IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.006, LIIR363
DOI: https://doi.org/10.52731/liir.v006.363

# Multi-User Activity Recognition in an Indoor Environment with Transformer Architectures

MD Irteeja Kobi [*], Pedro Machado [*], Ahmad Lotfi [*],
Daniyal Haider [*], Isibor Kennedy Ihianle[† ‡]

## Abstract

This paper proposes a device-free Human Activity Recognition (HAR) system, utilising Wi-Fi Channel State Information (CSI) to maintain the privacy of users in a multi-user environment. To achieve this goal, substantial annotated training data is required, which is often imbalanced with poor generalisability in complex, multi-user environments. To overcome these gaps, a hybrid deep learning approach is proposed that integrates signal pre-processing, targeted data augmentation, and a novel CNN incorporating a Transformer model. Experimental results show that the proposed model outperforms several baselines in single-user and multi-user contexts. Our findings demonstrate that combining real and augmented data significantly improves model generalisation in scenarios with limited labelled data.
*Keywords:* Human Activity Recognition, HAR, Channel State, CSI, Deep Learning, CNN, Transformer.

## 1 Introduction

Human Activity Recognition (HAR) algorithms based on information obtained from ambient sensors, wearable sensors or vision-based systems are successfully applied in detecting many basic human activities [1]. Although effective, these modalities pose considerable limitations; ambient sensors such as motion detectors do not provide accurate information about a specific activity, wearable devices demand user compliance, regular charging of batteries, and maintenance, while vision-based systems suffer from occlusion, varying illumination, and severe privacy concerns [1, 2]. To preserve the privacy of users, some research with thermal vision employing Thermal Sensor Arrays (TSA) has shown promising results [3]. None of the proposed approaches so far provides a holistic solution to the HAR, especially for an indoor multi-user environment.

[*] Corresponding authors: Ahmad Lotfi, Email: ahmad.lotfi@ntu.ac.uk
[†] Corresponding authors: Isibor Kennedy Ihianle, Email: isibor.ihianle@ntu.ac.uk
[‡] All authors are with Department of Computer Science, Nottingham Trent University, Nottingham, NG11 8NS, United Kingdom

To address the above limitations, device-free methods using Wi-Fi Channel State Information (CSI) have emerged. The CSI captures subtle radio signal changes caused by human motion, enabling passive, contactless, and privacy-preserving activity detection. This approach allows sensing through obstacles without user-worn devices. However, CSI-based HAR faces challenges like ambient noise, temporal drift, and multipath interference, often degrading performance in dynamic multi-user environments with overlapping activities.

The limited availability of labelled datasets, especially for diverse user setups and infrequent activities, hinders the development of generalisable models [4]. Multi-user HAR is further complicated by simultaneous, unpredictable movements and spatial entanglement, demanding models capable of extracting relevant features from noisy, intertwined signals while generalising across various spatial arrangements. Traditional Machine Learning (ML), relying on handcrafted features, often proves insufficient in these complex scenarios. Deep learning, particularly CNNs and attention-based models, offers a promising alternative. However, many existing studies treat pre-processing, feature extraction, and classification as separate stages, potentially missing the advantages of an integrated approach.

Recent deep learning advances highlight the potential of hybrid architectures combining CNNs for local pattern recognition and Transformers for capturing long-range dependencies, although their application in multi-user, device-free HAR remains largely unexplored [5]. Transformers, especially with relative positional encoding, are well-suited for modelling the complex temporal relationships inherent in overlapping human activities [6]. To address these limitations, we propose a unified, augmentation-aware approach for CSI-based HAR, specifically designed to tackle the challenges of multi-user recognition and limited training data. Our approach integrates signal denoising, advanced data augmentation, domain-informed feature engineering, and deep multimodal learning, featuring a custom CNN + Transformer model to learn both local and global patterns from CSI signals, trained on an augmented dataset that mimics real-world activity variability.

The remainder of the paper is structured as follows: Section 2 reviews related work, Section 3 details our proposed model and augmentation methods, Section 4 describes the experimental setup and results, Section 5 provides a discussion of our findings, and Section 6 concludes the paper.

