Human Activity Segmentation Challenge @ ECML/PKDD'23

Arik Ermshaus[1 [0000−0002−8138−3060]], Patrick Schäfer[1], Anthony Bagnall[2], Thomas Guyet[3], Georgiana Ifrim[4], Vincent Lemaire[5], Ulf Leser[1], Colin Leverger[6], and Simon Malinowski[7]

[1] Humboldt University of Berlin, Berlin, Germany
{ermshaua,patrick.schaefer,leser}@informatik.hu-berlin.de
[2] University of Southampton, Southampton, UK a.j.bagnall@soton.ac.uk
[3] Inria, Villeurbanne, France thomas.guyet@inria.fr
[4] University College Dublin, Dublin, Ireland georgiana.ifrim@ucd.ie
[5] Orange Labs, Lannion, France vincent.lemaire@orange.com
[6] Orange Innovation, Rennes, France colin.leverger@orange.com
[7] University of Rennes 1, Rennes, France simon.malinowski@irisa.fr

Abstract. Time series segmentation (TSS) is a research problem that focuses on dividing long multivariate sensor data into smaller, homogeneous subsequences. This task is critical for various real-world data analysis applications, such as energy consumption monitoring, climate change assessment, and human activity recognition (HAR). Despite its importance, existing methods demonstrate limited efficacy on real-world multivariate time series data. To advance the field, we organized the Human Activity Segmentation Challenge at ECML/PKDD and AALTD 2023, featuring 57 participants. Collaborating with 15 bachelor computer science students, we gathered and annotated 10.7 hours of real-world human motion sensor data. The challenge required participants to segment the resulting 250 multivariate time series into an unknown number of variable-sized activities. The top-8 approaches outperformed existing baselines, but show only limited improvements, capped at 1.9 percentage points. The segmentation of real-world mobile sensing recordings remains challenging. We release the labelled challenge data for future research.

Keywords: Ubiquitous Sensing · Human Activity Recognition · Data Mining · Unsupervised Learning · Time Series Segmentation

1 Introduction

The analysis of human behaviour can provide valuable insights into health status, fitness, or personal security [1]. This is relevant to various domains, including the medical sector [2], industrial applications [3], and military operations [4]. Wearable devices, such as smartphones, have low-cost sensors that capture the dynamics of human activities in the form of long consecutive segments within temporal data, commonly known as time series (TS) [5]. Such data can be used, for instance, to detect falls in the elderly [6], or to monitor patients with dementia or mental illness [7].

To accomplish such applications, the research field of human activity recognition (HAR) implements workflows that first segment TS motion data, then learn characteristic features and finally classify individual activities [1]. Most HAR systems process fixed-length subsequences extracted from sensor measurements, as opposed to processing the entirety of a single activity [8]. This leads to heterogeneity and performance losses in many downstream tasks [7]. The automatic partitioning of multivariate sensor signals into an unknown amount of variable-sized activity

