*M.J. Hornos et al. (Eds.)*

*doi:10.3233/AISE240030*

# Generating Data in an Intelligent Environment and Systematically Predicting User Activity via Bluetooth Low Energy Beacons and Machine Learning

Stefan Michael Anthony BENEDICT, Juan Carlos AUGUSTO and Omer KACAR
*Department of Computer Science*
*Middlesex University London, United Kingdom*
ORCiD ID: Juan Carlos Augusto https://orcid.org/0000-0002-0321-9150

**Abstract.** The field of Intelligent Environments is experiencing an upward growth trajectory. This investigation delves into the state-of-the-art technology associated with the Smart Spaces Lab at Middlesex University London, specifically its Bluetooth Low Energy beacons. This enables User Data Generation enabling applications including User insight generation and Activity Recognition via Machine Learning algorithms.

**Keywords.** intelligent environments, machine learning, activity recognition, data generation

## 1. Introduction

Intelligent Environment(s) (IE) is a growing industry associated with domains including Ambient Assisted Living and Smart Homes (SHs). Smart Homes (i.e., a form of IE), equipped with smart technology, allow customers to experience customised services. However, enabling User-specific customised services within a multiuser IE is driven on the identification of each User and their specific location within the IE [1]. An ancillary benefit of solving the challenge is the copious quantity of generated Data, giving insight into the User's behaviour (e.g., eating habit, sleeping pattern, activity pattern, etc.). This Data can provide valuable insights into the health (i.e., progression of diseases) of individuals with special needs, the energy utilisation patterns of the residents, enabling the optimisation of consumption [2,3].

The study was carried out at Middlesex University's Smart Spaces Lab utlises Bluetooth Low Energy (BLE) technology building on the approach described in [1], This study generated User Data, enabling User Insight generation and Activity Recognition (AR) via Machine Learning (ML). The remainder of the paper is segmented into three sections. Section 2 focusses on the System Configuration Methodology, Section 3 that is related to Data Collection while Section 4 is focused on AR and ML.

## 2. System Configuration Methodology

This investigation is built upon by leveraging the well-established Software (i.e., Web Servers, Databases, an Android application, etc.) and Hardware architecture (i.e., Smart sensors, smart switches, BLE Beacons, etc.) of the Smart Spaces Lab of Middlesex University London as described in [1]. The system configuration is displayed in Figure 1. Further, the computational hardware utilised in this segment of the investigation consisted of an 8th generation Intel Core i5 processor, 16 GB RAM and an 8 GB GPU.

