Balancing Accuracy and Energy: the Impact of Window Size on Sensor-based Human Activity Recognition

Chiara Contoli[1 [0000−0003−2389−2593]], Susanna Peretti[1 [0009−0000−4162−4301]], and Emanuele Lattanzi[1 [0000−0002−6568−8470]]

University of Urbino, Piazza della Repubblica 13, 61029 Urbino, Italy
{chiara.contoli,emanuele.lattanzi}@uniurb.it
s.peretti@campus.uniurb.it

Abstract. In wearable computing, data segmentation based on a sliding window approach is common, and the window segment size is crucial in determining activity recognition performances. The existing literature is indeed focused on investigating how the window size impacts accuracy neglecting, however, the impact on the energy consumption of low-power devices employed to perform the recognition task. We have performed an experimental analysis of the impact of the window size, coupled with feature selection, on the energy consumption of the ESP32 device. This paper describes how those two critical aspects affect performance evaluation. We consider three public datasets to provide useful insights on the best trade-off between accuracy and energy consumption. Results show that the best performance is usually obtained with windows longer than 2.56 seconds and features ranging from 10 to 20.

Keywords: Human Activity Recognition · Sliding Window Segmentation · Feature Selection · Energy Efficiency · Wearables · Constrained Devices

1 Introduction

Human Activity Recognition (HAR) using inertial sensors has emerged as a promising field with applications in various domains, such as healthcare, sports, and human-computer interaction. Accurate and efficient HAR systems are crucial for these applications, and energy consumption is a critical factor, especially for battery-powered wearable devices.

One of the critical phases that significantly impacts the energy efficiency and the classification accuracy of HAR systems is sensor signal segmentation. Techniques that accomplish this task have been surveyed in [7] and [11], and the sliding window approach is the most widely adopted because of its simplicity in implementation. The key parameter of this approach is the window size, which determines the length of the data segments used for feature extraction and classification. Understanding the best trade-off between accuracy and energy efficiency is not trivial. A smaller window size can reduce the classification phase's computational complexity and energy consumption. Still, it may compromise recognition accuracy, as shorter segments may not allow enough feature selection to capture information about (complex) activities. Conversely, a larger window size can improve accuracy but increase energy consumption due to increased computational overhead [4]. At the same time, it may compromise recognition accuracy as larger window may overlap with multiple activities. On the other hand, in a real-time scenario, when the recognition must be performed continuously, shorter windows entail frequent feature extraction and classification with an increase in energy consumption [9].

Furthermore, the objective of feature selection is to eliminate irrelevant or redundant features based on their predictive performance, thereby achieving an optimal balance between feature abundance and classification accuracy. Changing the window size impacts both classification accuracy and feature selection, which, in time, affects the device's power consumption. For this reason, balancing energy and accuracy, acting on both the set of features selected and the size of the window, takes the form of a non-trivial bi-objective optimization problem.

This paper extends our previous work [14, 15] by investigating how window size impacts the number and type of selected features, determining the trade-off between energy efficiency and classification accuracy in inertial sensor-based HAR systems. In particular, we aim to identify the optimal window size that balances energy consumption and recognition accuracy. To achieve this, we conduct a comprehensive study, varying: i) the window size across a wide range; ii) the number of selected features per window. We then evaluate the resulting energy consumption and classification performance by considering three different public datasets and the ESP32 device. The experimental results show that shorter windows and a higher number of features significantly increase energy demands due to frequent feature evaluations and higher computational costs. Moreover, considering the classification accuracy, the results reveal that features higher than 15, combined with sufficiently large window sizes, achieve performance close to the maximum. The best performance is usually obtained with windows longer than 2.56 seconds and features ranging from 10 to 20.

2 Related Work

Window sizing, or sliding window segmentation, has been previously studied in Human Activity Recognition. Fixed sliding windows (FSW) have been one of the most investigated approaches to address the problem of segmenting data sensor signals. According to FSW, sensor data are split into fixed time windows, and the window size moves over the time series to find and extract relevant information to recognize the activity.

