

A Multi-Sensor Approach for Soft Labeling in Human Activity Recognition Domain

Matteo Iervasi*, Cristian Turetta[†], Florenc Demrozi*, and Graziano Pravadelli[†]

**Dep. of Electrical Engineering and Computer Science, University of Stavanger, Norway, name.surname@uis.no*

[†]*Dep. of Engineering for Innovation Medicine, University of Verona, Italy, name.surname@univr.it*

Abstract—Manual annotation (MA) of sensor data for Human Activity Recognition (HAR) is labor-intensive, error-prone, and limits scalability. This paper proposes a multi-sensor methodology to automatically generate training labels (aka. soft labels) for HAR systems without human intervention. The approach integrates data from inertial measurement units glued to objects of daily life with the Received Signal Strength Indicator (RSSI) information derived from BLE beacon anchors for estimating both the performed activity and the subject’s location. We validate the quality of the generated soft labels against video-based MA ground truth. Experimental results show that a deep learning model for HAR trained on a Wi-Fi Channel State Information (CSI) dataset annotated with soft labels achieves comparable results with respect to the same model trained on the corresponding manually-annotated dataset.

Index Terms—automatic annotation, soft labels, manual annotation, human activity recognition.

I. INTRODUCTION

Human Activity Recognition (HAR) has become a key technology in smart environments, healthcare monitoring, active and assisted living, rehabilitation, sport, and wellbeing in general [1]. Despite its potential, the large-scale deployment of HAR systems remains challenging due to technological and methodological constraints [2].

Different sensing paradigms have been explored to overcome issues related to HAR approaches. Camera-based HAR provides detailed contextual information but is strongly constrained by privacy concerns, economic cost, computational requirements, and environmental conditions [3]–[5]. Wearable-based HAR enables direct motion sensing but requires users to continuously wear and maintain devices, which introduces cost, usability, and compliance challenges [2]. As a result, the practical applicability of these approaches in healthcare and everyday monitoring scenarios remains limited. Recently, wireless sensing techniques, and in particular WiFi Channel State Information (CSI), have emerged as a promising alternative, in particular, since it reduces limitations concerning privacy, durability, and usability. CSI provides fine-grained channel measurements that capture environmental changes caused by human presence and motion without requiring cameras or additional wearable devices [6]. Due to the ubiquity of WiFi infrastructure, CSI-based HAR offers a cost-effective and privacy-preserving solution [6]. However, despite significant progress in CSI-based sensing algorithms, their adoption is still limited by the major bottleneck that characterizes all HAR

solutions: the need for large amounts of labeled data to train the classification model [6].

Indeed, training supervised machine learning (i.e., Machine Learning (ML) or Deep Learning (DL)) models for HAR requires extensive labeled datasets with precise temporal boundaries of activities [1], [2]. Manual Annotation (MA) is labor-intensive and costly in terms of time and human resources, and it often relies on video recordings that raise privacy concerns and demand line-of-sight conditions. Video-based (semi) automatic methods exist to fasten the annotation task [7], [8], but they are also affected by inconsistencies, systematic errors in identifying activity transitions, and subjectivity when the feedback of a human observer is requested, which can negatively impact model performance and generalization [4]. These limitations translate into high economic costs and hinder the scalability of HAR systems [2], [9].

Paper contribution: In this work, we aim to address the data annotation challenge by introducing a new automatic methodology, based on multi-sensor data, to label HAR datasets in terms of “where the person is” and “what he/she is doing” within a daily living environment. The approach leverages Bluetooth Low Energy (BLE) Received Signal Strength Indicator (RSSI) data from both beacons located in the environment and smart objects, used/worn by the person, equipped with an Inertial Measurement Unit (IMU). During the activity annotation, RSSI is used to estimate the location of the person in the environment (e.g., near the table, near the kitchen, etc.), while smart objects are exploited to recognize activities performed in the environment (e.g., eating, writing, etc.). In this way, each sample in the dataset is automatically annotated with a soft label (SL) represented by the pair $\langle actions, locations \rangle$.

By generating reliable soft-label annotations (SL) without human intervention, our approach reduces the reliance on MA and constitutes a first step toward overcoming key limitations that currently hinder the design of robust HAR systems. It should be noted, however, that while our setup and evaluation, in the rest of the paper, applies the proposed soft-labeling approach to CSI-based HAR data, the methodology itself is agnostic to the dataset type. The creation of high-quality CSI datasets and the subsequent training of an HAR model lie beyond the scope of this paper.

Related work and research gap: To systematically identify

existing contributions related the aims of this paper, we conducted a structured literature search in Scopus employing query terms that integrate relevant keywords (synonyms and variations) from both “HAR”, “automatic labeling domains”, “sensor/multi-sensor data”, and “CSI and wireless sensing”.