## 2 Related Work

Advances in HAR using CSI have seen a notable shift from conventional ML methods to advanced deep learning models, with Transformer-based architectures emerging as a powerful alternative in recent years. This is driven by the increasing need for scalable, device-free, and privacy-conscious activity recognition systems, especially in smart homes and Iot-enabled environments [2].

Advances in Transformer-based architectures are emerging in CSI-HAR by incorporating self-attention to capture long-range temporal dependencies, offering a robust alternative to recurrence [5]. For example, a multichannel attention-based Transformer achieved high accuracy in HAR and a lightweight Transformer optimised for edge computing maintained competitive accuracy (92.4%) with reduced complexity [9]. However, attention-only models can struggle with fine-grained local features to discriminate similar activities. To overcome this, hybrid CNN-Transformer architectures have emerged, combining CNNs for spatial pattern extraction with Transformers for global contextual modelling. Incorporating

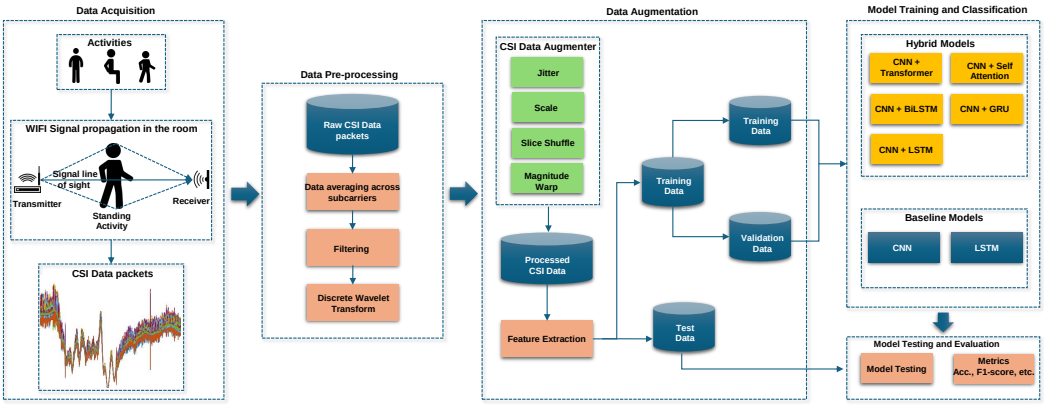


Figure 1: Overview of the proposed human activity recognition approach.

relative positional embeddings in these frameworks enhances temporal precision and activity segmentation [8]. These hybrid designs improve recognition accuracy and balance local and global sequence modelling, often a trade-off in Transformer-only systems.

The lack of consistent evaluation protocols in CSI-based HAR presents another significant challenge. The inconsistent use of evaluation protocols in the literature, including holdout sets, random splits, and k-fold cross-validation, impedes reproducibility and fair comparison of model performance across different CSI-based HAR studies. This is particularly problematic for benchmarking complex hybrid models where performance variability is more significant. Building upon recent advancements in hybrid deep learning for CSI-based HAR, a CNN-Transformer approach optimised for both single- and multi-user scenarios is proposed, utilising robust preprocessing and augmentation techniques to address these limitations and enhance accuracy, scalability, and real-world deployment potential.

## 3 Proposed Methods

The proposed HAR approach using CSI employs a four-stage process as depicted in Figure 1 designed to enhance recognition, especially in complex multi-user home environments. The pipeline includes data acquisition, pre-processing, augmentation, and model training/classification. Data acquisition involves Wi-Fi signal transmission, with motion-induced perturbations captured as CSI at the receiver, forming the raw dataset encoding static and dynamic movements.

A multi-stage preprocessing pipeline extracts meaningful information via subcarrier averaging, denoising, and Discrete Wavelet Transform. To enhance generalisability with limited CSI data, a random-transformation-based augmentation module generates diverse variations (jittering, scaling, slice shuffling, magnitude warping) while preserving activity semantics. The augmented data is then partitioned. The classification stage trains baseline (CNN, LSTM) and advanced hybrid deep learning models, CNN + Transformer, BiLSTM, GRU, LSTM to better capture spatial and temporal dependencies in CSI data, enhancing recognition of complex, multi-user activities.

