11th MALAYSIA STATISTICS CONFERENCE 2024

Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

Machine Learning Techniques to Classify Individual's Smart Homes Activities

CHONG KAR YIN[1], GAN CHEW PENG[1*], KOH SIEW KHEW[2], TAN YI-FEI[3]

[1] School of Accounting and Finance, Taylor's Business School, Taylor's University, 47500 Subang Jaya, Selangor Darul Ehsan, Malaysia.
karyin.chong@sd.taylors.edu.my, chewpeng.gan@taylors.edu.my
[2] Department of Mathematics, Xiamen University Malaysia, 43900 Sepang, Selangor, Malaysia.
siewkhew.koh@xmu.edu.my
[3] Faculty of Engineering, Multimedia University, Cyberjaya 63100, Selangor, Malaysia.
yftan@mmu.edu.my

Abstract:
With the rising elderly population and the increasing prevalence of cognitive difficulties like dementia, there is a critical need for innovative solutions to support independent living and reduce reliance on caregivers. This study explores the application of Machine Learning (ML) techniques in smart homes to classify and predict Activities of Daily Living, aiming to enhance healthcare monitoring and personalized assistance for the elderly. By leveraging the Internet of Things and ML techniques, smart home systems are ideally aimed to provide timely alerts and proactive interventions that can significantly improve the quality of life for elderly residents. Our research goes beyond traditional activity-based features by exploring if temporal features help to enhance model accuracy. We compare the effectiveness of several ML algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN) and Gradient Boosting (GB), in interpreting sensor data and identifying meaningful behaviour patterns. The results show that GB outperformed other ML algorithms, achieving an impressive 0.95 in F1 score and 0.93 in Matthews Correlation Coefficient for activity recognition. DT and RF, which share similar underlying logic with GB also performed well, each scoring above 0.92 in key metrics.

Keywords:
Activities of Daily Living; Supervised Learning; Feature Selection; Gradient Boosting

1. Introduction:
By 2040, the number of Malaysians aged 65 and above is projected to reach 6 million, nearly four times the 1.43 million recorded in 2010 (DOSM, 2016; UNESCAP, 2023). Despite longer lifespans due to medical advances, many elderly individuals face significant health issues, including a rise in dementia cases (Choi et al., 2024). In Malaysia, it is estimated more than 260,000 people, approximately 8.5% of the elderly population, are living with this form of disabilities (World Health Organization, 2020).

Unfortunately, this number is expected to grow due to longer life spans and lower birth rates in the country. This demographic shift impacts healthcare systems, economies, labour markets, social services, and infrastructure development. Low awareness of dementia-related issues has led to under-diagnosis and insufficient support in Malaysia, highlighting the need for innovative solutions like smart homes (Lai et al., 2022).

Smart home technologies provide healthcare monitoring and personalized assistance by connecting sensors throughout the home, allowing the identification of unusual behaviours or potential health issues without 24-hour caregiver support (Facchinetti et al., 2023). Human Activity Recognition (HAR) is crucial for these systems, detecting and identifying human behaviours to customize assistance and optimize elderly care. The complexity of analysing temporal and spatial data in smart homes requires sophisticated modelling, driven by advancements in IoT technologies (Nižetić et al., 2020).

To address these challenges, Machine Learning (ML) tools have become indispensable for interpreting the vast and varied data generated by interconnected sensors and devices in smart homes. ML techniques have proven essential for classifying and predicting Activities of Daily Living (ADLs), especially in the context of elderly care. A variety of algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Naive Bayes (NB), K-Nearest Neighbours (KNN), and Gradient Boosting (GB), have shown promise in activity recognition. The effectiveness of these algorithms varies based on specific classifier parameters, feature extraction methods, pre-processing techniques, and dataset characteristics.