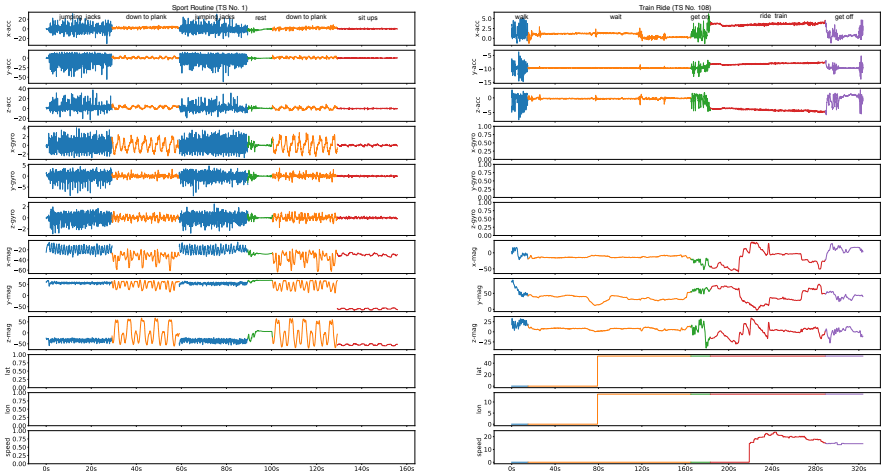

Fig. 1: Example TS of a sport routine from subject 1 (left, indoor) and a train ride from subject 1 (right, outdoor). Activity segments are coloured. Missing dimensions are displayed as empty cells.

segments is very challenging, and many open problems still exist, such as accurately locating activity transitions in multi-dimensional data and deciding if these are actually substantial or just emergent signal fluctuations.

The overarching task of activity segmentation is called time series segmentation (TSS), which is an unsupervised learning problem that seeks to discover variable-sized, distinguishable segments separated by change points (CPs) within TS [9,10]. TSS typically is not the final aim of data analysis, but serves as a preprocessing step to partition complex TS data for advanced analytics such as classification [11], anomaly detection [12] or motif discovery [13]. Accurate solutions need to be robust, segment a wide variety of different TS and handle imperfect and noisy multi-dimensional sensor recordings from different devices. Recently, specialized statistical methods [10] and modern data mining algorithms [14,15] have been employed to address this task. However, as highlighted by the survey of Aminikhanghahi et al. [9], accuracy is still limited.

To bridge this gap, we conducted an ECML/PKDD 2023 discovery challenge in collaboration with the 8th Workshop on Advanced Analytics and Learning on Temporal Data (AALTD@ECML)[8]. The competition aimed to increase the performance of multi-modal human activity segmentation and featured 57 participants. We provided a new mobile sensing data set from a daily setting, as opposed to the typical laboratory setup with intrusive and specialized sensor devices [16,17]. We collected and annotated 10.7 hours of multi-dimensional real-world TS data using heterogenous smartphone sensors capturing 100 typical human activities performed by 15 bachelor students in 6 motion sequences. See Figure 1 for two examples. The challenge task was to predict the amount and locations of activity transitions in the resulting 250 multivariate TS without any training or external data. Existing algorithms which served as baselines like BinSeg [10] or ClaSP [15] score low to medium F_1 scores (24,8% to 49,6%) on these data sets. The winning solutions improve the state of the art by up to 1.9 percentage points (pp) to 51.5%. This progress demonstrates the potential for further advancements in multivariate

[8] `https://ecml-aaltd.github.io/aaltd2023`

Motion ID	Group ID	Category	Activity Examples	Subject IDs
1	1	sport	jumping jacks, sit ups, plank, …	1,2,3,4,6,7,8
2	1	household	clear dishes, vacuum living room, push couch back, …	1,2,3,4,6,7,8
3	1	shopping	stand on escalator 1, change shoes, walk to Deichmann exit, …	1,3,4,5,6,7,8
1	2	commute	climb stairs, ride train (standing), wait for traffic lights, …	1,2,3,4,5,6,7
2	2	commute	go down stairs, wait, drive, …	1,2,3,4,5,6
3	2	sport	deep squat with arm reach, reverse plank hold, side stretch left, …	1,2,4,5,6,7

Table 1: List of motion sequences.