The initial query, which explicitly focused on “CSI and wireless sensing” in combination with automatic labeling, returned only four relevant articles. Among these, one survey [10] and three methodologies, not comparable with our approach (i.e., [11] video-based, [12] self-supervised, and [13] not related to soft labeling). This indicates that research on automatic label generation in the specific domain of CSI-based HAR remains largely unexplored. To broaden the scope and achieve a more comprehensive overview, we refined the query by removing the “CSI and wireless sensing” constraint, thus applying the search to the general HAR domain. This extended search retrieved 29 works, of which 17 were surveys or opinion papers. The remaining 12 works mainly focus on transfer learning [14], reinforcement learning [15], or semi-supervised approaches [16], which still require an initially annotated training set or human intervention/feedback to be effective. These findings align with the observations made in the recent systematic review presented in [9], which highlights the scarcity of studies addressing automated or weakly supervised labeling in HAR. As clearly discussed therein, only two works can be considered related to our approach, though they still differ substantially in their objectives and methodologies (i.e., in [17] the methodology is not tested on HAR scenario, and in [18] a location recognition system is addressed).

Paper organization: The rest of this paper is organized as follows. Section II details our methodology, including the data collection setup, preprocessing pipeline and soft labeling generation process. Section III presents experimental results. Finally, Section IV concludes the paper.

II. METHODOLOGY

Figure 1 depicts the automatic annotation methodology proposed in this paper. It is composed of three main steps: i) data collection, ii) data pre-processing, and iii) soft labeling. The effectiveness of the methodology has been evaluated by training a HAR model on a CSI dataset with the automatically-generated soft labels and comparing its performance against a model trained with ground-truth manual annotations, as presented in Section III.

A. Environment Setup

Two sets of equipment have been configured, the first for the collection of the CSI-based HAR dataset, which is used as an evaluation scenario, and the second for the soft-label annotation of the CSI dataset. The layouts of the overall set-up, for the two different settings used in the experimental analysis are illustrated in Figure 2.

CSI provides fine-grained measurements of the wireless channel during the communication between two devices at

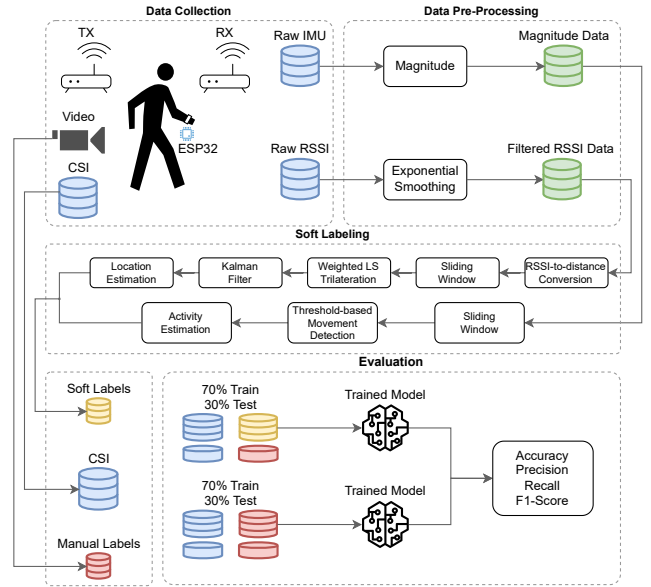


Fig. 1: Methodology overview.

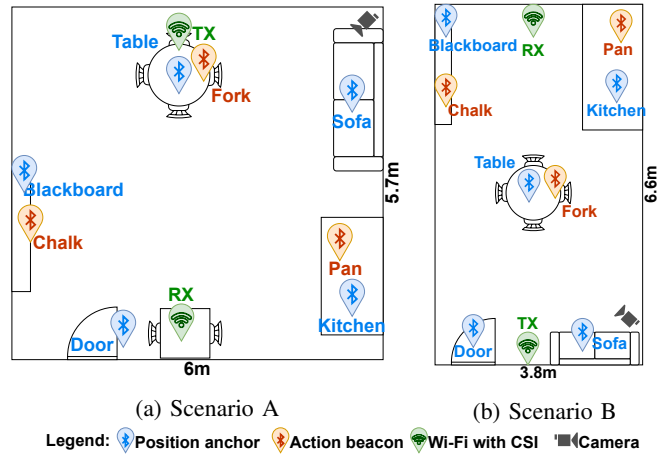


Fig. 2: Layout of the data collection and soft label annotation equipment located in two different environments.

the level of individual subcarriers. Then, it is used to capture amplitude and phase distortions caused by human motion and environmental dynamics. For generating the CSI dataset, two modified ASUS RT-AC58U routers were employed; one was configured as a packet transmitter (TX) and the other as a receiver (RX) using raw 802.11 packets in promiscuous mode. These devices run a special version of OpenWrt [19], which includes the `nexutil` extensions [20] that enable CSI data extraction. Special packets containing custom MAC addresses and specific payloads were sent by the TX at a fixed frequency of 100 ms, and the RX logged them into a `.pcap` file, which corresponds to the target CSI dataset.