Niazi *et al.* considered a set of varying window sizes (1, 2, 3, 5, and 10 seconds) coupled with varying sampling rates to investigate how they impact the accuracy of a trained classifier by considering data collected from a hip-worn accelerometer [12]. Similarly, Yamansavascilar and Guvensan investigated the impact of window size (4, 3, 2, and 1 seconds) and sampling rates on the accuracy by considering data collected from a smartphone [21]. Another study that investigated the impact of the window size (32, 64, and 128 seconds) on the accuracy by considering data collected from a smartphone is the one of Bashir *et al.* [5]. Identifying the window size that gives the optimum accuracy in HAR has also been investigated by Nurwulan and Jiang [13], who divided into equal-sized small sequences the wearable sensor signals considering sizes of (0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, and 3 seconds). More recently, Mekruksavanich *et al.* investigated how varying sliding window widths (5, 10, 20, 30, and 40 seconds) impact the accuracy of sensor-based HAR utilizing wristwatch sensors and deep learning methods [10].

It is, therefore, evident that many window sizes, ranging from 0.1 to 128 seconds, have been investigated and that there is no window size fitting and satisfying all requirements. Moreover, many other factors, such as sensor modality, diverse datasets, and recognition models, also affect such an impact [22]. However, all these works only focus on studying the effects of the window size on the recognition accuracy, thus neglecting the impact on the energy consumption. Only a few papers, e.g., [1] by Agac *et al.*, consider the effect of window size on energy consumption. Indeed, Agac *et al.* consider four different window sizes (5, 10, 20, and 30 seconds), different sampling rates, accelerometer and gyroscope, and only time-domain features to evaluate the accuracy and resource consumption for complex activity recognition, such as smoking, with smartwatches. In particular, to measure the energy usage, the authors run the experiments on a Microsoft Windows computer and use the WMI tool, which gives information about the battery, to measure the remaining capacity (in milliwatt hours) before and after the execution of their proposed algorithm.

In our work, we consider eight different window sizes, an increasing range of selected features per window (from 5 to 50), and one sampling rate, and we consider the accelerometer data only. However, compared to the existing literature, particularly to [1], we consider both time-domain and frequency-domain features, such as the fast Fourier transform (FFT). The choice of our window sizes is indeed partly motivated by the use of the FFT, as explained in the following section. We deployed our HAR application on a low-power ESP32 device to measure the energy consumption. Although we agree with [1] that time-domain features are more straightforward to compute, we believe that investigating the impact of frequency-domain features on recognition accuracy and energy consumption is a valuable contribution to identifying key parameters that influence the best balance between accuracy and energy. At the same time, the choice of a low-power device for energy measurement is motivated by the need to shed some light on the exact resource consumption values.

3 Methodology

In this section, we describe our proposed workflow to characterize the impact of window size, coupled with feature selection, with respect to recognition accuracy and energy consumption. The workflow includes different steps that will be described in subsection 3.2.

3.1 Preliminary Discussion

The sensor-based HAR process using shallow machine learning typically consists of four phases: i) data collection, ii) data pre-processing, iii) feature engineering, and iv) classification. Feature engineering involves two key steps: feature extraction and feature selection.

Feature extraction focuses on analyzing raw signals in the time, frequency, and time-frequency domains to derive distinctive features. On the other hand, feature selection aims to eliminate irrelevant or redundant features, thereby narrowing down to a subset of meaningful ones. The most commonly used methods for this step are filter-based, wrapper-based, and embedded approaches. Filter methods evaluate feature importance based on the training data's characteristics and operate independently of the learning algorithm. Wrapper methods iteratively search for the optimal subset of features based on their predictive performance using a specific learning algorithm, continuing until a stopping criterion is reached. Embedded methods combine the benefits of both by integrating feature selection into the model training process.

This work explores how changing the window size on top of which to extract features influences the number and the type of selected features, which, in time, affects the device's power consumption. This takes the form of a bi-objective optimization problem where one wants to maximize the accuracy of the classifier while minimizing the energy consumed by acting on both the set of features selected and the size of the window. It thus results in a non-trivial solvable system since the space of independent variables becomes quadratic, and, at the same time, the relationship between window size, number of selected features, and energy spent will be strongly nonlinear. For instance, reducing the size of the windows will reduce the time needed to compute each feature, consequently reducing energy consumption. On the other hand, for a continuous monitoring application, smaller window sizes result in more frequent feature computation, increasing energy expenditure.

