using a simple but yet efficient feature extraction descriptors and machine learning classification techniques. Using our tool, we are getting very promising results to predict the number of occupants present in a house.

II. METHODS AND MATERIALS

An automated prediction tool can be divided into two basic steps [7] [8] [9]:

1) feature extraction where discriminative features are generated from the raw data using feature extraction techniques
2) the classification which assigns a class label based on the extracted features in the first step and a trained classifier.

A. Features extraction

Despite the ability of recorded data to give useful information, it is not always captured in ready and adequate format for prediction which clearly shows the need for novel efficient methods to address this problem. Relevant and discriminative

features are of critical and fundamental importance to achieve high performances in any automatic pattern recognition system [8]. Feature extraction seeks to transform and fix the dimensionality of an initial input raw data to generate a new set of features containing meaningful information contributing to assign the observations to the correct corresponding classes.
In the following we introduced two simple but yet efficient feature descriptors namely, the average of activated sensors (AAS) and the average percentage of activated sensors (APAS).

1) The average of activated sensors (AAS)

The sensor's values from each thermostat were added to each other and the data was grouped by thermostat, day of the week and hour. Finally, the average value of the sum of sensors is calculated for each time of the week. The obtained feature is an average value of the number of active sensors for a given hour and day of the week.

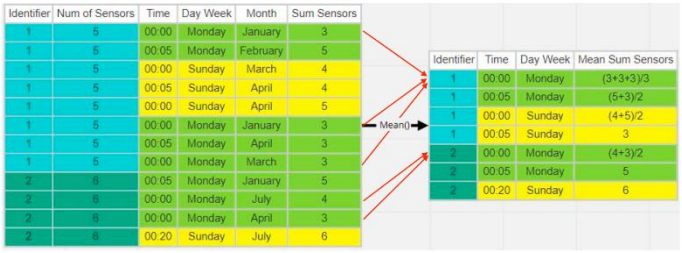

Figure 1. Calculation of the average of activated sensors

1) The average percentage of activated sensors (APAS)

The sum of activated sensors was calculated for each measure and then was divided by the number of sensors belonging to each identifier. The percentage of active sensors for each thermostat every 5 minutes of each day is obtained. Finally, the data is grouped by thermostat, time and day of the week, giving us the average of the percentage of active sensors (APAS).

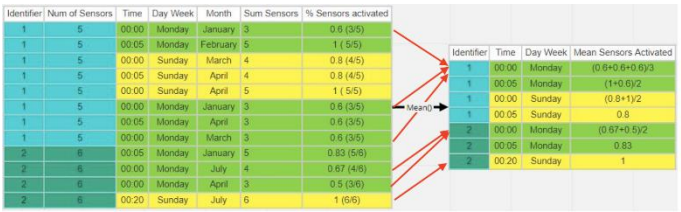

Figure 2. Calculation of the average percentage of activation

B. *Classification*

In the present work, we have used several conventional classifiers, which have shown their efficiency namely, random, forest (RF), support vector machine (SVM) and K nearest neighbours (KNN) [7] [9]. A short description of each classifier is summarized in Table 1.

TABLE I. COMPARISON BETWEEN THE THREE DIFFERENT CLASSIFICATION ALGORITHMS (RANDOM FOREST, KNN AND SVM).

Algorithm name	**Description**	**Advantages**	**Disadvantages**
Random Forest	- A set of decision trees in which the output of each is counted as a vote. -It combines the results of the decision trees. A decision tree models a hierarchy of tests to predict an outcome. [RF]	- Easy to interpret - Stable - Reduce the risk of overfitting - Good performance accuracies -	- Limited performance with small data sets - Requires large datasets for the training
K-Nearest Neighbors (KNN)	-Classify target points according to their distances from points forming the training sample	- Very simple - Provides good results	- Requires a lot of memory and processing resources - Tends to less perform in big datasets
Support Vector Machine (SVM)	Finds the best boundary between different classes to	- Efficient for high dimensional spaces - Memory efficient	- When the number of features is higher than the number of samples, it can lead to overfitting

III. RESULTS AND ANALYSIS

In this section, we evaluate the prediction performance on the Ecobee dataset. We predict the number of occupants in a house, as well as the geographical location. The evaluation is based on a 5-folds cross validation scheme and the performances are

measured by the number of well classified examples over the total number of examples.