To perform the soft labeling of the CSI dataset we exploit, instead, two information sources:

- BLE RSSI data referred to the communication between Nordic Semiconductor’s Thingy:52 devices, located in specific positions of the environment (e.g. kitchen, sofa, etc.) placed at approximately 1 m from the ground and

TABLE I: Devices used for data collection and labeling.

Device	Role	Data
ESP32	BLE scanner	Beacon's RSSI
Thingy:52	Position anchor	Beacon advertisements
Thingy:52	Action beacon	Accel/Gyro/Compass (IMU)
ASUS RT-AC58U	CSI transmitter	Wi-Fi packets
ASUS RT-AC58U	CSI receiver	CSI PCAPs
Smartphone	Camera	Video
Gateway	Broker/logger	RSSI, IMU, CSI

operating as beacons, thus acting as position anchors, and an Espressif ESP32, running a custom firmware, worn by the target subject on the arm with a velcro strip while performing daily life activities. RSSI quantifies the power of a received wireless signal; it is widely used for *localization*, since the signal strength decreases with distance, enabling coarse position estimates. Thus, RSSI data is used to create the soft labels for the subject location.

- IMU data provided by Thingy:52 devices glued to objects (e.g., fork, pan, etc.) used by the subject to perform daily life activities (e.g, eating, cooking). This data is used to create the soft labels for the performed activities.

Table I summarizes the devices used in the environment setup and their roles in the methodology.

B. Data Collection

The Espressif ESP32 is programmed to periodically scan the list of position anchors, with an interval of 150 ms and a 100 ms scan window. These devices are configured to perform a BLE beacon advertisement every 100 ms. When booted up, the ESP32 connects to a pre-defined Wi-Fi network and synchronizes its time using the Network Time Protocol (NTP). This step is fundamental to ensure that each collected sample has a correct timestamp. After synchronization, the device connects to the MQTT broker hosted on a central gateway, where it publishes tuples of the form $\langle \text{MAC}, \text{RSSI}, \text{timestamp} \rangle$ after every BLE scan. Action beacon devices glued to objects, on the other hand, are managed directly by the central gateway. A dedicated Python program establishes a direct connection with all of them and logs their IMU data. The central gateway is therefore used as a coordinator, it executes a Python program managing all components to collect both the CSI dataset and the corresponding RSSI/IMU data for soft labeling. When initialized, it first connects to the MQTT broker to log all the RSSI samples published by the ESP32. Then, it performs the required setup for CSI data collection connecting to the RX router via SSH. Finally, it establishes a direct BLE connection with the action beacons to log their IMU data.

C. Data Pre-Processing

The raw streams related to RSSI and IMU data are inherently noisy and prone to artifacts, which can negatively affect the quality of soft labelling. Pre-processing is therefore required to improve robustness and ensure features reflect the underlying phenomena. We apply filtering and transformation to RSSI and IMU data before soft-label generation.

For the RSSI measurements, we mitigate fluctuations caused by multi-path propagation, interference, and antenna orientation using exponential smoothing. An exponential moving average (EMA) filter computes each smoothed value \hat{s}_t from the current observation x_t and the previous value \hat{s}_{t-1} as:

$$\hat{s}_t = \alpha \cdot x_t + (1 - \alpha) \cdot \hat{s}_{t-1},$$

where $\alpha \in [0, 1]$ controls the trade-off between responsiveness and noise reduction. We empirically set $\alpha = 0.3$, which effectively suppresses noise while preserving signal dynamics. The filter is initialized with $\hat{s}_0 = x_0$ (the first RSSI value collected) to avoid startup transients.

For the IMU measurements, we process the raw tri-axial acceleration data by computing the vector magnitude, defined as $\sqrt{a_x^2 + a_y^2 + a_z^2}$. This transformation removes the dependence on sensor orientation, providing a rotation-invariant measure of movement intensity that is more robust across users and recording conditions.

D. Soft Labeling

To automatically generate soft labels, our approach leverages the fusion of RSSI and IMU data obtained after the pre-processing phase previously described. The process follows a two-stage pipeline: first, the estimation of location coordinates from RSSI signals is derived, and semantic location labels based on anchor deployment zones are accordingly assigned; secondly, IMU data related to object movements are elaborated to eventually associate an activity, thus obtaining the pair $\langle \text{actions}, \text{locations} \rangle$.