Similarly, the relationship between window size and the composition of the best feature set for the classification task is far from linear. A particular set of features can optimally describe a human activity if computed on a specific window size, whereas it might fail with different sizes. In this case, the feature selection algorithm would select a distinct set of features, affecting the energy consumed due to different computational complexity. For instance, the accelerometer's mean and standard deviation can help discriminate between *walking* and *running* if computed over a window containing a few steps, as they can capture differences in signal magnitude. While computing these over a long-lasting window can incur a dilution problem of the peculiarities of each activity. Some time-domain features, however, might be more representative when computed over larger windows. For example, the dominant frequency can easily discriminate between running and walking but requires several step repetitions to be determined rigorously.

The following subsection describes the methodology proposed to empirically evaluate the relationships between the independent variables and the energy consumption and accuracy to guide the researcher in identifying the best trade-off solution.

3.2 Windowing and Features Selection

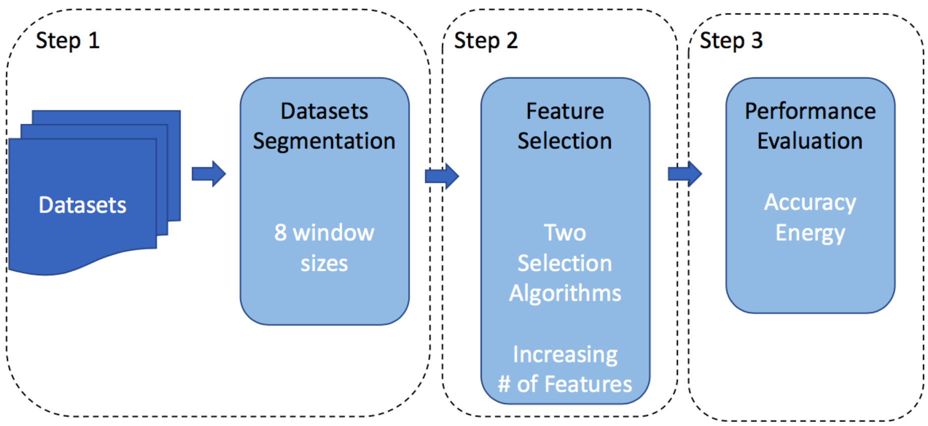

Fig. 1: Proposed methodology for accuracy and energy trade-off evaluation.

Figure 1 shows the general proposed workflow. During Step 1, we segment the datasets according to each of the eight considered window sizes. In Step 2, we perform feature selection for each dataset and window size using two different methods. Features are selected by varying the number from a minimum of five to a maximum of fifty. Then, in Step 3, we explore the bi-dimensional parameter space by evaluating each window's accuracy and energy coupled with each number of selected features.

In this work, we consider a sampling rate of 100 Hz. Although we know that the sampling frequency impacts recognition accuracy and energy consumption, we decided to investigate primarily the impact of the window size while varying the selected features.

As window dimensions, we consider different sizes: 0.16, 0.32, 0.64, 1.28, 2.56, 5.12, 10.24, and 20.48 seconds, corresponding to the following number of samples: 16, 32, 64, 128, 256, 512, 1024, 2048. The reason for such a choice is twofold: on the one hand, the activities contained in the HAR datasets (described in subsection 4.1) span from simple static and dynamic locomotion activities (such as standing and walking) to more complex activities (such as washing plate or reading book). Therefore, a broad range of window sizes, from short to wide, is helpful in capturing the nature of the underlying dynamics (e.g., more complex activities typically require wider windows). On the other hand, we use the FFT as one of the considered frequency-domain features, which requires data of power of 2 lengths for efficient computation.

Often, the accelerometer is enough to recognize activities of daily living [3, 17]. However, more complex activities could benefit from the adoption of the gyroscope. We consider adopting only data from a triaxial accelerometer, leaving the gyroscope for future investigation.

We extract features from both the time and frequency domain. This choice is motivated by the fact that, although time-domain features are computationally less expensive than frequency-domain ones, those letters are essential to capture more complex dynamics that may not be easily identifiable in time-domain analysis. We consider statistical indexes, such as mean and standard deviation, and other relevant metrics, such as root mean square, interquartile range, autoregressive coefficients, and signal magnitude area in the time domain. We consider the FFT and other metrics, such as peaks, energy skewness, and kurtosis in the frequency domain. It is worth noting that some features are selected both in the time and frequency domain. Table 1 reports all the features considered in this work.