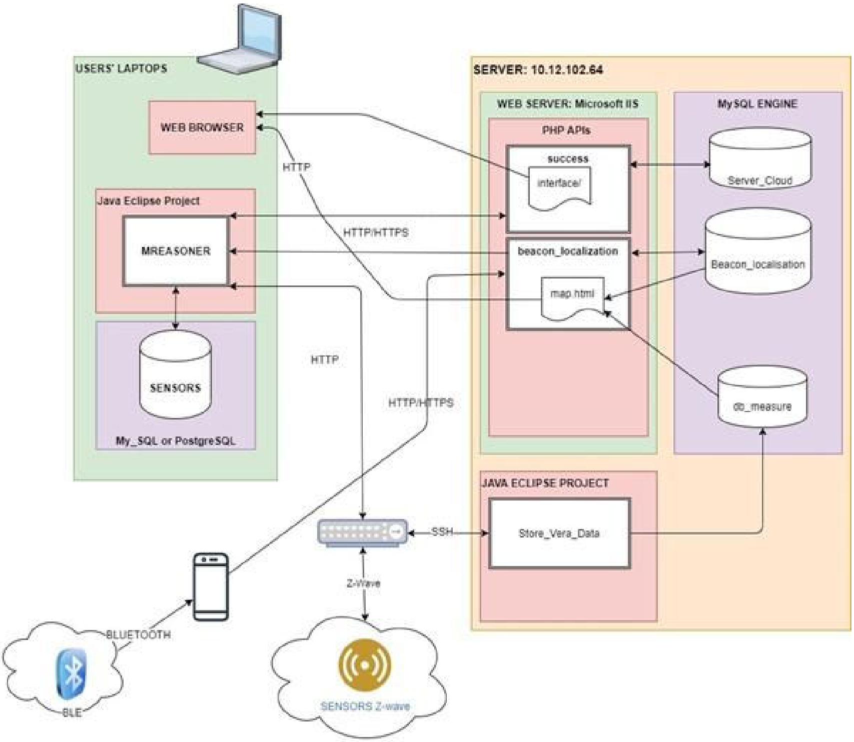

**Figure 1.** The system architecture of the Smart Spaces Lab.

## 3. Data Collection

All locations, except for Room 04, Room 05, and the Shower, are used for Data Collection. The Data that can be generated in the Smart Spaces Lab utilising its technology, as displayed in Figure 2. The technology within the Lab includes a multitude of sensors which include but is not limited to motion sensors, proximity door sensors as well as BLE Beacons to determine User/ Resident location. It should be noted that the "Multi Sensor" (s) provides temperature readings within the SH, but this was excluded from the investigation due to functionality issues. Table 1 illustrates a sample of the type of Data used/final feature space within this research by utilising the sensory equipment in the Lab and their respective descriptions. It is crucial to note that the Received Signal Strength (RSSI) algorithm used to determine the User location via BLE Beacons is based on [1]. However, any improvements to the algorithm are considered out of scope for this investigation. Moreover, this investigation only considers the final output of the Resident/User's location. The complete version of the collected Data can be found in [4] under the file "Compiled User Data.csv".

**Table 1.** Data collection/ type sample generated at the Smart Spaces Lab at Middlesex University London.

| Feature Name | Description | Unit/ Type |
|---|---|---|
| Specific User | The Specific User within the IE | N/A |
| Time Stamp | Time of the activity | s/min |
| Type of Day | Weekday/Weekend | N/A |
| Activity | The activity (i.e., eating, sleeping, etc.) that the User carries out | N/A |
| Location | Location of the User | N/A |
| Routine | Morning/Evening/Night | N/A |
| **Bedroom Data** | | |
| Bed Pressure Sensor | Senses if the User is on the bed | Binary |
| Bedroom Motion Sensor | Senses the movement of the User | Binary |
| Bedroom Door Sensor | Senses if the door has been opened | Binary |
| Bedroom Light Switch | Senses if the lamp is turned on | Binary |
| Energy Sensor 1 | Senses if an electrical appliance is turned on | Binary |
| Bedroom BLE Beacon | Used to determine the location of the User | mW |
| **Living room Data** | | |
| Couch Pressure Sensor | Senses if the User is on the couch | Binary |
| Energy Sensor 2 | Senses if an electrical appliance is turned on | Binary |
| Living room Motion Sensor | Senses the movement of the User | Binary |
| Living room Door Sensor | Senses if the door has been opened | Binary |
| Garden Door Sensor | Senses if the door has been opened | Binary |
| Living room Light Switch | Senses if the lamp is turned on | Binary |
| Living room BLE Beacon(s) | Used to determine the location of the User | mW |