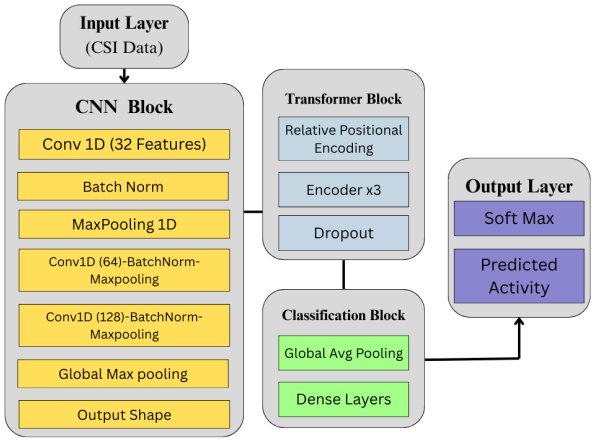


Figure 2: Architecture of the proposed time-series analysis model.

### 3.1 Data Acquisition

A publicly available CSI-based HAR dataset, specifically designed to investigate the feasibility of indoor activity recognition using fine-grained CSI from wireless signals, is utilised. The dataset was collected in a controlled laboratory environment, simulating a typical indoor room measuring 3 metres by 2.8 metres [7].

The experimental setup involved two Universal Software Radio Peripheral devices: a USRP X300 as the transmitter and a USRP X310 as the receiver. Both were equipped with VERT2450 omnidirectional antennas and operated at a frequency of 3.75 GHz within the 5G sub-6 GHz band. The transmitter and receiver are positioned diagonally opposite each other in the room to establish a reliable communication link and maximise spatial signal coverage. Within the defined activity zone, consisting of four chairs arranged in a 1-metre grid, participants performed various daily activities. During these activities, the USRP devices continuously collected CSI data using the GNU Radio software environment.

### 3.2 The Proposed CNN + Transformer Model

The proposed architecture shown in Figure 2 combines CNN and Transformer encoders for effective local and global feature extraction from time-series CSI data, capturing complex temporal features for HAR by using CNNs for local dependencies and Transformers for long-range context. The model has three parts: a CNN feature extractor, a Transformer encoder with relative positional encoding, and a classification head. Input sequence: $\mathbf{X} = \{x_0, x_1, x_2, \ldots, x_T\}, \quad x_t \in \mathbb{R}^d$ is reshaped to $(T, 1)$ for univariate processing.

**CNN Feature Extractor:** Three 1D convolutional layers (32, 64, 128 filters, kernel size 3, ReLU, Batch Normalisation, MaxPooling1D pool size 2) followed by Global Max Pooling and reshaping to (1,128).

**Transformer Encoder:** Vector-based relative positional encoding $\mathbf{X}_{\text{pos}} = \mathbf{X} + \mathbf{E}_{\text{pos}}$. Three encoder layers with multi-head self-attention - $\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$, residual connections, Layer Normalisation, and feed-forward layers (1D convolutions of size (256,d), ReLU).

**Classification Head:** Transformer output through Global Average Pooling and two Dense layers (256, 128 units, ReLU, Dropout 0.5 and 0.3). The final softmax output is $\hat{y} =$

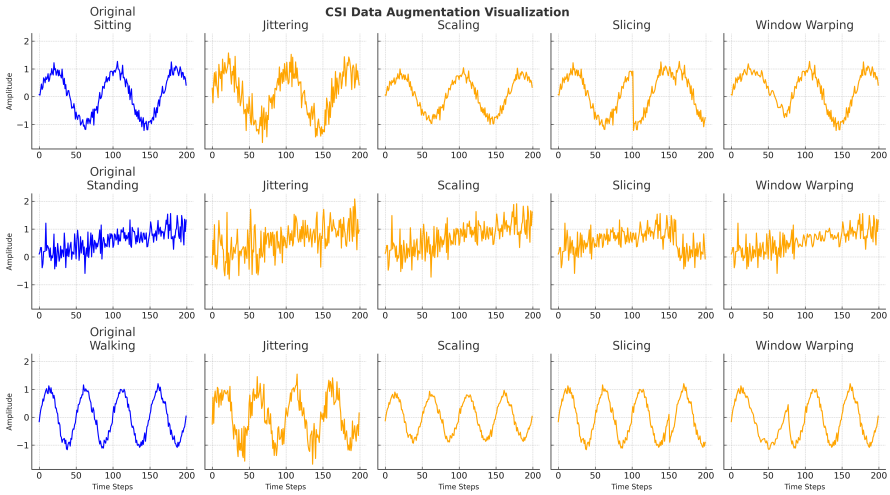


Figure 3: Augmentation impact of random-transformation techniques on the CSI data with each row displaying the original and augmented signals.