Table 1: List of extracted features.

Features
Mean
Min
Max
Median
Standard Deviation
N Peaks
Peak-to-Peak Amplitude
Interquartile Range
Autocorrelation
Energy
Autoregressive coefficients
Signal Magnitude Area
Root Mean Square
FFT
Spectral Mean
Skewness
Kurtosis
Growth factor

We use Recursive Feature Elimination combined with logistic regression (LR-RFE) and the Select From Model method with RidgeCV regularization (SFM-RidgeCV) as a feature selection algorithm. LR-RFE is a wrapper-based method that gradually reduces the number of features by iteratively removing the least relevant ones. SFM-RidgeCV is an embedded-based method consisting of a set of RidgeCV models, each containing a different set of selected features and leveraging cross-validation performance metrics to choose the best feature subset, which is subsequently used to train the final model. For recognition accuracy evaluation, we use a Random Forest classifier.

4 Experimental Evaluation

In this section, we provide a thorough description of the experimental setup and the datasets; we present the results of the space parameter exploration and highlight the best parameter configuration.

4.1 Datasets

We use three different publicly available datasets, namely, Real World (HAR) 2016 [16, 18], WatchHAR [2, 19], and MobiAct [20, 6].

Real World (HAR) 2016 consists of 15 subjects (eight males and seven females, aged between 19 and 43, height between 167 cm and 179 cm, and weight between 61 kg and 87 kg). Each subject performs about 10 minutes of *climbing stairs down and up, lying, standing, sitting, running/jogging, and walking* activities, and about 1.7 minutes of jumping activity. Data are collected from six sensors (accelerometer, GPS, gyroscope, light, magnetometer, and sound) placed on the body in wearable devices that include all relevant device positions to recognize the above-mentioned activities.

watchHAR consists of 13 subjects (both males and females aged between 23 and 67). Subjects perform between 1 and 3 minutes of *brushing teeth, preparing sandwich, reading book, typing, using phone, using remote control, walking freely, walking holding a tray, walking with handbag, walking with hands in pockets, walking with object underarm, washing face and hands, washing mug, washing plate, and writing* activities. Data are collected from IMU recordings (3-axial acceleration, 3-axial gyroscope, and 3-axial magnetometer) with a Sony Smartwatch 3.

MobiAct consists of 57 subjects (42 males and 15 females, aged between 20 and 47, height between 160 cm and 189 cm, and weight between 50 kg and 120 kg). Fifty subjects perform activities of daily living (ADLs) such as *standing, walking, jogging, jumping, stairs up, stairs down, sit chair, car in, and car out.* In our work, we consider the extended version of the former MobiAct dataset, which includes 4 different types of falls and 12 different ADLs from a total of 66 subjects. It is worth mentioning that we did not consider the falls in our experiments. Data are collected from the accelerometer, gyroscope, and orientation sensors of a Samsung Galaxy S3 smartphone.

4.2 Setup

Feature selection algorithms, model training, and testing have been executed in a workstation featuring two Intel ® Xeon® Silver 4314, CPU 2400 MHz 1351632 processors, RAM 512GB, 16 DDR4 2666 MHz 32768 MB.

To characterize the energy impact of each window size, we developed a sensor-based HAR application, which is responsible for collecting data from a triaxial accelerometer. The software was compiled for the ESP32 platform, connected to an MPU6050 triaxial accelerometer [8].

For each window size, we measured both the energy consumption and the execution time of the features listed in Table 1. These measurements were obtained using a National Instruments NI-DAQmx PCI-6251, a 16-channel data acquisition board connected to the ESP32.

To ensure result reliability, we averaged the execution time over five measurements and used it to compute the energy consumed for each feature. The power consumption of each feature was then determined by dividing the energy consumption by the window size and the sampling frequency of the accelerometer.