This research aims to examine and contrast various ML methodologies for classifying and predicting ADLs in smart home environments. While previous studies mainly focus on activity-related characteristics such as activity duration, activity count, and sensor activation, we are adopting a novel approach by investigating temporal features. It is hypothesized that the timing of activities, daily routines, and resident behaviours follow recurring patterns on daily, weekly, or monthly cycles. The incorporation of temporal features, such as time windows and day of the week, is aimed to uncover new patterns and relationships in elderly activities that are not captured by conventional activity-based features. This approach may facilitate a more detailed comprehension of residents' behaviours and requirements, thereby enhancing the precision and dependability of activity recognition models. In order to ensure a balanced assessment of model performance, our framework employs a diverse set of metrics that address common challenges in ADLs classification, such as data imbalance and overlapping activities. By focusing on these temporal dynamics, this research aims to advance HAR in smart home environments, thereby facilitating the development of more accurate, reliable, and user-friendly systems for elderly care.

2. Methodology:

In this paper, the Aruba dataset, which is part of the CASAS (Center for Advanced Studies in Adaptive Systems) project carried out by the WSU (Washington State University), is selected among the public smart home datasets to evaluate the proposed algorithm approach (Cook, 2012). This dataset is collected from a single-residential smart home testbed of a volunteer woman for 7 months. The raw dataset is represented in date, time, sensor ID, sensor status and activity annotation as shown in Figure 1. There are 11 activities annotated within the Aruba dataset and consist of over 1.7 million rows of data. For motion and door sensors, the status is recorded in binary data such as “ON/OFF” and “OPEN/CLOSE”, else for temperature sensors, data is presented in numerical data with one decimal point.

	timestamp	sensor_id	sensor_value	activity_annotation
0	2010-11-04 00:03:50.209589	M003	ON	Sleeping="begin"
1	2010-11-04 00:03:57.399391	M003	OFF	None
2	2010-11-04 00:15:08.984841	T002	21.5	None
3	2010-11-04 00:30:19.185547	T003	21	None
4	2010-11-04 00:30:19.385336	T004	21	None
...	...	...	...	...
1719553	2011-06-11 23:42:59.028507	T002	25.5	None
1719554	2011-06-11 23:48:02.888409	T001	23.5	None
1719555	2011-06-11 23:48:02.988798	T002	25	None
1719556	2011-06-11 23:53:06.004292	T002	25.5	None
1719557	2011-06-11 23:58:10.004834	T002	25	Sleeping="end"

Figure 1 Sample of CASAS Aruba Raw Datasets

Data pre-processing is essential to align the raw data with our proposed algorithm. After cleaning the data to remove irrelevant information and handle missing values, we extracted several features for activity recognition. Two main tables were prepared for this purpose: one based on conventional activity-based features and another incorporating temporal features.

The first table focuses on activity-based features. The "Duration" feature captures the total time spent on specific activities in minutes. The "Sensor Value" feature indicates the total activations of each motion and door sensor during the activity period, while temperature sensors were excluded to keep the model focused on the most relevant data. The "Activities per Day" feature shows the frequency of each activity, and the "Activity" feature consists of 11 different activities as the target variable.

To explore if the addition of temporal features improves the performance of various ML models, we created a second table with time-based features. This table includes the same activity-based features but adds temporal elements. The "Day" feature records the *n*-th day of the experiment, aiding in data separation during cross-validation. The "Day of Week" feature accounts for weekly activity patterns, which may influence activity recognition. The "Time Window" feature slices the day into 1-minute intervals, indexing each timestamp based on the time of day. Managing activities that span across midnight is crucial for accurate temporal feature representation. When an activity extends past midnight, it starts a new record for the new day. This approach ensures continuity and precise recording of overnight activities while maintaining daily-based feature integrity.

Figure 2 reveals significant variability and consistency in activity durations. "Relax," "Meal Preparation," and "Work" display high variability and many outliers, contrasting with the more consistent activities like “Sleeping” and "Housekeeping." The skewed distributions of most activities, except for "Sleeping," "Respirate," and "Housekeeping," indicate that ML models may benefit from non-linear transformations or specialized architectures to capture these asymmetric patterns effectively.