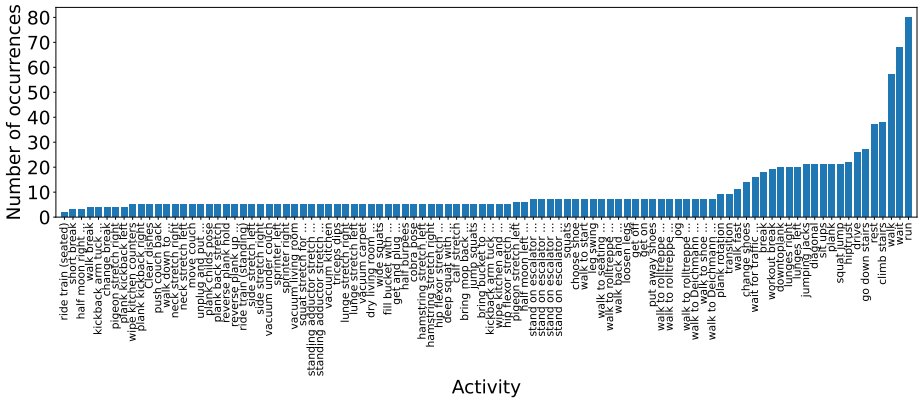

Fig. 2: Number of occurrences for single activities in the challenge data.

TSS research in general and HAR in particular. We make the labelled challenge data freely available [18] to encourage comparative evaluations in the field.

2 Challenge Data

In collaboration with 15 bachelor computer science students (see Section 5), we created a multimodal data set comprising 40 twelve-dimensional multivariate smartphone sensor recordings. These capture 6 distinct human motion sequences designed to represent pervasive behaviour in realistic indoor and outdoor settings. Data were collected using built-in smartphone sensors placed in the subjects' front right trouser pockets. We annotated the activities performed and their transitions in the recordings, resampled the data at a constant rate of 50 Hz, and segmented it to yield 250 multivariate TS. This data set serves as a benchmark for evaluating machine learning workflows.

The subsequent subsections provide detailed information on the data set's design (Subsection 2.1), collection process (Subsection 2.2), annotation and preprocessing workflow (Subsection 2.3), specifications (Subsection 2.4), and availability (Subsection 2.5).

Subject ID	Group ID	Gender	Age	Size (in cm)	Weight (in kg)
1	1	M	25	180	74
2	1	F	23	155	50
3	1	M	23	179	83
4	1	M	24	167	68
5	1	F	26	166	67
6	1	F	22	180	65
7	1	F	23	170	58
8	1	F	30	172	57
1	2	M	29	183	96
2	2	M	23	183	65
3	2	M	24	182	130
4	2	M	31	180	100
5	2	M	42	171	62
6	2	M	21	186	66
7	2	M	27	186	75

Table 2: List and characteristics of participants.

2.1 Data Set Design

Two independent groups from Humboldt-Universität zu Berlin, each consisting of either 8 or 7 bachelor computer science students, recorded 3 motion sequences in 2022. These sequences covered a total of 100 activities, with the first group focusing on indoor activities and the second group targeting outdoor behaviours. The primary objective was to capture natural human behaviour. A summary of the motion sequences is provided in Table 1, and specific activity annotations are linked to individual TS.

The student cohort included 10 males and 5 females, ranging in age from 21 to 42. Further details are presented in Table 2. Within a group, each student performed up to 3 preconceived motion routines, consisting of different and partly recurring activities, the distribution of which is visualized in Figure 2. The data collection yielded 40 multivariate recordings that were subsequently cut to create a data set of 250 multi-dimensional TS. Recordings were made using 5 different smartphones from 4 brands (Huawei, Motorola, Samsung, and Xperia) and were placed in the front right trouser pocket of (almost) all participants. The ”Physics Toolbox Sensor Suite” application was employed to capture sensor data from a triaxial accelerometer, gyroscope, and magnetometer, as well as latitude, longitude, and speed when available. The resulting TS feature 12 dimensions, with 9 filled sensor data dimensions and 3 empty ones, as illustrated in Figure 1. The empty dimensions are due to different sensors in the smartphones. To ensure continuous recording and prevent data loss in standby mode, the ”Touch Protector” application was also used. This smartphone placement and sensor configuration is consistent with common practices in human activity recognition (HAR) research [19,20]. Ground truth behaviour was captured through additional recording using another smartphone or action camera.