$\text{softmax}(\mathbf{W}x + b)$. This architecture provides an optimal balance between representational power and computational efficiency, making it particularly suitable for deployment in real-time or edge-based HAR systems.

# 4 Experiments and Results

The CNN + Transformer model was evaluated against baselines, focusing on data augmentation's impact, multi-user performance, and accuracy. The original dataset which contained 1777 instances was augmented as depicted in Figure 3 using factors up to 10. This increased the instances for example factor 3 yielded 5331 instances which introduced variability while preserving temporal structure for better generalisation. Following augmentation, 34 features per instance were extracted.

The model was evaluated using accuracy and F1-score to address class imbalance in CSI-based HAR datasets. F1-score provides a balanced performance assessment beyond potential skew in result due to class imbalance.

## 4.1 Experiment I: Sensitivity Analysis of Augmentation

Sensitivity analyses were conducted to compare the original and augmented dataset. The primary objective was to determine if augmentation introduces statistically significant deviations in CSI feature distributions. Towards this, the following hypothesis is formulated:

- Null Hypothesis ($H_0$): The distributions of the original and augmented data are statistically identical; i.e., augmentation does not significantly alter feature distributions.
- Alternative Hypothesis ($H_1$): The distributions of the original and augmented data differ significantly.

Normality tests were performed to determine the appropriateness of parametric versus non-parametric testing. The Mann–Whitney U test was used to examine whether they are

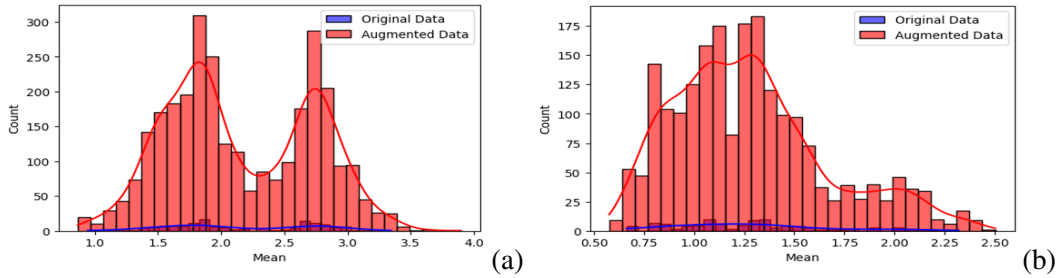


Figure 4: Sensitivity analysis of original and augmented CSI data for two multi-user activity scenarios.

Table 1: Mann–Whitney U and Levene's Test Results for Original vs. Augmented Data.

| Feature | Label | Mann–Whitney U | p-value | Levene's W | p-value | Interpretation |
|---|---|---|---|---|---|---|
| Mean | Empty | 144584.0 | 0.9139 | 0.1772 | 0.6739 | Fail to reject $H_0$ |
| Mean | 1Subject-1Sit | 206671.5 | 0.9325 | 0.0449 | 0.8322 | Fail to reject $H_0$ |
| Mean | 2Subjects-1Sit-1Stand | 105480.5 | 0.9384 | 0.0819 | 0.7747 | Fail to reject $H_0$ |
| Mean | 3Subjects-2Sit-1Stand | 105100.5 | 0.9871 | 0.0858 | 0.7696 | Fail to reject $H_0$ |
| Mean | 4Subjects-2Sit-2Stand | 106725.5 | 0.7811 | 1.0095 | 0.3151 | Fail to reject $H_0$ |

Table 2: Experiment II Results: Accuracy across different augmentation factors.