The first stage, to estimate the location coordinates, relies on fixed position anchors with known positions throughout the environment, each broadcasting at a calibrated transmission power of -59 dBm. For location estimation, we employ a log-distance path loss model with an exponent of 2.7, which accounts for indoor propagation characteristics. The RSSI-to-distance conversion follows the equation $d = 10^{(P_{tx} - \hat{s}_t)/10n}$, where P_{tx} is the reference power at 1 meter and n is the path loss exponent. The recording of this value, at the passing of time, creates a time series of distances, tsd . Then, on tsd we use a sliding window approach with a width of 60 samples, and a 5 sample stride to balance temporal resolution with noise reduction. Within each window, we apply weighted least squares trilateration, which assigns higher weights to stronger RSSI signals, thereby reducing the influence of distant or occluded anchors. The optimization minimizes the weighted sum of squared distance residuals between predicted and measured distances across all position anchors. To handle failed estimations due to poor signal conditions or geometric dilution of precision, we compute the centroid of all successful position estimates within each window, providing robust aggregation against outliers.

The estimated positions then undergo Kalman filtering to smooth the trajectory and reduce estimation noise. We configure the Kalman filter with a process noise standard deviation of 0.7 meters, and initial position uncertainty of 0.5 meters.

Finally, we discretize the continuous position estimates into semantic labels corresponding to the anchor deployment zones (e.g., kitchen, sofa, etc.). This discretization enables classification-based approaches and provides interpretable location categories for analysis. We assign each estimated position to its nearest anchor location using Euclidean distance in the 2D plane. For each pair of position coordinates (x_i, y_i) we compute the distances with respect to all N anchor locations and identify the minimum distance anchor using the following formula:

$$l = \arg \min_{j \in [1, N]} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

To handle positions that fall outside the reliable coverage area of any anchor, we set a distance threshold of 1.5 meters. Positions within this threshold are assigned the corresponding anchor’s identifier as their location label, effectively creating circular zones of radius 1.5 meters around each anchor. Positions exceeding this threshold from all anchors are labeled as *Not Marked*, indicating areas of uncertain localization or transition zones between anchor coverage areas.

At this point, the semantic label corresponding to the position is coupled with the label related to the performed activity. Each activity is associated to a specific object used to perform it (e.g., a pan for cooking, a chalk for writing, etc.). A Thingy:52 (aka. action beacon) is glued to each object to collect IMU signals while using it. To extract the activity labels from the IMU signals, we adopt a sliding window approach. Windows of 60 samples (corresponding to 1 s at the device’s sampling rate) are processed with a 30% overlap to ensure temporal continuity. For each window, we compute the maximum–minimum difference of the IMU acceleration vector magnitude. If this difference is below an empirically defined threshold of 0.1 m/s^2 , the object is considered stationary (the label then will be *no action*). Otherwise, the object is labeled as *activity_i* (e.g., cooking, writing, etc.), indicating that the object was used to perform the action corresponding to *activity_i*.

At the end of this phase, each sample in the target CSI dataset is assigned a pair $\langle \text{actions}, \text{locations} \rangle$, where $\text{location} \in \{l \mid l \text{ is a target location}\} \cup \{\text{not marked}\}$ and $\text{activity} \in \{a \mid a \text{ is a target action}\} \cup \{\text{no action}\}$, which represent its final soft label.

III. EXPERIMENTAL RESULTS

We evaluate the effectiveness of the proposed methodology for generating soft labels using a two-fold approach. At first, by directly comparing the soft labels with manual annotations (Section III-C). Then, by comparing the performance of the deep learning model described in Section III-D when trained with soft labels versus manual labels (Section III-E).

A. Dataset Overview

Two male participants (1.90 m and 1.80 m) took part in two scenarios, represented in Figure 2, providing informed consent prior to the study. Scenario A included both participants (102,792 samples, 25.7 min), while Scenario B involved only

the 1.80 m subject (341,498 samples, 85.3 min). The dataset comprises 444,290 (i.e., 111 min) synchronized samples distributed across four activity classes: *Writing* (53,287), *No Action* (291,615), *Eating* (48,258), and *Cooking* (51,130). Location annotations span six classes: *Table* (101,315), *Sofa* (107,766), *Kitchen* (101,872), *Door* (29,451), *Blackboard* (64,354), and *Transition* (39,532). Each session followed a scripted path: door \rightarrow blackboard \rightarrow kitchen \rightarrow table \rightarrow sofa \rightarrow door, with at least 30 seconds spent at each location (max 2 and 20 minutes for, respectively, Scenario A and B). The cycle was repeated four times, twice with interaction with smart objects and twice without, ensuring balanced coverage of activities and zones. Ultimately, Scenario B produces a more unbalanced dataset than Scenario A due to the significant duration disparity between activities and transitions. In Scenario B, activities performed in specific locations can last up to 20 minutes, while unmarked transitions between locations take only a few seconds, creating a substantial imbalance. In contrast, Scenario A maintains a better balance since activities are limited to 2 minutes in duration.