2.2 Data Collection

The student groups conducted the data collection over several days in the fall and winter of 2022, with tasks, roles, and responsibilities delegated among smaller teams. Prior to recording,

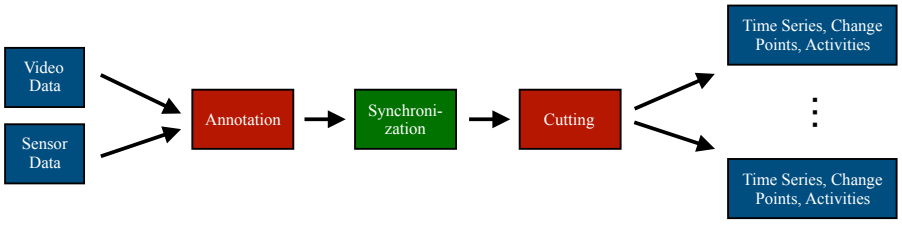

Fig. 3: The preprocessing workflow for a single recording.

instructors briefed the subjects on the motion sequences and time commitments involved. During data collection, participants initiated sensor recording, placed the phone in their front right pocket, performed the specified motions, and then ceased recording upon completion. Additional students guided subjects through the correct motion sequences and filmed the activities to be used for annotations. All recordings were subsequently reviewed for data quality and securely stored.

The data collection process encountered several challenges. Both groups experienced data loss due to hardware failure, necessitating re-recordings. We only used uninterrupted TS in the challenge data. Organizational difficulties also arose due to illnesses among team members. In one case, a phone had to be taped onto a subject's pants due to a lack of pockets.

2.3 Preprocessing

A basic preprocessing pipeline was applied to the challenge data, as illustrated in Figure 3. Student groups annotated the recordings with activity labels and transitions to establish a ground truth, which is used for evaluation (Subsection 3.3). They manually analysed the video footage in conjunction with the sensor data to do so. Subsequently, we synchronized the sensor dimensions using linear interpolation, a prerequisite for most TS analysis algorithms. A constant sample rate of 50 Hz was employed, deemed appropriate for human behaviour detection [21]. Finally, each of the 40 recordings was cut at randomly selected activity transitions to yield a data set of 250 multivariate TS, encompassing diverse problem settings.

2.4 Data Set Overview

The data set comprises 250 twelve-dimensional TS, capturing 15 participants performing up to three motion sequences each. The TS range from 7 seconds to 14 minutes in duration (median 100 seconds) and contain between 1 and 15 segments; 76% of TS have 5 or fewer segments (see Figure 4, top right). Activity durations vary from half a second for waiting to 10 minutes for running, with generally small variances between subjects and individual executions (see Figure 4, bottom).

Figure 1 presents two example TS from the challenge data, displaying all 12 dimensions. Activities are represented as coloured segments, and missing dimensions appear as empty cells. The sports routine (left) reveals abrupt transitions between activities, while the train ride sequence (right) shows gradual transitions and significant variations in activity duration.

2.5 Data Availability

We make all the challenge data publicly available, complete with labels, meta-information, and a Python data loader, on our website [18]. The data is licensed under CC-BY-NC-SA, allowing