A. *Data Set Description*

Ecobee manufactures smart thermostats that include a motion sensor, to which it is possible to connect wireless remote sensors. The number of remote sensors varies depending on how many the user has purchased. All the information captured by the different sensors is sent to the main thermostat via Wi-Fi. These data are accessible thanks to the "Donate Your Data" program[1], where users can donate their data which is then anonymized) [5].

Ecobee's database contains the data of 111296 thermostats and for each thermostat we know the associated user's identifier, the number of remote sensors, the number of people living in the house, the number of floors in the house, country, state, city, among other information. Each thermostat can belong to one user's account, and the same user can have several identifiers. Each sensor measures temperature and movement, every 5 minutes. The values returned by the motion sensors are 1 if motion has been detected and 0 if not. These sensors are placed by the user and the user can move them at any time. In this analysis, we are only interested in the movement data from the thermostat and its remote sensors.
For our analysis, we only selected the data from 2019 measurements and removed non-representative values. The thermostats come mainly from the US, but also from Canada, most of the thermostats have less than three sensors.

B. *Prediction of the number of occupants*

At the beginning, we tried to predict the number of occupants in each house based on their movement information. We have created four classes (houses with one, two, three and four occupants). We gradually started from a binary classification problem to discriminate between houses (with one and two occupants), houses (with three and four occupants) to finally a multi-class classification problem to discriminate between houses (with one, two, three and four occupants). In total we worked with 1878 thermostats for every class.

1) *Results based on AAS features*

The performances to differentiate between houses with one and two occupants carried out using AAS features and combined with Random Forest, KNN and SVM were. 0.65 (+/- 0.03), 0.61 (+/- 0.03), and 0.63 (+/- 0.03) respectively. The corresponding confusion matrices are shown in Figure 3. When it came to discriminating between houses with three and four occupants, the obtained performances were slightly lower as follows: 0.59 (+/- 0.03), 0.57 (+/- 0.05) and 0.61 (+/- 0.03). The confusion matrices are depicted in Figure 4. Finally, to differentiate between houses with one, two, three and four occupants, the best performance was obtained using SVM 0.43 (+/- 0.03), while KNN and Random Forest obtained 0.40 (+/- 0.01) and 0.41 (+/- 0.03) respectively. The corresponding confusion matrices are also shown in Figure 5.

2) *Results based on APAS features*

The same experiments have been carried out using the second feature descriptor namely APAS introduced in order to reduce the loss of information of the first descriptor (AAS). The obtained results to discriminate between houses with one and two occupants were very close to the ones obtained using AAS features: 0.66 (+/- 0.03), 0.65 (+/- 0.04), and 0.65 (+/- 0.04). The corresponding confusion matrices are shown in Figure 6.

To distinguish between houses with three and four occupants, we reached 0.60 (+/- 0.06), and 0.57 (+/- 0.05) and 0.61 (+/- 0.06). The corresponding confusion matrices are shown in Figure 7. Finally, to distinguish between houses with one, two, three and four occupants, we obtained the following accuracies, 0.42 (+/- 0.04), 0.41 (+/- 0.04) and 0.44 (+/- 0.04). The confusion matrices are in Figure 8.

When analysing our dataset, we have observed that the number of thermostats is different from a house to another one. Furthermore, in some specific moments, a large part of the thermostats was not able to record information. These missing information are more likely due to company maintenance moments or internet connection failures. Further work can be envisaged in order to tackle the problem of missing data using different techniques such as the interpolation.

When comparing the obtained results using the two different feature descriptors, namely AAS and APAS, it can be seen that the two descriptors were able achieve very close and similar results. The four classes' problem (one, two, three and four occupants) remains very challenging. It should be noted that the occupants' habits based on the season and geographical localisation influence the obtained results and makes the problem very challenging. Low sampling information is a further cause (one information is sent each five minutes). In other terms, our classification problem suffers from large intra-class variations due to the over mentioned reasons.

IV. Conclusion

A simple but yet efficient tool based on feature extraction and machine learning has been proposed in order to predict house

[1] [Ecobee, *Donate your Data : Smart homes and thermostats*, https://www.ecobee.com/donate-your-data/]

occupancy. Preliminary obtained results on Ecobee dataset using two different feature descriptors and three conventional classifiers have shown very promising results. Further future works are envisaged in order to tackle the problem of intra-classification due the missing information, geographical localisation, season and the low sampling.