4.3 Exploring the Space Parameters

In Figure 2, we report the heatmaps representing the misclassification rate (MCR) and the power consumption measured on the quadratic parameter space for each dataset. In particular, concerning the MCR (first column), more than 15 features associated with sufficiently large window sizes are needed to reach performance close to the maximum in all datasets. In the realWorld2016 (top row) and Watch_HAR (middle row) datasets, a window containing 512 samples (corresponding to 5.12 seconds) seems to be the best choice in terms of classification accuracy. On the other hand, the MobiAct (bottom row) results in higher performance when segmented with even longer windows, within 10 seconds or more. This result is probably due to the intrinsic characteristics of the MobiAct dataset, which contains two quasi-static activities (*standing* and *sitting on chair*) accounting for a significant part of the data. Indeed, static activities tend to be easily recognized compared to those involving dynamic movements using long windows.

Concerning the average power consumption (the second column in Figure 2), it is evident that very short windows and many features generate high power values. Indeed, more features entail more power spent on computation, and short windows lead to a more frequent feature evaluation. Although we recognize this common general trend, the values do not coincide precisely across the three datasets. This result is because, by setting the same number of features, we are likely to select the features that best classify the specific characteristics of each dataset, which will not necessarily be the same. As a consequence, different features may correspond to a different computational cost.

To evaluate the best configuration of the number of features and of the window size, we use the Pareto charts of Figure 3,4, and 5.

A Pareto chart is valuable for visualizing the relationship between different cost metrics and identifying the optimal trade-off points. In particular, in this case, the optimal points simultaneously minimize the misclassification rate (MCR) and the power consumption (i.e., the points closest to the origin of the axes). For both realWorld2016 and watch_HAR, a window size of 512 samples

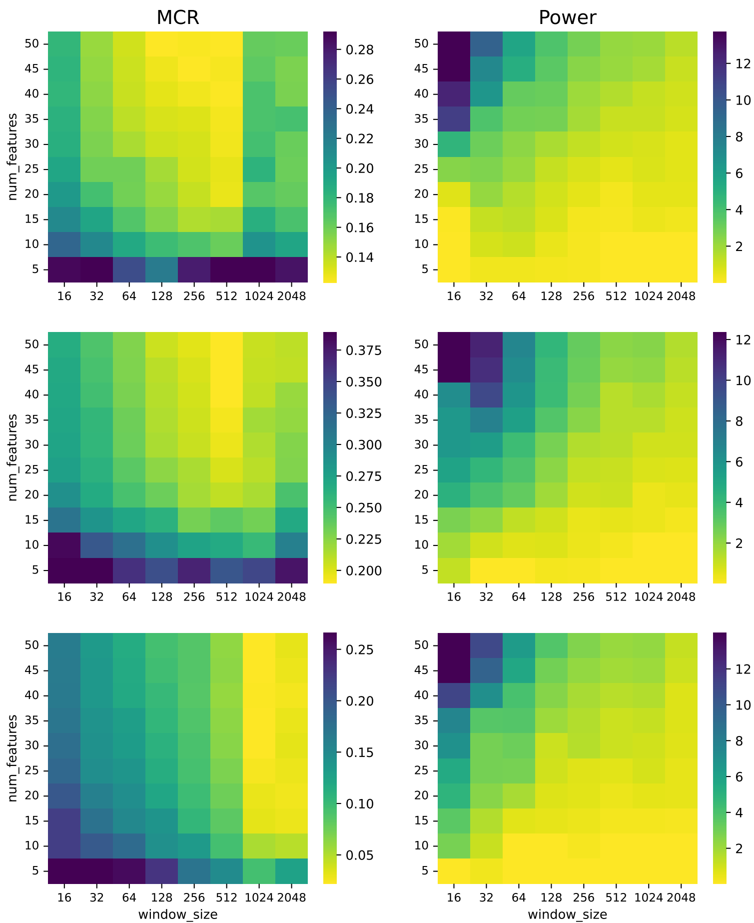

Fig. 2: Missclassification rate (MCR) and power consumption resulting from exploring the parameters space: realWorld2016 (top row), watch_HAR (middle row), and MobiAct (bottom row) dataset results, respectively.

(i.e., 5.12 seconds) and features between 10 and 20 identify the best point in space. On the other hand, concerning the MobiAct dataset, we have to choose a window size between 1024 and 2048 samples (i.e., between 10.24 and 20.48 seconds) to minimize power and classification errors.