| Model | Aug. Factor | Phase 1 | Phase 2 | Phase 3 | Phase 4 | All Activities |
|---|---|---|---|---|---|---|
| CNN + Transformer | 0 | 0.869 | 0.840 | 0.899 | 0.869 | 0.757 |
| | 1 | 0.963 | 0.927 | 0.957 | 0.885 | 0.817 |
| | 3 | 0.988 | 0.970 | 0.976 | 0.938 | 0.910 |
| | 5 | 0.994 | 0.979 | 0.987 | 0.959 | 0.939 |
| | 7 | 0.994 | 0.986 | 0.990 | 0.965 | 0.954 |
| | 10 | 0.994 | 0.989 | 0.991 | 0.973 | 0.963 |
| CNN + BiLSTM | 0 | 0.807 | 0.852 | 0.885 | 0.871 | 0.766 |
| | 1 | 0.968 | 0.925 | 0.957 | 0.874 | 0.805 |
| | 3 | 0.987 | 0.968 | 0.974 | 0.933 | 0.908 |
| | 5 | 0.992 | 0.977 | 0.986 | 0.955 | 0.937 |
| | 7 | 0.992 | 0.984 | 0.990 | 0.964 | 0.948 |
| | 10 | 0.995 | 0.985 | 0.992 | 0.972 | 0.960 |
| CNN + GRU | 0 | 0.762 | 0.729 | 0.832 | 0.881 | 0.772 |
| | 1 | 0.965 | 0.932 | 0.960 | 0.875 | 0.799 |
| | 3 | 0.985 | 0.969 | 0.979 | 0.935 | 0.906 |
| | 5 | 0.993 | 0.978 | 0.985 | 0.954 | 0.930 |
| | 7 | 0.993 | 0.985 | 0.989 | 0.963 | 0.952 |
| | 10 | 0.994 | 0.987 | 0.991 | 0.973 | 0.958 |
| CNN + LSTM | 0 | 0.746 | 0.729 | 0.841 | 0.869 | 0.761 |
| | 1 | 0.962 | 0.927 | 0.952 | 0.872 | 0.794 |
| | 3 | 0.984 | 0.966 | 0.974 | 0.933 | 0.901 |
| | 5 | 0.990 | 0.973 | 0.983 | 0.954 | 0.927 |
| | 7 | 0.993 | 0.982 | 0.988 | 0.965 | 0.949 |
| | 10 | 0.995 | 0.985 | 0.989 | 0.973 | 0.959 |

statistically significant differences in the means of the two datasets. Levene's test was also used to assess the homogeneity of variances. As summarised in Table 1, the p-values from both tests exceeded 0.05 for all classes suggesting no statistically significant differences in central tendency or variance with "fail to reject" the null hypothesis in all cases. Figure 4 is included as a confirmation showing histograms for the two subjects (one sitting and one standing) and three subjects (two sitting and one standing) classes. While the augmented data (in red) displays broader tails, the distributions remain centred around the same mean as the original data (in blue), suggesting that augmentation preserves core statistical structure.

Table 3: Comparison of classification performance across different model architectures.

| Model Variant | Accuracy | F1-Score | Precision | Recall |
|---|---|---|---|---|
| CNN [7] | 0.857 | – | – | – |
| CNN | 0.893 | 0.891 | 0.893 | 0.892 |
| LSTM | 0.900 | 0.894 | 0.894 | 0.894 |
| Proposed CNN+Transformer | 0.939 | 0.933 | 0.940 | 0.934 |

### 4.2 Experiment II: Impact of Augmentation Factors

This experiment examined the effect of varying data augmentation factors (0-10) on the proposed CNN + Transformer and three baseline hybrid models (CNN + BiLSTM, CNN + GRU, CNN + LSTM). Table 2 shows the accuracy improved consistently from factor 1 to 5 across all models and activity phases, indicating the benefit of a moderate increase in CNN + Transformer: 0.757 to 0.939 at factor 5; CNN + GRU: 0.772 to 0.930 at factor 5. Performance plateaued beyond factor 5 as illustrated by the CNN + LSTM model, where accuracy increased only marginally from 0.927 at factor 5 to 0.959 at factor 10. Notably, no model performance declined by a factor of 10, confirming the robustness of our augmentation. The lowest performance at factor 0 (no augmentation) highlights data augmentation's crucial role in enhancing deep learning HAR models, especially for sparse or imbalanced datasets.

