Data Extraction and Transformation Methodology for Biometric Signals from Wearable Devices

Patricia Concheiro-Moscoso, Iago Fernández-Garrido, Jerónimo Pardo-Rodríguez, María del Carmen Miranda-Duro, and Betania Groba

CITIC - TALIONIS Group, Elviña Campus, Faculty of Health Sciences, Universidade da Coruña (University of A Coruña), 15071 A Coruña, Spain.
Correspondence: iago.fgarrido@udc.es, j.pardor@udc.es

DOI: *https://doi.org/10.17979/spu.23.c37*

Abstract: The increasing use of consumer wearable devices, such as smart bands, provides new opportunities for health monitoring and clinical research. However, access to high-resolution data is often limited by proprietary formats and aggregated summaries that are unsuitable for detailed analysis. This work presents a reproducible methodology for data extraction and transformation from Xiaomi devices, applied in a clinical study with 179 participants suspected of obstructive sleep apnea. A two-stage pipeline was developed to convert exported files into structured, minute-level datasets, accessible through a graphical interface designed for non-technical researchers. The approach was evaluated in terms of data quality, robustness, and utility, successfully generating key metrics such as heart rate, oxygen saturation, respiration, steps, stress, and sleep stages. Results show that the methodology facilitates standardized access to physiological signals, supporting visualization and analysis in clinical and interdisciplinary research contexts.

1 Introduction

The increasing availability of consumer-grade wearable devices has opened new opportunities for health research, particularly in the field of sleep monitoring Guillodo et al. (2020). Wrist-worn devices such as smart bands can collect physiological signals (e.g., heart rate, respiration rate, blood oxygen saturation) and activity patterns with minimal burden on participants. This makes them attractive for large-scale and long-term studies, complementing traditional approaches such as polysomnography Concheiro-Moscoso et al. (2023).

However, despite their potential, the use of wearables in research is often limited by technical constraints. Data are usually stored in proprietary formats, accessible only through vendor applications, and provided as aggregated summaries that are not suitable for detailed analysis. Researchers face difficulties in extracting raw or fine-grained information, and available third-party tools are often restricted to a single brand or fail to produce interoperable outputs Elfouly et al. (2025).

In this work, we present a practical approach to data extraction and transformation from Xiaomi devices, applied within a clinical study on sleep. A total of 179 participants with suspected obstructive sleep apnea (OSA) wore the devices during 24-hour sessions, including overnight with concurrent nocturnal polygraphy. After each session, devices were returned

to the hospital, synchronized through an application, and exported as raw CSV files. We developed a two-stage script to convert these files: (i) a first version that generated simplified CSV summaries, and (ii) a second version that decomposed the data into multiple tables with minute-by-minute resolution. This transformation enables domain experts in sleep research to access detailed metrics in a standardized and interpretable format.

The main contributions of this paper are:

- A reproducible pipeline for extracting, cleaning and transforming wearable data into structured tables;
- Preliminary validation of the approach with real data from 179 participants;
- Identification of strengths and limitations, and discussion of future directions towards a generalized framework for multi-device integration.

2 Related work

The use of wearable devices for health monitoring has been widely explored in recent years, particularly in the context of sleep research. Commercial smart bands and smartwatches provide an affordable and user-friendly alternative to clinical-grade devices, offering continuous tracking of activity, heart rate, respiration, oxygen saturation and estimated sleep stages. Several studies have reported the potential of such devices to complement clinical tools or to support large-scale observational studies at home Guillodo et al. (2020). Nevertheless, their accuracy remains limited, and wearable-derived metrics should be interpreted with caution when compared to gold-standard measurements Birrer et al. (2024).
From a technical perspective, data accessibility is another major challenge Elfouly et al. (2025). While some vendor platforms restrict access to daily summaries, others do provide exports but in formats that are heterogeneous, incomplete, or difficult to integrate. Researchers attempting to retrieve minute-by-minute information often face cumbersome workflows, manual synchronizations, or APIs that require additional steps and deliver partial results.
Different tools and libraries have been proposed to overcome these issues Lee et al. (2023). Open-source projects and third-party software often focus on specific brands (e.g., Fitbit, Garmin), but their scope is limited and cross-device integration is rarely achieved. Moreover, existing solutions generally lack the ability to transform raw records into structured datasets tailored to clinical analysis. In particular, no available software was able to handle the Xiaomi device files used in our study and deliver the level of temporal granularity required Concheiro-Moscoso et al. (2025).
This gap motivated the development of a dedicated pipeline to extract and standardize data from Xiaomi device, with the goal of making them accessible and interpretable by domain experts in sleep research. Our work builds upon prior attempts to bridge consumer wearables and clinical research, but emphasizes reproducibility and extensibility to other devices.