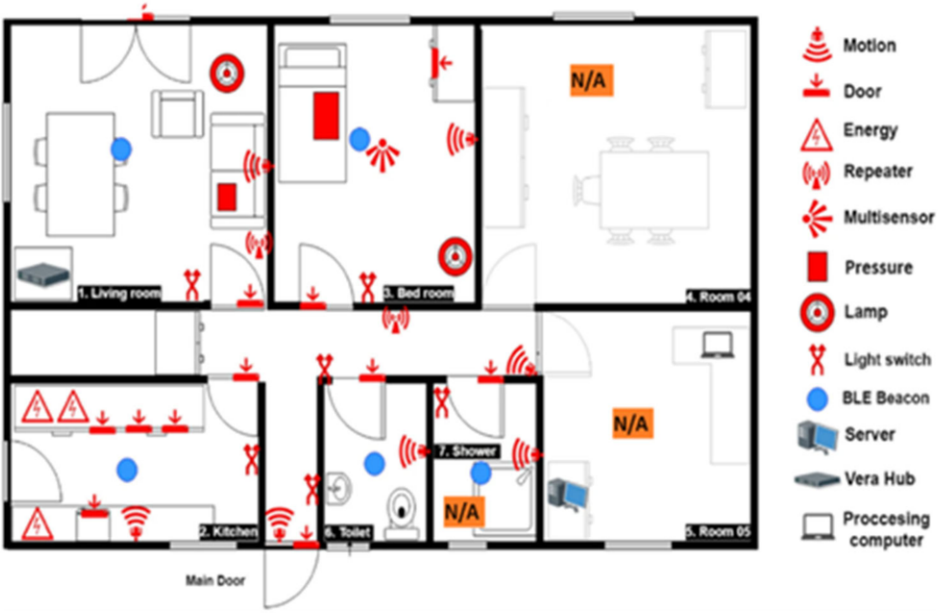

**Figure 2.** Smart Spaces Lab location utilisation [5].

### *3.1. Data Collection Overview and its Applications*

The first phase/input of the Data Collection revolved around formulating Hypothetical scenarios (e.g., User wakes up and rests on the bed, User heats food and eats it, etc.) within an IE, specifically a SH. Hypothetical scenarios had to be used due to the residential restrictions placed by Middlesex University London on the Smart Spaces Lab. The generated scenarios account for a multitude of elements, including two Users, type of day (i.e., Weekday or Weekend), a variety of activities, a selection of Routines, and so forth. This specific Data (i.e., Hypothetical scenarios) can be accessed via the "Data Collection Scenarios.xlsx" file in [4]. Next, the Hypothetical scenarios were acted out in person in the Smart Spaces Lab, and the generated Data, which was logged on to the relevant Databases, was downloaded via a cross-platform Database software specifically "DBeaver" and then compiled using a secondary software. The final feature space consisted of 33 features and 2678 instances. Due to the number of observations being more significant than ten times that of the features, the Data is considered low dimensional, which is beneficial for ML. This is observed and discussed in the upcoming segments of this investigation. The generated Data can be obtained from [4]. The Data can be further explored to generate User insights which includes User habits such as the time spent carrying out an activity by the User, and the time spent in a specific location. A selection of insights can be found in [4]. The next section discusses the ML element of the investigation, beginning with its implementation and concluding with its evaluation.

## 4. Activity Recognition Leveraging Machine Learning

The ML component of this investigation is based on six models, specifically a Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), a Dummy classifier (DC) model, and a tuned RF model. The DC model utilised follows a uniform strategy and was the baseline model. The classes predicted from the models included Eating, Resting, Sleeping, and Miscellaneous activities. After the models were executed, their performances were evaluated to distinguish the best model. The best model in this investigation was quantified as the RF model and was tuned. The pipeline was implemented using the Python programming language. The Data is then split into the independent and Target variables (i.e., User Activity) as this is a classification-based ML challenge. Due to the reason that ML models work purely on numerical input Data, the Data had to be preprocessed before the implementation of the ML pipeline. The preprocessing was executed by converting the Boolean variables (i.e., Lights, motion sensors, etc.) of the Data into numeric format (i.e., from True or False to 1 or 0) and using the approach of Label encoding for the remaining independent categorical variables, which include the "Weekday/Weekend", "Location" and "Routine" attributes. It was observed that there are 2,678 instances corresponding to 32 independent variables. Next, the Data was split into Train and Test Data with a training size of 70% corresponding to 1874 instances. The next and final phase of this was to implement the ML pipeline of the DT, RF, LR, and NB models. It is imperative to emphasize that the data was further preprocessed before executing the DT through scaling the Data for better performance. This is not required for the remaining models. Moreover, a random state parameter was included in the relevant models, including the LR, DT, and RF models, ensuring the models' reproducibility. The DT model obtained 100% accuracy on the training Data and 99.8% on the test Data, whereas the RF models achieved 99.9% on both the train and test sets. The LR model scored 95.7% on the training set and 96.4% on the test set, whereas naïve Bayes NB achieved 73.5% accuracy on training and 72.0% on test Data and the DC model fared with an accuracy of just 24.1%. A visual representation of the accuracy variation between the DC model and the tuned RF model can be found in [4], under the "User Activity Interface (Machine Learning.mp4" file. Moreover, Figure 3 displays the most important features of the RF model in descending order. The top three attributes associated with the high influence on the model are the Location, Time, and Bedroom Bed Pressure. The contributions are over 25%, just under 20%, and just over 10%, respectively. Strikingly, only five features are associated with a contribution of over 5%.