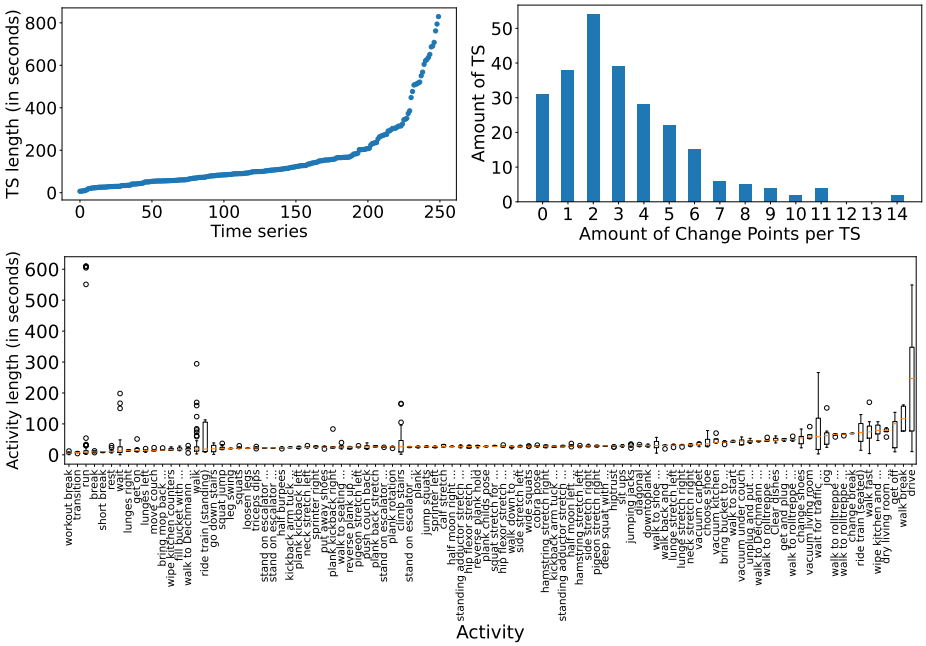

Fig. 4: Top: TS length and amount of change points. Bottom: Single activities and their lengths.

users to share and adapt the content, provided they give appropriate credit, use the same licence, and refrain from commercial use.

3 Challenge Organisation

We organized the contest as a discovery challenge for ECML/PKDD 2023, in collaboration with the 8th Workshop on Advanced Analytics and Learning on Temporal Data (AALTD@ECML). The competition ran from April 11 to June 11, 2023, concluding at 23:59 UTC. Registration was open until June 2, 2023. Following the competition's end, we requested the top-ranking solutions from competitors and, after a final review, released the scores on June 16, 2023. In total, the challenge attracted 57 registrations, with 17 active participants submitting 240 entries. Two winners were awarded free tickets to ECML/PKDD 2023, oral presentations of their approaches at both the conference and the workshop, as well as publications in its proceedings.

Subsequent subsections will detail the technical aspects (Subsection 3.1), rules (Subsection 3.2), evaluation measure (Subsection 3.3), and competition results (Subsection 3.4).

3.1 Technical Details

We hosted an invite-only community competition on Kaggle [9] to disseminate challenge information, data, baselines, and to maintain public and private leaderboards. Interested individuals could

[9] `https://www.kaggle.com/competitions/human-activity-segmentation-challenge`

access the competition through an invitation link, provided upon request via a questionnaire. We supplied Jupyter notebooks featuring an exploratory data analysis and six state-of-the-art algorithms for TSS [18], including BinSeg [10], ClaSP [15], FLUSS [14], GGS [22], IGTS [23], and STRAY [24]. Participants submitted their predictions as a CSV file, containing predicted activity transitions for each of the 250 TS. These submissions were automatically scored and ranked by the Kaggle platform.

3.2 Rules

Participants had to adhere to specific rules to join the challenge. Each participant was allowed up to three daily submissions, using only reproducible and deterministic methods subject to verification on request by the organizers to prevent cheating. We deemed a solution deceitful if it relied on manually labelled annotations or machine learning algorithms that used such annotations. Only fully unsupervised solutions were permitted to ensure a fair competition. Additionally, the use of external data or metadata-based manual tuning of hyperparameters was prohibited. Parameters had to be either universally set or data-driven.

The top-3 competitors were required to submit their code for a final inspection and hand-in a report that describes their approach. Failure to fulfil these obligations resulted in forfeiture of the award and winning status, as was the case for one participant. The challenge organizers were ineligible to submit entries.