## 5 Discussion

The experimental results, from sensitivity analysis to varying augmentation factors, demonstrate our approach's capability in handling diverse data and complex multi-user environments. Combining augmentation and Transformer mechanism significantly enhances in modelling temporal dependencies and isolating key features. Multi-user experiments confirmed the CNN + Transformer model's robustness against signal interference and overlapping activities. In a 2-user setup, performance was good with high precision and recall, especially for dynamic activities like walking, reflecting the effective capture of temporal patterns by the hybrid CNN and Transformer layers.

Increased user configurations posed challenges, particularly for static activities (standing, sitting), which saw a precision drop, often misclassified, especially with fewer users. This likely resulted from reduced motion variance during concurrent sedentary activities, creating similar sensor patterns difficult even for the attention mechanism to differentiate, as seen in mixed-user settings. However, the Transformer model still effectively distinguished more kinetic activities like walking, showcasing its strength in capturing global temporal dependencies. This suggests the Transformer does well with dynamic activities in concurrent scenarios, but static activity classification with overlap requires further enhancement. Comparative analysis presented in Table 3 reinforces this. A standalone CNN achieved 89.3% accuracy, improving on prior CNN-only results (85.7% [7]). LSTM slightly improved to 90.0%. However, our CNN+Transformer achieved the highest performance: 93.9% accuracy, 93.3% F1-score, 94.0% precision, and 93.4% recall. This significant gain stems from the hybrid model's integration of spatial feature extraction with long-range temporal dependency learning, providing a balanced activity representation. The high F1-score indicates a strong balance between precision and recall, crucial for reliable HAR applications.

## 6 Conclusion

This paper presents a robust and scalable CSI-based HAR approach tailored for complex multi-user indoor environments. The method integrates multi-stage preprocessing, data augmentation, and a CNN + Transformer hybrid model, alongside other deep learning architectures, to effectively capture spatiotemporal CSI dependencies. Extensive experiments validate the approach's effectiveness. Sensitivity analysis confirmed that augmentation preserves the underlying data distribution. Moderate augmentation—specifically at factor 5 optimised model performance across all variants, with CNN + Transformer consistently outperforming others. Overall, the proposed CSI-based HAR system exhibits notable improvements in robustness, accuracy, and scalability, forming a strong foundation for future intelligent activity recognition systems in ambient settings. Key limitations include non-adaptive augmentation, lack of explicit user-level separation, and the Transformer's computational demands. Future work will focus on adaptive augmentation, finer-grained user identification, lightweight Transformer variants, and deployment across diverse real-world environments and populations.

## References

[1] A. Lotfi, C. Langensiepen, et al., "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour." *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, pp. 205–218 (2012).

[2] H. Chen, et al., "WiFi CSI-Based Activity Recognition Using Attention Mechanisms," *Sensors*, vol. 21, no. 2, p. 452, 2021.

[3] A. Naser, A. Lotfi, et. al, "Privacy-Preserving, Thermal Vision With Human in the Loop Fall Detection Alert System" *IEEE Trans. on Human-Machine Systems*, Vol. 53, no. 1, pp. 164-175, 2023.

[4] A. Doherty, H. Yuan, et. al, Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality, 2023.

[5] M. Ezzeldin, S. Ghoneim, et. al, 2024. Multi-modal hybrid hierarchical classification approach with transformers to enhance complex human activity recognition. *Signal, Image and Video Processing*, 18(12), pp.9375-9385, 2024.

[6] L. Wang, et al., "Privacy Considerations in HAR," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, pp. 1–26, 2021.

[7] A.M. Ashleibta, A. Taha, et. al, "5g-enabled contactless multi-user presence and activity detection for independent assisted living," *Sci. Reports*, 11(1), p.17590, 2021.

[8] H. Lee, J. Park, and Y. Nam, "Temporal positional encoding for HAR using CSI and transformer architectures," *IEEE Sensors J.*, vol. 24, no. 3, pp. 4567–4578, 2024.

[9] W. Han and J. Yu, "Algorithm for interference filtering of Wi-Fi gesture recognition," *Int. J. Inf. Commun. Technol.*, vol. 24, no. 1, pp. 1–20, 2024.

[10] Q. Chen, Y. Lin, and S. Zhao, "Entropy-based features for HAR using CSI," *Sensors*, vol. 18, no. 4, pp. 1223–1231, 2018.