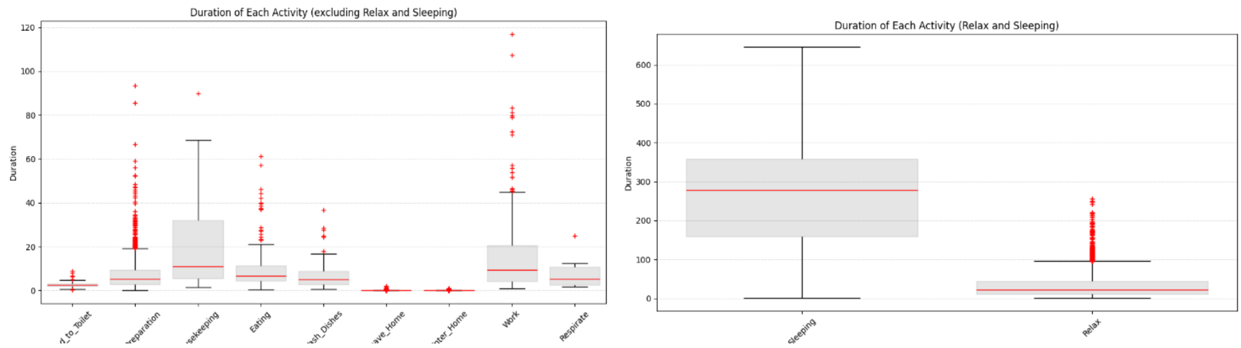

Figure 2 Duration for Each Activity (Boxplot)

Figure 3 uncovers the frequency patterns of daily activities. "Enter Home" and "Leave Home" show similar distributions, which may pose challenges for classification. "Meal Preparation" and "Relax" exhibit more outliers, indicating greater variability in their daily occurrence. Other activities demonstrate more consistent patterns, with fewer outliers and narrower ranges in their daily frequencies.

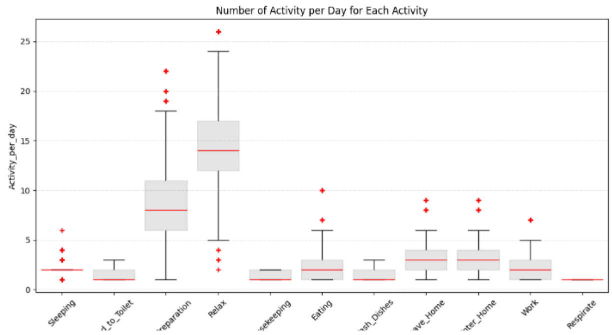

Figure 3 Number of Activity per Day for Each Activity (Boxplot)

Figure 4 visualizes daily activity patterns using 1-minute intervals over 220 days. Sleep cycles and relaxation periods show consistent routines, while most other activities lack fixed patterns. This contrast presents an intriguing challenge for activity classification models. Notably, "Leave Home" and "Enter Home" activities show similar distributions and sparse occurrences during certain time windows, potentially complicating the model's ability to distinguish between these actions and periods of genuine inactivity.

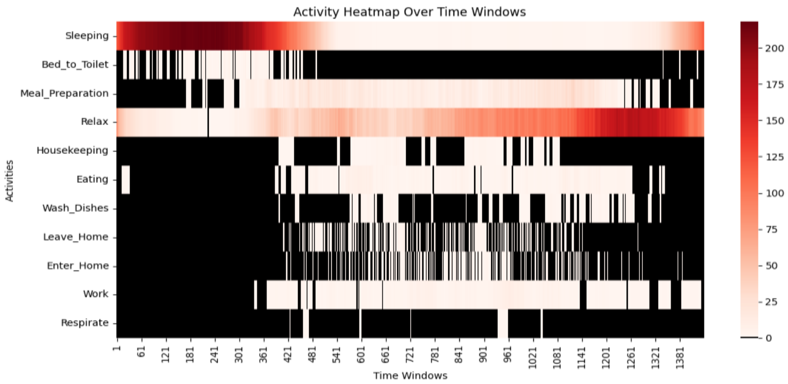

Figure 4 Activity Heatmap Over Time Windows (Heatmap)

These two tables are applied in various ML techniques to classify the ADLs in smart home environments. In this paper, we focus on supervised learning models, including SVM, DT, RF, NB, KNN, and GB. Each model will be trained and validated using a 70:30 train-validation split. For activity-based features table, 70% of instances from each activity are selected for training, while for daily-based features table, 70% of the 220 days are randomly selected for training.