B. Manual Annotation Protocol

Manual labels are obtained through video annotation performed by an operator. Because the video recordings do not provide absolute timestamps, synchronization with the sensor data is achieved using a visual reference: at the beginning of each session, the LED of the Thingy:52 (aka. action beacon) device turned red. This event, clearly visible in the video, is used as a temporal synchronization point to align manual annotations with the automatically generated soft labels. Two annotation levels were defined:

- 1) **Activity-level:** labels are assigned when the subject is actively manipulating the corresponding object (i.e., during motion). Idle states of the objects are annotated as *no action*.
- 2) **Position-level:** labels are assigned when the subject is stationary within a semantic zone associated with one of the deployed position anchors. Transient passages through a zone are annotated as *not marked*.

As a result, short-lived transitions (e.g., when moving, in Figure 2a, from blackboard to kitchen passing nearby table) are not included in the manual labels. While this design choice simplifies the annotation procedure, it inevitably introduces gaps in label coverage.

C. Soft and Manual Label Alignment

We evaluated the quality of the soft-label annotation (SL) by comparing it against the MA. The assessment was conducted using standard classification metrics for time series—accuracy, precision, recall, and F1-score—where the MA time series served as the reference and the SL time series as the predicted labels. As an illustrative example, Figure 3 shows a data collection session from Scenario A, highlighting the alignment between SL and MA for both activities and locations.

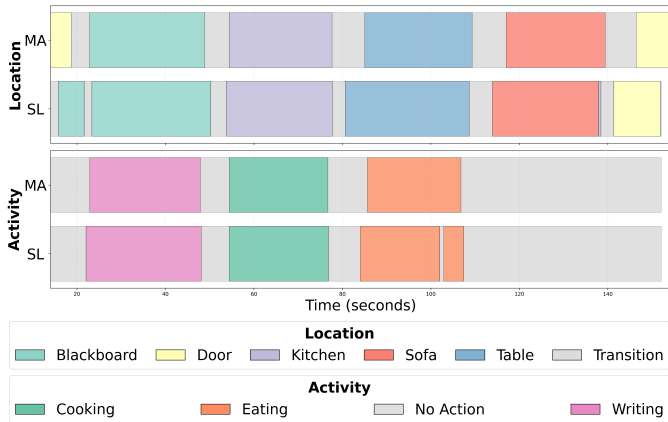


Fig. 3: Example of alignment between soft labels and manual labels for scenario A.

TABLE II: Performance comparison of soft-label annotation against manual annotation (ground truth) using confusion-matrix metrics. Results are expressed in percentage (%).

Metrics	Scenario A		Scenario B	
	Activity	Location	Activity	Location
Accuracy	97.54	77.04	98.75	92.29
Macro Precision	96.00	75.10	99.40	78.00
Macro Recall	95.60	70.20	97.50	80.00
Macro F1-Score	95.70	70.60	98.40	78.70
Weighted Precision	97.50	78.10	98.80	87.60
Weighted Recall	97.50	77.00	98.80	92.30
Weighted F1-Score	97.50	76.40	98.70	89.60

For each scenario in Fig. 2, we computed the classification performance of SL against MA for activity recognition and location estimation, as summarized in Table II.

Across both scenarios, soft labeling aligns more closely with manual annotation for activities than for locations. This gap reflects the instability of RSSI-based transitions, where small boundary shifts remain a dominant error source, particularly in Scenario A, as the duration of performed activities and locations are relatively small (max 2 minutes each). Scenario B shows higher performance, benefiting from a longer data collection period and more clearly separated activity phases (up to 20 minutes each), which reduced ambiguity in location transitions. Overall, these results confirm that the proposed pipeline achieves high accuracy for structured activities with distinct motion patterns, across different participants, and during stationary location phases. In contrast, passive behaviors and rapid transitions remain the most challenging conditions for reliable soft-label generation.

A deeper evaluation of the effectiveness of the proposed SL against MA has been conducted by training a DL classification model on a CSI dataset. The model has been trained twice: once by using the soft labels, and once by using the manual labels, and tested in both cases against the manual annotation (Figure 1, bottom part). The goal was to evaluate how the imprecision of SL versus MA reflects on the performance of the DL model in correctly recognizing activities and locations. The next section then presents the adopted DL model, while

the SL versus MA evaluation is reported in Section III-E.

D. Deep Learning Model Training

We adopted a multi-task learning approach for the simultaneous recognition of activities and locations by using Wi-Fi CSI data. The DL network employs a hierarchical feature extraction backbone consisting of six convolutional layers organized in three stages with progressive channel expansion. Each convolutional block combines a 2D convolution with batch normalization and ReLU activation. The model processes temporal information through a snapshot-wise approach in which each CSI snapshot is initially treated as an independent sample through the convolutional backbone, extracting spatial-frequency features. Adaptive average pooling then aggregates the temporal dimension, producing a unified representation that encodes both spatial-frequency characteristics and temporal dynamics. The extracted features pass through a shared fully-connected network with two hidden layers, using dropout regularization ($p = 0.3$) at each layer. This shared representation then branches into task-specific heads for activity and location classification. Each head consists of a two-layer Multi-Layer Perceptron (MLP) with ReLU activation and dropout, projecting the shared features to the respective output spaces. This architecture enables the model to learn complementary representations, where action patterns inform location prediction and vice versa.