TABLE II. SUMMARY OF THE OBTAINED RESULTS OF THE DIFFERENT ANALYSIS

Features	**Classes**	**Accuracy**		
		KNN	SVM	Random Forest
AAS	1,2 occupants	0.61 +- 0.03	0.63 +- 0.03	**0.65 +- 0.03**
	3,4 occupants	0.57 +- 0.05	**0.61 +- 0.05**	0.59 +- 0.03
	1,2,3,4 occupants	0.40 +- 0.01	**0.43 +- 0.03**	0.41 +- 0.03
APAS	1,2 occupants	0.65 +- 0.04	0.65 +- 0.04	**0.66 +- 0.03**
	3,4 occupants	0.57 +- 0.05	**0.61 +- 0.06**	0.60 +- 0.06
	1,2,3,4 occupants	0.41 +- 0.03	**0.44 +- 0.01**	0.42 +- 0.03

ACKNOWLEDGMENTS

The authors thank, M. Plinio Morita and University of Waterloo for the access to data from Ecobee sensors. This research work was funded by université de technologie de Compiègne in the framework of the project ADL identification.

REFERENCES

[1] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). A survey on sensor networks. IEEE Communications magazine, 40(8), 102-114.

[2] Xue, Y., Ju, Z., Xiang, K., Chen, J., & Liu, H. (2018). Multimodal human hand motion sensing and analysis—A review. IEEE Transactions on Cognitive and Developmental Systems, 11(2), 162-175.

[3] Lu, J., Sookoor, T., Srinivasan, V., Gao, G., Holben, B., Stankovic, J., ... & Whitehouse, K. (2010, November). The smart thermostat: using occupancy sensors to save energy in homes. In Proceedings of the 8th ACM conference on embedded networked sensor systems (pp. 211-224).

[4] Oetomo, A., Jalali, N., Costa, P. D. P., & Morita, P. P. (2022). Indoor temperatures in the 2018 heat wave in Quebec, Canada: exploratory study using Ecobee smart thermostats. JMIR formative research, 6(5), e34104.

[5] Sahu, K. S., Oetomo, A., Jalali, N., & Morita, P. P. (2021, June). Household and population-level behavioural changes due to Covid-19 pandemic: A smart thermostat based comparative data analysis. In Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care (Vol. 10, No. 1, pp. 1-6). Sage CA: Los Angeles, CA: SAGE Publications.

[6] Sahu, K. S., Oetomo, A., & Morita, P. P. (2020). Enabling remote patient monitoring through the use of smart thermostat data in canada: exploratory study. JMIR mHealth and uHealth, 8(11), e21016.

[7] Rida, I., Al-Maadeed, N., Al-Maadeed, S., & Bakshi, S. (2020). A comprehensive overview of feature representation for biometric recognition. Multimedia Tools and Applications, 79, 4867-4890.

[8] Rida, I. (2018). Feature extraction for temporal signal recognition: An overview. arXiv preprint arXiv:1812.01780.

[9] Rida, I. (2017). Temporal signals classification (Doctoral dissertation, Normandie).

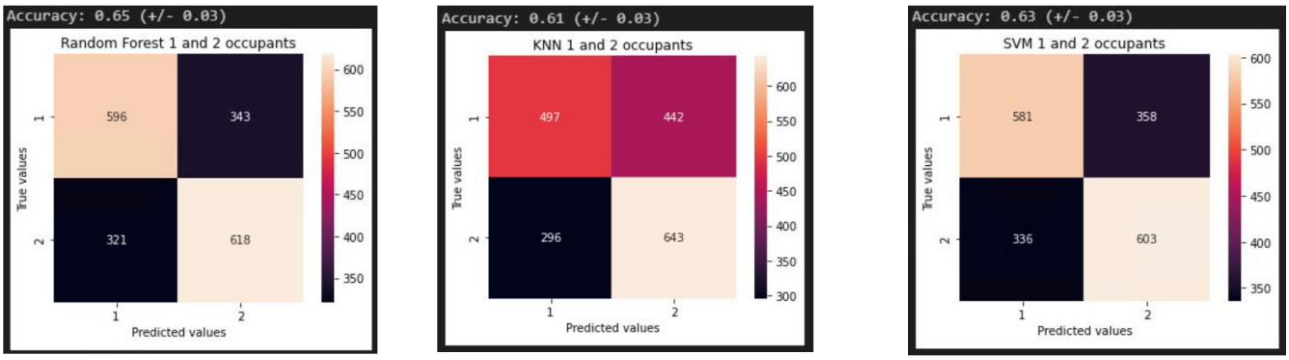

Figure 3. Confusion matrices for the classification of houses with one or two occupants using Random Forest, KNN and SVM