3 Materials and methods

The data analyzed in this work were collected as part of an ongoing clinical study on effectiveness of the Xiaomi device as a complementary tool for monitoring health status and daily activity in individuals with OSA. A total of 179 participants have been recruited to date. The protocol involved the use of Xiaomi Smart Band 9 devices, which were distributed in parallel to three participants at a time. Each participant wore the device continuously for 24 hours, including a nocturnal polygraphy conducted simultaneously overnight. After the session, the devices were returned to the hospital, synchronized with an application, and prepared for the next group of participants Concheiro-Moscoso et al. (2025).

Data were exported from the Xiaomi ecosystem in the form CSV files. These files contained raw logs and aggregated records, but were not directly suitable for detailed analysis. Manual handling of these files was impractical given the number of participants and the need for temporal granularity.

3.1 Data extraction and transformation pipeline

To address this, we developed a two-stage pipeline to convert the raw exports into structured datasets:

1. **Simplified CSV summaries.** A first version of the script produced single CSV files containing condensed daily information. This reduced complexity and allowed preliminary inspection of the data, but did not preserve full temporal resolution.
2. **Structured minute-level tables.** The second version of the pipeline decomposed the data into multiple CSV tables, each corresponding to a physiological metric (e.g., steps, heart rate, respiration, oxygen saturation, sleep stages). Each table provided minute-by-minute values aligned by timestamp, enabling a detailed and standardized representation of the signals.

3.2 Implementation details

The pipeline was implemented in Python (version 3.12), using common libraries for data manipulation (e.g., pandas). Input files were validated to ensure temporal consistency and the removal of duplicated or missing entries. Outputs were generated as CSV files with a normalized schema, accompanied by a data dictionary describing the metrics and units.
On top of this core functionality, we developed a lightweight graphical user interface (GUI) using Tkinter. This interface enables researchers to select which datasets or metrics they wish to generate (e.g., heart rate, respiration, sleep stages) and to exclude others if not required. The GUI reduces unnecessary processing, simplifies file management, and provides a user-friendly workflow for investigators who may not have technical expertise. In this way, the pipeline can be operated directly by domain experts, lowering the barrier to adoption in clinical research environments.

3.3 Evaluation approach

In order to assess the pipeline, we defined a set of technical and usability criteria:

- **Data quality:** completeness of records, absence of duplicates, detection of out-of-range values.
- **Robustness:** percentage of input files successfully parsed, average processing time per file.
- **Utility:** coverage of metrics required by the sleep expert, and example visualizations of transformed data.

Although a full clinical validation was beyond the scope of this work, a subset of participants with simultaneous polysomnography was identified for future comparisons.

4 Results

The proposed pipeline was applied to the data collected from 179 participants. All exported CSV files from the Xiaomi devices were processed, although minor manual adjustments were occasionally required (e.g., correcting unexpected column headers or handling missing entries). The transformation yielded structured CSV tables organized by metric, with minute-level granularity. This organization provided a clearer and more interpretable view of the

physiological signals recorded during the study, while still depending on the accuracy and limitations of the consumer-grade sensors.