3.3 Evaluation Measure

In this challenge, participants were tasked with predicting the offsets of activity transitions for all 250 twelve-dimensional TS in our data set. Apart from the TS, sensor names, and overall sample rate, no further information was provided. Ground truth annotations, kept confidential, served as the basis for evaluating the predicted segmentations. To score submissions and generate leaderboards, the data set was randomly partitioned into public and private sets, each containing 125 TS. No stratified sampling was applied, as TSS had to be performed for single TS without training or external data. Final performance was assessed on the private set, yielding the ultimate leaderboard and rankings.

For evaluating segmentation performance, we employed a well-established benchmarking score from existing literature. Drawing inspiration from an image segmentation challenge [10], we combined classification and clustering metrics to calculate the average F_1 score across different thresholds. Specifically, for each TS, we calculated the intersection over union (IoU) between predicted and ground truth segments to yield a normalized score (higher is better). A threshold was then applied to determine sufficient overlaps, generating a confusion matrix from which the F_1 score was computed. This process was iteratively applied for multiple thresholds (ranging from 0.5 to 0.95 in steps of 0.05), and the results were averaged to generate the final normalized score. This measure was calculated for each of the 250 TS and averaged per leaderboard to measure the quality of a participant and infer the public and private rankings.

3.4 Competition Results

Table 3 displays the final rankings and F_1 scores for the top-10 competitors. The top-2 winning solutions achieved F_1 scores exceeding 50%. Both utilized the ClaSP algorithm in a multivariate

[10] `https://www.kaggle.com/competitions/airbus-ship-detection/overview/evaluation`

Rank	F_1 Score (in %)	Participant	No. of Entries
1	51.5	gh	46
2	50.7	Koular	12
3	49.8	Panos	14
4	49.8	infoxin	15
5	49.8	kojimar	7
6	49.8	Shayekh Islam	4
7	49.8	fuge	5
8	49.8	laffrent	11
	49.6	***ClaSP***	
9	49.6	pjmathematician	16
10	49.1	ALLAccept	11
	24.8	***BinSeg***	
	23.9	***FLUSS***	

Table 3: The final private leaderboard with top-10 best-ranking competitors and 3 baselines (in bold / italic). The top-8 approaches outperform the best baseline ClaSP.

setting, employing strategies for selecting relevant sensor dimensions, hyper-parameter tuning, and change point merging. Their detailed methodologies and code are available in the respective publications [25,26]. The top-8 competitors outperformed the highest-ranking baseline, ClaSP, which scored 49.6%. However, the performance improvement, capped at 1.9 pp, highlights the inherent challenge of segmenting real-world mobile sensing data in a fully unsupervised manner.

4 Conclusion

We presented an overview and results of the Human Activity Segmentation Challenge at ECML/PKDD and AALTD 2023. The contest utilized 10.7 hours of mobile sensing data recorded with 15 bachelor students, which is now publicly available for future human activity recognition research. In the challenge, 17 active participants competed, with the top-2 achieving F_1 scores over 50% for the segmentation task. However, the overall performance on this data set remains limited and requires significant improvement for TSS to be a viable component in human activity recognition workflows.

Based on this challenge, we identify several avenues for future research: (a) exploring sensor fusion within multivariate TSS, as opposed to current methods that segment TS channels independently and merge resulting change points; (b) investigating dimension selection for multivariate TSS to potentially improve accuracy; and (c) advancing domain-specific data denoising, normalization, and preprocessing, particularly to facilitate the segmentation process.

5 Acknowledgements

We would like to thank Jonas Albrecht, Alexandria Arnold, Leo Baur, Malte Borgmann, Simon Bosse, Sinan Genc, Alina Hartwich, Isabel Heise, Jan Evert Hinrichs, Hoai Ngoc Ho, Malte Hückelkempkes, Wei Jin, Elida Sengül, Gerrit Slomma and Katharina Winde for their work in creating, recording and annotating the motion sequences for the challenge data.