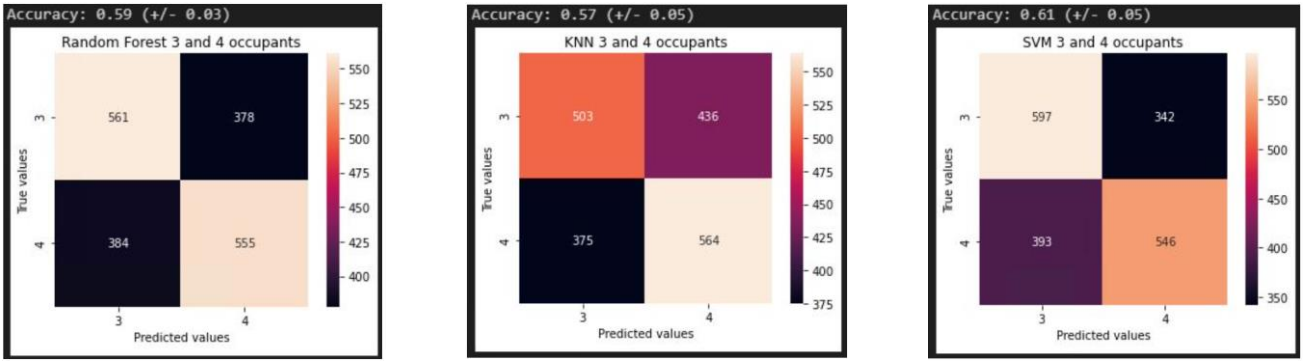

Figure 4. Confusion matrices for the classification of houses with three and four occupants using Random Forest, KNN and SVM

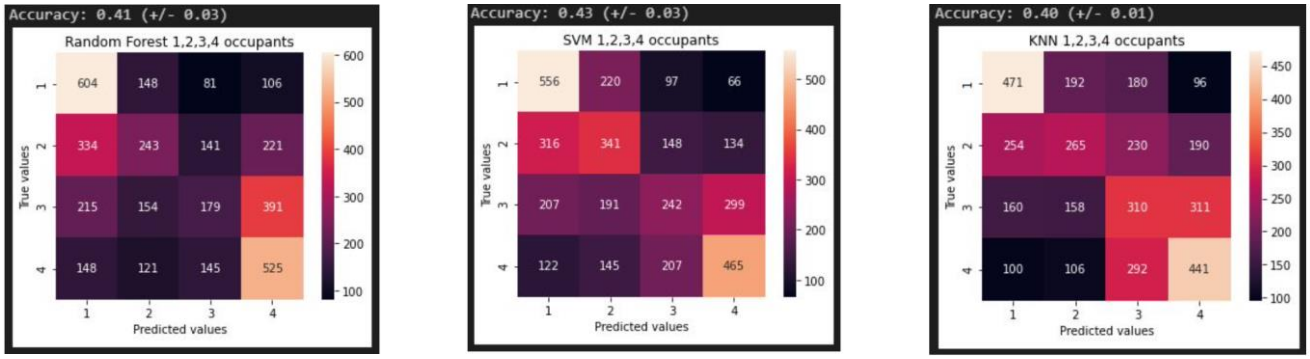

Figure 5. Confusion matrices for the classification of houses with one, two, three and four occupants using Random Forest, KNN and SVM