We optimize the joint objective function $\mathcal{L} = \mathcal{L}_{act} - \mathcal{L}_{loc}$ using AdamW with learning rate 3×10^{-4} and weight decay 1×10^{-4} for L2 regularization. Both classification tasks employ cross-entropy loss, with equal weighting to balance the multi-task learning. Training proceeds by splitting data into train (70%), validation (10% of the training set), and test (30%) sets. We trained the model for up to 150 epochs with early stopping (patience=20) based on validation loss (\mathcal{L}_{val}). To ensure a comprehensive evaluation, we repeated this process five (i.e., 5) times changing the random seed used to split the data.

E. Impact of SL-MA Misalignment on DL Model Performance

While the DL model described above was employed to evaluate the quality of SL against MA, it should be emphasized that defining the optimal DL architecture for activity and location recognition is beyond the scope of this paper. Instead, our focus is on analyzing how misalignment between SL and MA affects the ability of the DL model to accurately predict the activities and locations of the target subjects.

Interesting observations derived by analyzing Table III and Table IV, where columns report (results from 5 repetitions), for both Scenarios A and B, the classification metrics obtained by training the DL model twice—once with the SL dataset and once with the MA dataset—and testing it, in both cases, against the MA ground truth. The absolute error is computed as the difference Δ between the metrics obtained with MA and SL training datasets, and the corresponding relative error Δ_r is also reported.

In Table III, performances are computed with the *macro method*, while in Table IV with the *weighted method*; the

TABLE III: Performances of the DL model trained by using SL and MA datasets for the two scenarios of Figure 2 computed by considering the *macro method*. Results are expressed in percentage.

Type	Metrics	Scenario A				Scenario B			
		MA	SL	Δ	Δ_r	MA	SL	Δ	Δ_r
Activity	Accuracy	90.67 (\pm 0.47)	91.15 (\pm 1.01)	0.48 (\pm 1.01)	0.53 (\pm 0.54)	99.84 (\pm 0.14)	99.84 (\pm 0.07)	-0.00 (\pm 0.07)	0.00 (\pm 0.85)
	Precision	86.64 (\pm 1.64)	88.08 (\pm 1.76)	1.44 (\pm 1.76)	1.66 (\pm 0.06)	99.79 (\pm 0.19)	99.83 (\pm 0.03)	0.05 (\pm 0.03)	0.05 (\pm 5.98)
	Recall	86.76 (\pm 1.10)	87.62 (\pm 2.22)	0.87 (\pm 2.22)	1.00 (\pm 0.50)	99.75 (\pm 0.26)	99.75 (\pm 0.11)	0.00 (\pm 0.11)	0.00 (\pm 1.28)
	F1-Score	86.34 (\pm 1.11)	87.62 (\pm 1.38)	1.29 (\pm 1.38)	1.49 (\pm 0.20)	99.77 (\pm 0.21)	99.79 (\pm 0.07)	0.02 (\pm 0.07)	0.02 (\pm 1.93)
Location	Accuracy	86.31 (\pm 1.59)	77.31 (\pm 1.48)	-9.00 (\pm 2.53)	10.42 (\pm 0.07)	98.93 (\pm 0.73)	93.24 (\pm 0.49)	-5.69 (\pm 0.63)	5.75 (\pm 0.49)
	Precision	85.72 (\pm 1.78)	72.17 (\pm 5.68)	-13.55 (\pm 7.32)	15.80 (\pm 0.69)	98.31 (\pm 1.05)	78.25 (\pm 0.19)	-20.07 (\pm 1.11)	20.41 (\pm 4.61)
	Recall	84.97 (\pm 2.06)	68.94 (\pm 1.46)	-16.02 (\pm 3.31)	18.85 (\pm 0.41)	98.41 (\pm 1.23)	82.23 (\pm 1.01)	-16.18 (\pm 1.05)	16.45 (\pm 0.21)
	F1-Score	85.21 (\pm 1.89)	68.56 (\pm 2.26)	-16.65 (\pm 4.14)	19.54 (\pm 0.16)	98.35 (\pm 1.15)	80.05 (\pm 0.53)	-18.30 (\pm 0.99)	18.61 (\pm 1.18)

TABLE IV: Performances of the DL model trained by using SL and MA datasets for the two scenarios of Figure 2 computed by using the *weighted method*. Results are expressed in percentage.