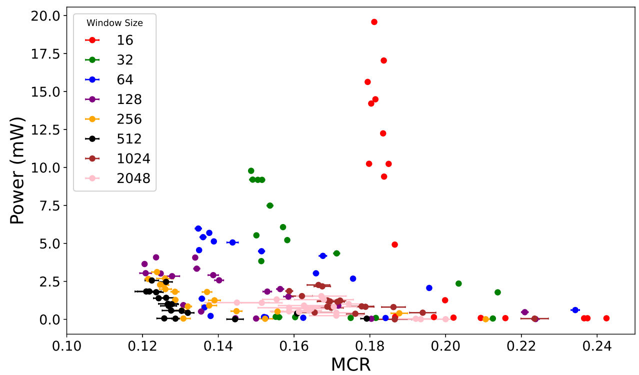

Fig. 3: Pareto charts comparing the (MCR) with the average power consumption for realWorld2016, when varying the number of features and the window size.

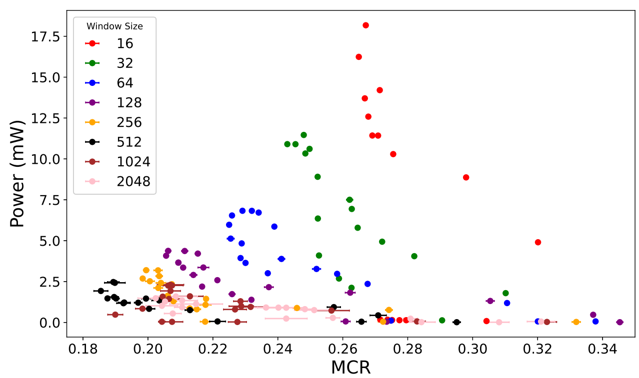

Fig. 4: Pareto charts comparing the (MCR) with the average power consumption for watch_HAR, when varying the number of features and the window size.

To summarize, the best HAR solution, representing an optimal trade-off between accuracy and power, should be designed considering that windows that are too short combined with a high number of features result in excessive energy consumption, which is not reflected in improved classification performance. Indeed, the best performance is usually obtained with windows longer than 2.56 seconds and features ranging from 10 to 20.

5 Conclusion

This study analyzed the interplay between classification performance and power consumption in Human Activity Recognition systems by exploring a quadratic parameter space defined by the window size and the number of features.

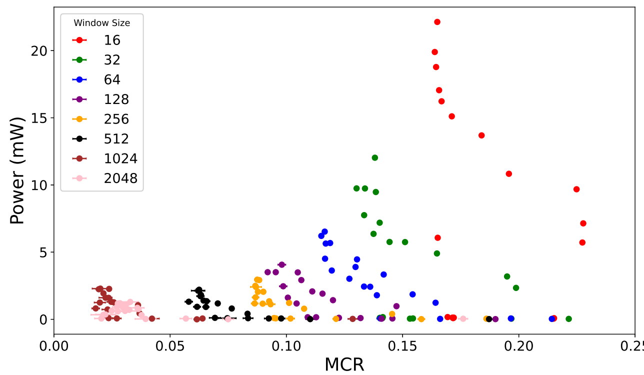

Fig. 5: Pareto charts comparing the (MCR) with the average power consumption for MobiAct, when varying the number of features and the window size.

Power consumption analysis revealed that shorter windows and a higher number of features significantly increase energy demands due to frequent feature evaluations and higher computational costs. Moreover, considering the classification accuracy, the results reveal that features higher than 15, combined with sufficiently large window sizes, achieve performance close to the maximum.

In summary, this study demonstrates that optimizing window size and feature count is crucial for balancing classification accuracy and power consumption in HAR systems and suggests that optimal configurations typically use windows longer than 2.56 seconds and 10–20 features.

Acknowledgment

This work was partially funded by MIMIT, under FSC project "Pesaro CTE SQUARE", CUP D74J22000930008.