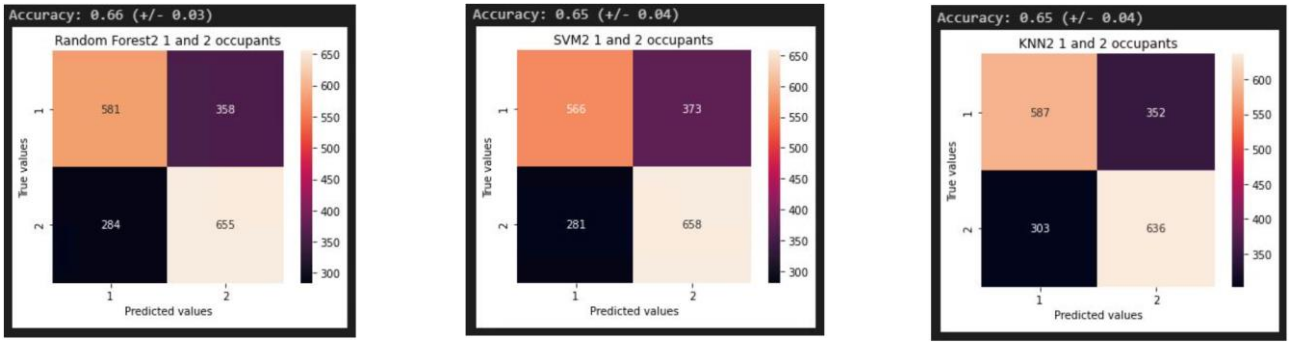

Figure 6. Confusion matrices for the classification of houses with one and two occupants using Random Forest, SVM and KNN

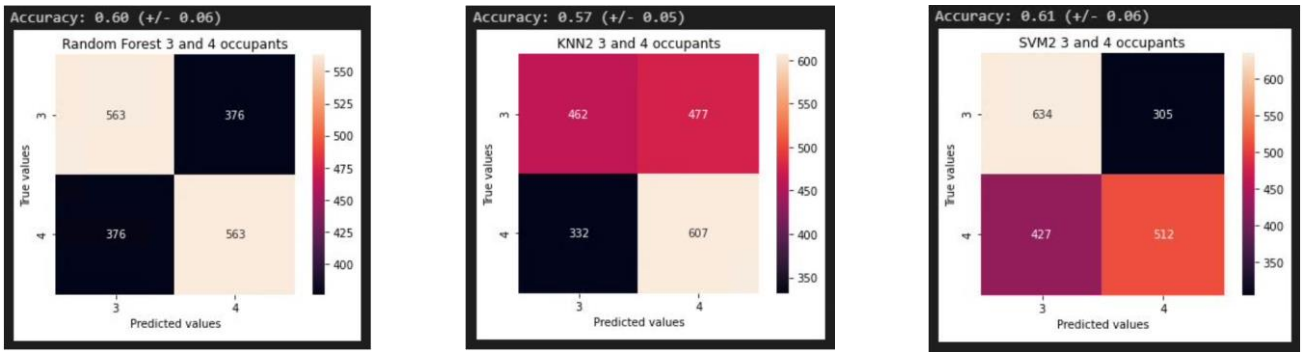

Figure 7. Confusion matrices for the classification of houses with three and four occupants using Random Forest, KNN and SVM

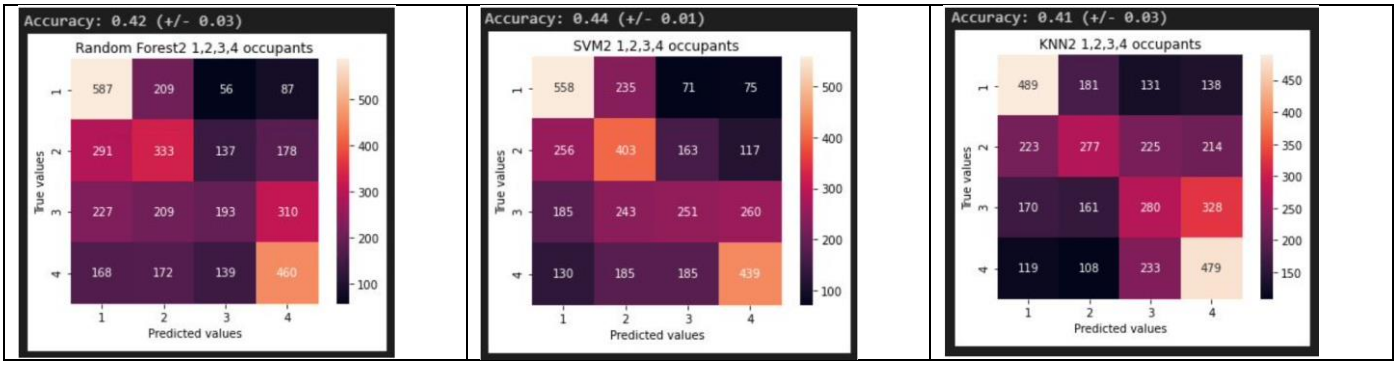

Figure 8. Confusion matrices for the classification of houses with one, two, three or four occupants fusing Random Forest, SVM and KNN