To address imbalanced datasets, Matthews Correlation Coefficient (MCC) and F1 score for evaluation. MCC considers true positives, true negatives, false positives, and false negatives, providing a balanced measure. The F1 score, defined as the harmonic mean of Precision and Recall, ensures robust evaluation beyond accuracy alone.

3. Results:

Figure 5 reveals the performance comparison of six ML classifiers using datasets with and without time-related features. Overall, algorithms like GB and RF consistently outperformed other algorithms, with GB achieving the highest F1 score of 0.95 and MCC of 0.93 in the activity-based dataset. DT and KNN also performed well, while SVM and NB showed lower performance.

The addition of time-related features generally improved model performance, albeit to varying degrees. SVM showed the most significant improvement with temporal features, suggesting its effectiveness in capturing time-dependent patterns. RF slightly outperformed GB when temporal features were included, achieving the highest overall scores (F1: 0.95, MCC: 0.93). GB, while still performing excellently, showed minimal improvement with temporal features, indicating its already robust performance on activity-based features alone. Notably, NB's performance decreased significantly with temporal features, highlighting its limitations in handling complex, interdependent data.

Analysis of individual activities revealed that most classifiers excelled at recognizing 6 out of 11 activities: "Relax", "Meal_Preparation", "Sleeping", "Eating", "Work", and "Bed_to_Toilet". Temporal features particularly enhanced the recognition of activities with strong time-of-day associations, such as "Sleeping" and "Eating". However, they had minimal impact on distinguishing between similar activities like "Leave Home" and "Enter Home". Activities like "Wash Dishes" and "Housekeeping" proved challenging for most classifiers, likely due to inconsistent sensor activations, though RF and GB showed improved performance for these activities with the inclusion of time-related features.

In conclusion, while adding time-related features generally improved model performance, the enhancement was modest for top-performing algorithms like GB and RF. This suggests that sophisticated tree-based methods can effectively capture activity patterns even without explicit temporal information, but the inclusion of such features can provide incremental improvements in overall accuracy and specific activity recognition. The study highlights the potential of temporal features in enhancing activity recognition in smart home environments, particularly for activities with distinct temporal patterns.