Type	Metrics	Scenario A				Scenario B			
		MA	SL	Δ	Δ_r	MA	SL	Δ	Δ_r
Activity	Accuracy	90.67 (\pm 0.47)	91.15 (\pm 1.01)	0.48 (\pm 1.01)	0.53 (\pm 0.54)	99.84 (\pm 0.14)	99.84 (\pm 0.07)	-0.00 (\pm 0.07)	0.00 (\pm 0.85)
	Precision	91.00 (\pm 0.77)	91.35 (\pm 0.98)	0.35 (\pm 0.98)	0.38 (\pm 0.22)	99.84 (\pm 0.14)	99.84 (\pm 0.07)	-0.00 (\pm 0.07)	0.00 (\pm 0.85)
	Recall	90.67 (\pm 0.47)	91.15 (\pm 1.01)	0.48 (\pm 1.01)	0.53 (\pm 0.54)	99.84 (\pm 0.14)	99.84 (\pm 0.07)	-0.00 (\pm 0.07)	0.00 (\pm 0.85)
	F1-Score	90.69 (\pm 0.50)	91.13 (\pm 1.00)	0.44 (\pm 1.00)	0.48 (\pm 0.50)	99.84 (\pm 0.14)	99.84 (\pm 0.07)	0.00 (\pm 0.07)	0.00 (\pm 0.84)
Location	Accuracy	86.31 (\pm 1.59)	77.31 (\pm 1.48)	-9.00 (\pm 2.53)	10.42 (\pm 0.07)	98.93 (\pm 0.73)	93.24 (\pm 0.49)	-5.69 (\pm 0.63)	5.75 (\pm 0.49)
	Precision	86.52 (\pm 1.55)	76.86 (\pm 2.96)	-9.66 (\pm 4.43)	11.16 (\pm 0.48)	98.94 (\pm 0.73)	88.26 (\pm 0.46)	-10.68 (\pm 0.61)	10.79 (\pm 0.58)
	Recall	86.31 (\pm 1.59)	77.31 (\pm 1.48)	-9.00 (\pm 2.53)	10.42 (\pm 0.07)	98.93 (\pm 0.73)	93.24 (\pm 0.49)	-5.69 (\pm 0.63)	5.75 (\pm 0.49)
	F1-Score	86.28 (\pm 1.55)	75.99 (\pm 1.58)	-10.29 (\pm 2.88)	11.93 (\pm 0.02)	98.92 (\pm 0.74)	90.51 (\pm 0.50)	-8.41 (\pm 0.63)	8.50 (\pm 0.48)

difference is that the first calculates metrics independently for each class and then takes the unweighted mean, treating all classes equally regardless of their frequency. In contrast, the weighted method computes the same metrics for each class but takes the average weighted by the number of instances (support) in each class, giving more influence to classes with higher representation in the dataset. This makes the weighted method more sensitive to performance on frequent classes, while the macro method provides equal importance to all classes, making it particularly useful for detecting poor performance on minority classes. Given that classes in Scenario B are unbalanced while Scenario A features a more equal distribution, we employed both evaluation methods to comprehensively demonstrate that SL performs effectively under both balanced and unbalanced conditions.

Focusing on the **activity recognition**, Δ and Δ_r between the model trained on manual labels and the one trained on soft labels are minimal. This indicates that both manual and soft labels allow the CSI-based model to effectively discriminate between the scripted activities, confirming that SL provides supervision of comparable quality to MA. Indeed, the misalignment error introduced by using SL instead of MA (see Table II) is largely compensated by the DL model during the training process (see Table III and Table IV, where Δ and Δ_r are almost negligible), demonstrating that for activity recognition, our automatic soft-labeling approach represents a valuable alternative to labor-intensive manual annotation.

The compensatory effect of the DL model is even more pronounced for **location estimation**. While the misalignment between SL and MA is non-negligible for locations—due to RSSI-related issues discussed in Section III-C (see Table II)— Δ and Δ_r in model performance are significantly reduced. For instance, in Scenario A, the macro (weighted) F1-score for location shows a direct loss of approximately 30% (24%) when comparing SL and MA directly, whereas Δ

and Δ_r in macro (weighted) F1-score after training the DL model drops to approximately 16% and 19% (10% and 11%) in Table III (Table IV). This demonstrates that the DL model can partially compensate for the initial labeling inaccuracies through its learning process.

Similar observations hold for Scenario B, though the compensatory effect is less pronounced. This occurs because the DL model achieves higher baseline performance in Scenario B for both SL and MA, leaving less room for improvement through error compensation. Despite this ceiling effect, SL remains effective for training DL models even when substantial misalignment with ground truth exists, as demonstrated in Scenario A.