References

1. Agac, S., Shoaib, M., Incel, O.D.: Context-aware and dynamically adaptable activity recognition with smart watches: A case study on smoking. Computers & Electrical Engineering **90**, 106949 (2021)
2. Alevizaki, A., Trigoni, N.: watchhar: A smartwatch imu dataset for activities of daily living (Sep 2022). https://doi.org/10.5281/zenodo.7092553, https://doi.org/10.5281/zenodo.7092553
3. Atallah, L., Lo, B., King, R., Yang, G.Z.: Sensor positioning for activity recognition using wearable accelerometers. IEEE transactions on biomedical circuits and systems **5**(4), 320–329 (2011)
4. Banos, O., Galvez, J.M., Damas, M., Pomares, H., Rojas, I.: Window size impact in human activity recognition. Sensors **14**(4), 6474–6499 (2014)

5. Bashir, S.A., Doolan, D.C., Petrovski, A.: The effect of window length on accuracy of smartphone-based activity recognition. IAENG International journal of computer science **43**(1) (2016)
6. Biomedical Informatics Laboratory, Hellenic Mediterranean University: The MobiFall and MobiAct datasets. https://bmi.hmu.gr/the-mobifall-and-mobiact-datasets-2/, accessed: January 03, 2025
7. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR) **46**(3), 1–33 (2014)
8. InvenSense Inc.: Mpu-6050 product specification (2023), https://invensense.tdk.com/products/motion-tracking/6-axis/mpu-6050/, last accessed 2025-01-05
9. Lee, J., Kim, J.: Energy-efficient real-time human activity recognition on smart mobile devices. Mobile Information Systems **2016**(1), 2316757 (2016)
10. Mekruksavanich, S., Jantawong, P., Jitpattanakul, A.: Effect of sliding window sizes on sensor-based human activity recognition using smartwatch sensors and deep learning approaches. In: 2024 5th International Conference on Big Data Analytics and Practices (IBDAP). pp. 124–129. IEEE (2024)
11. Najeh, H., Lohr, C., Leduc, B.: Dynamic segmentation of sensor events for real-time human activity recognition in a smart home context. Sensors **22**(14), 5458 (2022)
12. Niazi, A.H., Yazdansepas, D., Gay, J.L., Maier, F.W., Ramaswamy, L., Rasheed, K., Buman, M.P.: Statistical analysis of window sizes and sampling rates in human activity recognition. In: HEALTHINF. pp. 319–325 (2017)
13. Nurwulan, N.R., Jiang, B.C.: Window selection impact in human activity recognition. International Journal of Innovative Technology and Interdisciplinary Sciences **3**(1), 381–394 (2020)
14. Peretti, S., Contoli, C., Lattanzi, E.: Complexity-aware features selection for wrist-worn human activity recognition. In: 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). pp. 799–804. IEEE (2024)
15. Peretti, S., Contoli, C., Lattanzi, E., et al.: An experimental study on the energy efficiency of feature selection for human activity recognition with wrist-worn devices. In: Proceedings of International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence. pp. 1–15. Springer (2024)
16. Sztyler, T., Stuckenschmidt, H.: On-body localization of wearable devices: An investigation of position-aware activity recognition. In: 2016 IEEE International Conference on Pervasive Computing and Communications (PerCom). pp. 1–9 (2016). https://doi.org/10.1109/PERCOM.2016.7456521
17. Twomey, N., Diethe, T., Fafoutis, X., Elsts, A., McConville, R., Flach, P., Craddock, I.: A comprehensive study of activity recognition using accelerometers. In: Informatics. vol. 5, p. 27. MDPI (2018)
18. University of Mannheim: DataSet – RealWorld (HAR) (2016). https://www.uni-mannheim.de/dws/research/projects/activity-recognition/dataset/dataset-realworld/ (2016), accessed: December 30, 2024
19. University of Oxford: watchHAR: A Smartwatch IMU dataset for Activities of Daily Living. https://zenodo.org/records/7092553 (2022), accessed: December 30, 2024
20. Vavoulas, G., Chatzaki, C., Malliotakis, T., Pediaditis, M., Tsiknakis, M.: The mobiact dataset: Recognition of activities of daily living using smartphones. In: International conference on information and communication technologies for ageing well and e-health. vol. 2, pp. 143–151. SciTePress (2016)

21. Yamansavascilar, B., Guvensan, A.: M.: Activity recognition on smartphones: efficient sampling rates and window sizes, 1–6 (2016) (2016)
22. Yi, M.K., Lee, W.K., Hwang, S.O.: A human activity recognition method based on lightweight feature extraction combined with pruned and quantized cnn for wearable device. IEEE Transactions on Consumer Electronics **69**(3), 657–670 (2023)