4.1 Example outputs

Figure 1 shows a sample output table generated by the pipeline, illustrating minute-by-minute heart rate values for a single participant. Similar tables were created for respiration rate, oxygen saturation, and sleep stages. The separation of metrics into different files simplified navigation and downstream analysis, compared to the aggregated format provided by the vendor.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Uid	Sid	Time	UpdateTime	avg_breath	avg_hr	avg_spo2	bedtime	sleep_deep_duration	device_bedtime	device_wake_up_time	sleep_light_duration	max_hr	max_spo2	min_hr	min_spo2	protoTime
2	8279514403	850623333	1748487300	1748781334	100	630	950	17484678600	970	17484678600	17484873000	1520	820	970	510	910	17484873000
3	8279514403	850623333	1748407200	1748781334	130	690	970	17483800800	1130	17483800800	17484072000	1740	880	990	490	930	17484072000
4	8279514403	850623333	1748359560	1748781334				17483546400	0	17483546400	17483595600	0					17483595600
5	8279514403	850623333	1748321040	1748781334	130	620	970	17483020800	1130	17483020800	17483210400	920	750	990	550	930	17483210400
6	8279514403	850623333	1747978800	1748193194	110	590	950	17479555200	1160	17479555200	17479788000	1480	720	990	500	920	17479788000
7	8279514403	850623333	1747929240	1748193194				17479230000	0	17479230000	17479292400	0					17479292400
8	8279514403	850623333	1747890900	1748193194	190	590	940	17478773400	730	17478773400	17478909000	830	880	980	480	920	17478909000
9	8279514403	850623333	1747806420	1748193194	180	710	960	17477814000	1240	17477814000	17478064200	1810	1170	1000	600	920	17478064200
10	8279514403	850623333	1747717020	1748193194	180	700	950	17476963200	1180	17476963200	17477170200	1410	840	970	620	930	17477170200
11	8279514403	850623333	1747378500	1747413441	180	660	950	17473591800	880	17473591800	17473785000	1560	970	980	530	900	17473785000
12	8279514403	850623333	1747194840	1747413441				17471849400	0	17471849400	17471948400	0					17471948400
13	8279514403	850623333	1747151280	1747413441				17471464800	0	17471464800	17471512800	0					17471512800
14	8279514403	850623333	1747108620	1747413441	170	560	950	17470882800	1060	17470882800	17471086200	1480	710	980	460	920	17471086200
15	8279514403	850623333	1746763140	1746894149				17467612200	0	17467612200	17467631400	0					17467631400
16	8279514403	850623333	1746757560	1746894149				17467533600	0	17467533600	17467575600	0					17467575600
17	8279514403	850623333	1746738660	1746894149				17467284600	0	17467284600	17467386600	0					17467386600
18	8279514403	850623333	1746591120	1746894149	160	780	940	17465762400	860	17465762400	17465911200	870	860	980	680	900	17465911200
19	8279514403	850623333	1746507600	1746894149	140	610	970	17464858800	970	17464858800	17465076000	1780	740	990	520	940	17465076000
20	8279514403	850623333	1745994000	1746365789				17459865600	0	17459865600	17459940000	0					17459940000
21	8279514403	850623333	1745926660	1746365789	150	530	940	17459657400	1010	17459657400	17459826600	1470	670	980	490	900	17459826600
22	8279514403	850623333	1745938020	1746365789				17459340000	0	17459340000	17459382000	0					17459382000
23	8279514403	850623333	1745854200	1746365789				17458512600	0	17458444200	17458542000	0					17458542000
24	8279514403	850623333	1745849280	1746365789				17458444200	0	17458444200	17458542000	0					17458492800
25	8279514403	850623333	1745558160	1745590685	160	740	940	17455357800	1110	17455357800	17455581600	1630	1000	980	610	900	17455581600
26	8279514403	850623333	1745386380	1745590685				17453847600	0	17453704800	17453863800	0					17453863800
27	8279514403	850623333	1745382420	1745590685				17453781600	0	17453704800	17453863800	0					17453824200
28	8279514403	850623333	1745377500	1745590685				17453745000	0	17453704800	17453863800	0					17453775000
29	8279514403	850623333	1745373840	1745590685				17453704800	0	17453704800	17453863800	0					17453738400
30	8279514403	850623333	1744780500	1744891013	180	740	960	17447562000	1580	17447562000	17447805000	1860	840	980	640	910	17447805000
31	8279514403	850623333	1744730400	1744891013				17447244600	0	17447244600	17447304000	0					17447304000
32	8279514403	850623333	1744699920	1744891013	210	760	950	17446990200	1110	17446990200	17446995200	2700	1000	1000	660	910	17446995200
33	8279514403	850623333	1744344360	1744891013				17443381800	0	17443232400	17443443600	0					17443443600
34	8279514403	850623333	1744337520	1744891013				17443352400	0	17443232400	17443443600	0					17443375200
35	8279514403	850623333	1744332720	1744891013				17443268400	0	17443232400	17443443600	0					17443327200
36	8279514403	850623333	1744326120	1744891013				17443232400	0	17443232400	17443443600	0					17443261200
37	8279514403	850623333	1744258740	1744891013	110	630	960	17442436200	720	17442436200	17442587400	1080	850	990	520	910	17442587400
38	8279514403	850623333	1744176060	1744891013	120	650	950	17441496600	1340	17441496600	17441760600	1970	770	990	580	900	17441760600
39	8279514403	850623333	1743742560	1743778890	200	830	960	17437165200	1080	17437165200	17437425600	2440	1070	990	760	910	17437425600

Figure 1: Processed CSV capture

The Tkinter-based graphical interface (Figure 2) allowed researchers to choose which metrics to export. While basic, this functionality reduced unnecessary processing and helped non-technical users interact with the pipeline without needing to edit code directly.