IV. CONCLUSION

This paper introduces a multi-sensor methodology that leverages BLE beacons and IMU sensors to automatically annotate soft labels on datasets used for training HAR systems without manual annotation. Experiments on a CSI-based dataset collected on two environments show that the generated labels are highly reliable, and that a deep learning model can compensate for the error introduced by automatic annotation, such that when trained with soft labels, the model achieves performance comparable to that obtained by training it with a manually-labeled dataset. These findings demonstrate the potential of automatic annotation to replace costly manual labeling and enable scalable HAR systems. Indeed, it is worth noting that manually annotating the considered dataset (1.85 h of recordings) required approximately 5 h of video labeling.

Future work will explore the application of this soft-labeling framework beyond WiFi CSI. Since the methodology is agnostic to the underlying sensing modality, it can be extended to other radio-frequency technologies such as mmWave or Ultra Wide Band, which provide higher spatial resolution and are increasingly investigated for human sensing tasks.

REFERENCES

- [1] W. Qi, X. Xu, K. Qian, B. W. Schuller, G. Fortino, and A. Aliverti, "A review of aiot-based human activity recognition: From application to technique," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 4, pp. 2425–2438, 2024.
- [2] Y. Yin, L. Xie, Z. Jiang, F. Xiao, J. Cao, and S. Lu, "A systematic review of human activity recognition based on mobile devices: Overview, progress and trends," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 890–929, 2024.
- [3] G. Bholra and D. K. Vishwakarma, "A review of vision-based indoor har: state-of-the-art, challenges, and future prospects," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 1965–2005, 2024.
- [4] T. F. N. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujaali, and A. Jalal, "A review of video-based human activity recognition: theory, methods and applications," *Multimedia Tools and Applications*, vol. 84, no. 17, pp. 18 499–18 545, 2025.
- [5] P. Climent-Pérez and F. Florez-Revuelta, "Protection of visual privacy in videos acquired with rgb cameras for active and assisted living applications," *Multimedia Tools and Applications*, vol. 80, no. 15, pp. 23 649–23 664, 2021.
- [6] I. Bisio, C. Fallani, C. Garibotto, F. Lavagetto, A. Sciarrone, and M. Zerbino, "Analysis of csi-based human activity recognition for contactless patients monitoring," in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2024, pp. 438–443.
- [7] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.
- [8] A. Berg, J. Johnander, F. Durand de Gevigney, J. Ahlberg, and M. Felsberg, "Semi-automatic annotation of objects in visual-thermal video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [9] F. Demrozi, C. Turetta, F. A. Machot, G. Pravadelli, and P. H. Kindt, "A comprehensive review of automated data annotation techniques in human activity recognition," *arXiv preprint arXiv:2307.05988*, 2023.
- [10] M. A. Hossen and P. E. Abas, "Machine learning for human activity recognition: State-of-the-art techniques and emerging trends," *Journal of Imaging*, vol. 11, no. 3, p. 91, 2025.
- [11] B. Sheng, C. Sun, F. Xiao, L. Gui, and Z. Guo, "Muat-va: Multi-attention and video-auxiliary network for device-free action recognition," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10 870–10 880, 2023.
- [12] A. K. Koupai, M. J. Bocus, R. Santos-Rodriguez, R. J. Piechocki, and R. McConville, "Self-supervised multimodal fusion transformer for passive activity recognition," *IET Wireless Sensor Systems*, vol. 12, no. 5-6, pp. 149–160, 2022.
- [13] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal csi-based human activity recognition using gans," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17 345–17 355, 2021.
- [14] M. Thukral, H. Haresamudram, and T. Ploetz, "Cross-domain har: Few-shot transfer learning for human activity recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, pp. 1–35, 2025.
- [15] Y. Cui, S. K. Hiremath, and T. Ploetz, "Reinforcement learning based online active learning for human activity recognition," in *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 23–27.
- [16] A. Jiang and J. Ye, "Selfvis: Self-supervised learning for human activity recognition based on area charts," *IEEE Transactions on Emerging Topics in Computing*, vol. 13, no. 1, pp. 196–206, 2024.
- [17] F. Demrozi, M. Jereghi, and G. Pravadelli, "Towards the automatic data annotation for human activity recognition based on wearables and ble beacons," in *2021 IEEE International Symposium on Inertial Sensors and Systems (INERTIAL)*. IEEE, 2021, pp. 1–4.
- [18] T. Dissanayake, T. Maekawa, T. Hara, T. Miyanishi, and M. Kawanabe, "Indolabel: Predicting indoor location class by discovering location-specific sensor data motifs," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5372–5385, 2021.
- [19] "Home — Asuswrt-Merlin — asuswrt-merlin.net," <https://www.asuswrt-merlin.net/>, [Accessed 27-08-2025].
- [20] "GitHub - seemoo-lab/nexmon: The C-based Firmware Patching Framework for Broadcom/Cypress WiFi Chips that enables Monitor Mode, Frame Injection and much more — github.com," <https://github.com/seemoo-lab/nexmon>, [Accessed 27-08-2025].