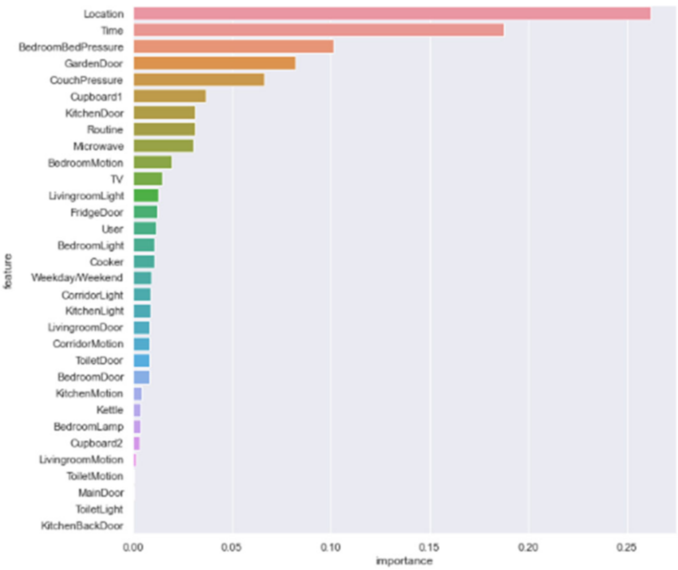

**Figure 3.** The most important features of the Random Forest Machine Learning Model.

## 5. Conclusion

The investigation carried out at the Smart Spaces Lab at Middlesex University London demonstrates the value potential of Data which can be generated with an Intelligent Environment and its high potential in Machine Learning based applications. The Data generated can be further used in other investigations for further including Insight Generation and alternative Machine Learning based applications. The main caveat of the investigation is that the Data is based on Hypothetical scenarios due to the restrictions placed on the residents by the University to implement a more realistic Data Collection approach.

## References

[1] M. Quinde, J.G. Gimenez Manuel, C.L. Leonard and J.C. Augusto, "Achieving multi-user capabilities through an indoor positioning system based on BLE beacons," in - 2020 16th International Conference on Intelligent Environments (IE), 2020, DOI: 10.1109/IE49459.2020.9155011.

[2] J.G. Gimenez Manuel, J.C. Augusto and J. Stewart, "AnAbEL: towards empowering people living with dementia in ambient assisted living," Universal Access in the Information Society, vol. 21, (2), pp. 457-476, 2022.

[3] M. Z. Fakhar, E. Yalcin and A. Bilge, "A survey of smart home energy conservation techniques," Expert Syst. Appl., vol. 213, pp. 118974, 2023, DOI: 10.1016/j.eswa.2022.118974.

[4] S. M. A. Benedict, "Distinguishing between multiple Users in an Intelligent Environment and their Activities while gaining insights into their Behaviour: A Data-driven Approach". Middlesex University, 29-Sep-2023, doi: 10.22023/mdx.24216723.v1.

[5] A Smart Environments Architecture (SEArch), J. C. Augusto, J. G. Gimenez-Manuel, M. Quinde, Ch. Oguego, M. Ali, C. James-Reynolds. Applied Artificial Intelligence, 34 (2) pp. 155-186. ISSN 0883-9514. Taylor and Francis. 2020.