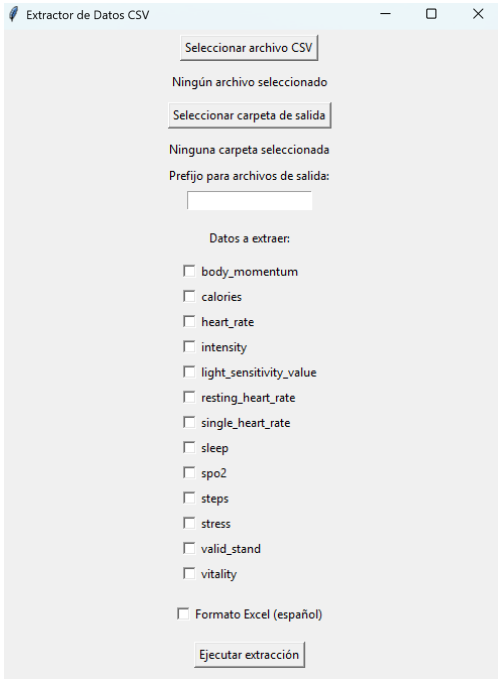

Figure 2: Interface capture

4.2 Data visualization

To illustrate potential applications, exploratory plots were generated from the standardized outputs. Figure 3 shows an example of aggregated sleep stage data across one night, together with the corresponding heart rate trend. These examples demonstrate that the pipeline can produce data suitable for visualization and preliminary analysis, but further validation is required before drawing clinical conclusions.

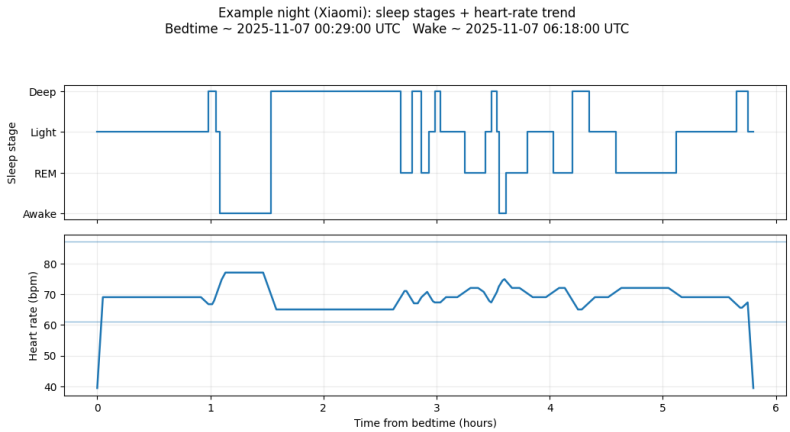

Figure 3: Example of sleep and heart rate data

4.3 Preliminary evaluation

The technical evaluation of the pipeline highlighted both strengths and limitations:

- **Data quality:** the majority of files were processed without duplication errors, although occasional missing values were present and required filtering.
- **Robustness:** all CSV exports could be parsed with the current script, which required on average **1 minute and 20 seconds** per file to complete the transformation process.
- **Utility:** the pipeline successfully provided all the core metrics requested by the sleep expert. Separate minute-level datasets were generated for steps, heart rate, blood oxygen saturation (SpO_2), stress, activity intensity, and detailed sleep data (including stage-level annotations). Validation confirmed that the script extracted these metrics correctly and consistently across participants. However, the current interface remains basic and could be improved for more complex research needs.

Feedback from the domain expert confirmed that the outputs were more convenient than the raw Excel exports. At the same time, the accuracy of the underlying device and the need for manual synchronization were recognized as major constraints of the approach.

5 Discussion

The results demonstrate that a relatively simple pipeline can make wearable-derived data more accessible to researchers, especially in interdisciplinary contexts where clinicians and computer scientists work together. By transforming vendor-specific exports into structured tables with minute-level resolution, the approach facilitates visualization and further processing, and reduces the technical burden on non-specialist users.