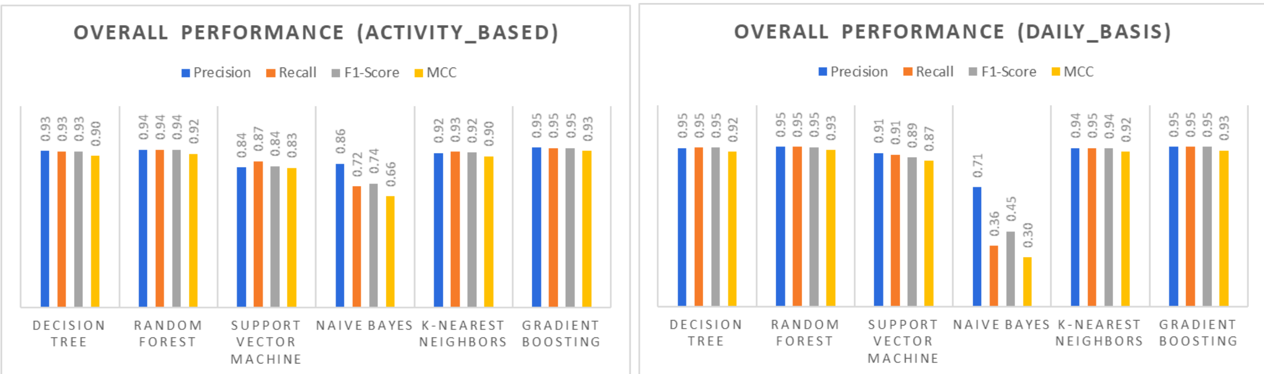

Figure 5 Comparison of Different Classifiers for Both Feature Table

4. Discussion and Conclusion:

In conclusion, this study demonstrates the potential of ML in classifying ADLs within smart home environments. The comparison of various ML algorithms reveals that tree-based methods, particularly GB, RF and DT, excelled in activity recognition tasks. These algorithms' success can be attributed to their ability to capture complex, non-linear relationships in sensor data and handle the inherent variability of human behaviours. Among three, GB resulted the best with 0.95 F1 score and 0.93 MCC. KNN's result is nearly behind the 3 algorithms. However, SVM and NB showed lower performance in activity recognition. Upon analysis, NB's poor performances can be attributed to its assumption of feature independence, which doesn't align with the interrelated sensor activations in ADLs. Similarly, SVM's attempt to create strict classification boundaries proved inadequate for ADLs, which often have overlapping characteristics and variable sensor activation patterns.

The incorporation of temporal features generally enhanced model performance, especially for activities with strong time-of-day associations. However, the modest improvement in top-performing algorithms suggests that sophisticated methods can effectively capture activity patterns even without explicit temporal information. This insight is valuable for future smart home system designs, indicating that while temporal features can provide incremental improvements, they may not be critical for achieving high accuracy in activity recognition. Future research should focus on addressing the limitations in recognizing similar activities and exploring more advanced feature engineering techniques to further enhance the reliability and adaptability of smart home systems for elderly care.

References:

1. Choi, M., Sempungu, J. K., Lee, E. H., & Lee, Y. H. (2024). Living longer but in poor health: Healthcare system responses to ageing populations in industrialised countries based on the findings from the global burden of disease study 2019. BMC Public Health, 24(1). https://doi.org/10.1186/s12889-024-18049-0

2. Cook, D. (2012). Learning setting-generalized activity models for smart spaces. IEEE Intelligent Systems, 27(1), 32–38. https://doi.org/10.1109/mis.2010.112

3. Department of Statistics Malaysia (DOSM, 2016). Population Projection (Revised), Malaysia, 2010-2040. Retrieved July 7, 2024, from https://www.dosm.gov.my/portal-main/release-content/population-projection-revised-malaysia-2010-2040 .

4. Facchinetti, G., Petrucci, G., Albanesi, B., De Marinis, M. G., & Piredda, M. (2023). Can smart home technologies help older adults manage their chronic condition? A systematic literature review. International Journal of Environmental Research and Public Health, 20(2), 1205. https://doi.org/10.3390/ijerph20021205

5. Lai, W.-H., Hon, Y.-K., Pang, G. M.-H., Chong, E. M.-G., Nordin, N., Tiong, L.-L., Tan, S. H., Sapian, R. A., Lee, Y.-F., & Rosli, N. (2022a). Dementia of the ageing population in Malaysia: A scoping review of published research. Aging and Health Research, 2(2), 100077. https://doi.org/10.1016/j.ahr.2022.100077

6. Nižetić, S., Šolić, P., López-de-Ipiña González-de-Artaza, D., & Patrono, L. (2020). Internet of things (IOT): Opportunities, issues and challenges towards a smart and sustainable future. Journal of Cleaner Production, 274, 122877. https://doi.org/10.1016/j.jclepro.2020.122877

7. United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP, 2023). "Malaysia | Demographic Changes." UNESCAP, Retrieved July 7, 2024 from https://www.population-trends-asiapacific.org/data/MYS .

8. World Health Organization. (2020). Supporting people living with dementia in Malaysia. Retrieved July 23, 2024, from https://www.who.int/westernpacific/news-room/feature-stories/item/support