References

1. O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 1192–1209, 2013.
2. L. Zhou, E. Fischer, C. M. Brahms, U. Granacher, and B. Arnrich, "Duo-gait: A gait dataset for walking under dual-task and fatigue conditions with inertial measurement units," *Scientific Data*, vol. 10, 2023.
3. N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: a narrative review," *Artificial Intelligence Review*, vol. 55, pp. 4755 – 4808, 2022.
4. A. Mukherjee, S. Misra, P. Mangrulkar, M. Rajarajan, and Y. Rahulamathavan, "Smartarm: A smartphone-based group activity recognition and monitoring scheme for military applications," *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–6, 2017.
5. A. Ermshaus, S. Singh, and U. Leser, "Time series segmentation applied to a new data set for mobile sensing of human activities," in *EDBT/ICDT Workshops*, 2023.
6. J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 1082–1090, 2008.
7. M. A. R. Ahad, A. D. Antar, and M. Ahmed, "Iot sensor-based activity recognition - human activity recognition," in *Intelligent Systems Reference Library*, 2021.
8. O. Baños, J. M. Gálvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, pp. 6474 – 6499, 2014.
9. S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, pp. 339–367, 2017.
10. C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020.
11. M. Middlehurst, P. Schafer, and A. Bagnall, "Bake off redux: a review and experimental evaluation of recent time series classification algorithms," *ArXiv*, vol. abs/2304.13029, 2023.
12. S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," *Proc. VLDB Endow.*, vol. 15, pp. 1779–1797, 2022.
13. P. Schäfer and U. Leser, "Motiflets - simple and accurate detection of motifs in time series," *Proc. VLDB Endow.*, vol. 16, pp. 725–737, 2022.
14. S. Gharghabi, C.-C. M. Yeh, Y. Ding, W. Ding, P. R. Hibbing, S. R. LaMunion, A. Kaplan, S. E. Crouter, and E. J. Keogh, "Domain agnostic online semantic segmentation for multi-dimensional time series," *Data Mining and Knowledge Discovery*, vol. 33, pp. 96 – 130, 2018.
15. A. Ermshaus, P. Schäfer, and U. Leser, "Clasp: parameter-free time series segmentation," *Data Mining and Knowledge Discovery*, vol. 37, pp. 1262 – 1300, 2023.
16. A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012.
17. R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. del R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, pp. 2033–2042, 2013.
18. Challenge Supporting Materials. `https://github.com/patrickzib/human_activity_segmentation_challenge`, 2023.
19. G. Bieber, J. Voskamp, and B. Urban, "Activity recognition for everyday life on mobile phones," in *Universal Access in Human-Computer Interaction*, 2009.
20. S. A. Elkader, M. Barlow, and E. Lakshika, "Wearable sensors for recognizing individuals undertaking daily activities," *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018.
21. O. Baños, C. Villalonga, R. García, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *BioMedical Engineering OnLine*, vol. 14, pp. S6 – S6, 2015.
22. D. Hallac, P. Nystrup, and S. P. Boyd, "Greedy gaussian segmentation of multivariate time series," *Advances in Data Analysis and Classification*, p. 727–751, 2019.

23. A. Sadri, Y. Ren, and F. D. Salim, "Information gain-based metric for recognizing transitions in human activities," *Pervasive Mob. Comput.*, vol. 38, pp. 92–109, 2017.
24. P. D. Talagala, R. J. Hyndman, and K. Smith-Miles, "Anomaly detection in high-dimensional data," *Journal of Computational and Graphical Statistics*, vol. 30, pp. 360 – 374, 2019.
25. G. Harańczyk, "Change points detection in multivariate signal applied to human activity segmentation," in *AALTD@ECML/PKDD*, 2023.
26. T.-J. Huang, Q.-L. Zhou, H.-J. Ye, and D.-C. Zhan, "Change point detection via synthetic signals," in *AALTD@ECML/PKDD*, 2023.