5.1 Strengths

One of the main strengths of the pipeline is reproducibility: the same script can be applied consistently to all exported files, ensuring homogeneous outputs across participants. The ability to separate metrics into different tables also improves interpretability, allowing domain experts to focus on specific signals such as respiration or sleep stages. Furthermore, the inclusion of a lightweight graphical interface lowers the barrier for adoption, enabling researchers without programming skills to operate the pipeline.

5.2 Limitations

Several limitations must be acknowledged. First, the accuracy of consumer-grade wearables such as Xiaomi smart bands is inherently lower than that of clinical devices like polysomnography Concheiro-Moscoso et al. (2023). This means that the transformed data should be interpreted with caution and primarily used for exploratory or large-scale observational purposes. Second, the process still depends on manual synchronization of devices and exports, which limits scalability in larger cohorts. Third, while the pipeline was robust for the data collected in this study, it remains tailored to a single brand and format; applying it to other devices would require additional adaptation.

5.3 Future directions

Building on this initial work, several improvements are planned. A key priority is to extend the pipeline with adapters for multiple wearable brands, moving towards a generalizable framework for heterogeneous data sources. Integration with interoperability standards (e.g., HL7 FHIR HL7 International (2019), Open mHealth Estrin and Sim (2010)) would further enhance compatibility with clinical systems. From a usability perspective, the graphical interface can be

expanded to provide richer configuration options and more advanced visualizations. Finally, future studies will compare wearable-derived outputs with simultaneous polysomnography to assess concordance and validate the utility of the transformed data.

6 Funding

This publication is part of the project 'Quality of life for caregivers through a person-centred technological solution'(TED2021-130127A-I00), funded by *MCIN/AEI/10.13039/501100011033* and by the European Union 'NextGenerationEU'/PRTR. Also, this work was supported by University of A Coruña (Universidade da Coruña), Xunta de Galicia and CITIC, which is funded by the department of Education, Science, Universities and Vocational Training of the Xunta de Galicia. TALIONIS research group of the University of A Coruña (grants for the consolidation and structuring of competitive research units (ED431B 2025/23)). CITIC is a centre accredited for excellence within the Galician University System and a member of the CIGUS Network (ED431G 2023/01). Additionally, it is cofinanced by the European Union through the FEDER Galicia 2021-2027 operational programme. PC-M also received funding for postdoctoral training from the Xunta de Galicia (ED481B-2023-125).

Bibliography

V. Birrer, M. Elgendi, O. Lambercy, and C. Menon. Evaluating reliability in wearable devices for sleep staging. *npj Digital Medicine*, 7, 2024. doi: 10.1038/s41591-024-01016-9.

P. Concheiro-Moscoso, B. Groba, D. Álvarez Estevez, M. d. C. Miranda-Duro, T. Pousada, L. Nieto-Riveiro, F. J. Mejuto-Muiño, and J. Pereira. Quality of sleep data validation from the xiaomi mi band 5 against polysomnography: Comparison study. *Journal of Medical Internet Research*, 25:e42073, 2023. ISSN 1438-8871. doi: 10.2196/42073. URL *https://www.jmir.org/2023/1/e42073*.

P. Concheiro-Moscoso, J. Pereira, M. Mosteiro-Añón, et al. Restech project on xiaomi wearable devices for monitoring and detecting obstructive sleep apnoea: observational study protocol. *BMJ Open*, 15(8):e101824, 2025. doi: 10.1136/bmjopen-2025-101824.

T. Elfouly et al. A comprehensive study on the efficacy of a wearable sleep. *Electronics* (*MDPI*), 14(17):3443, 2025. doi: 10.3390/electronics14173443.

D. Estrin and I. Sim. Open mhealth architecture: An engine for health care innovation. *Science*, 330(6005):759–760, 2010. doi: 10.1126/science.1196187.

E. Guillodo et al. Clinical applications of mobile health wearable–based sleep. *JMIR mHealth & uHealth*, 8:e10733, 2020. doi: 10.2196/10733.

HL7 International. Fhir release 4. *https://www.hl7.org/fhir/*, 2019. Accessed: 2025-09-24.

T. Lee et al. Accuracy of 11 wearable, nearable, and airable consumer sleep trackers: Prospective multicenter validation study. *JMIR mHealth & uHealth*, 11:e50983, 2023. doi: 10.2196/50983.