Identifying Subject Bias in WiFi-based Human Activity Recognition Evaluation Methods

Amany Elkelany[1,2,*,†], Robert Ross[1,2,†] and Susan Mckeever[1,2,†]

[1] *Technological University Dublin, Ireland*
[2] *ADAPT Research Centre, Ireland*

Abstract

WiFi-based Human Activity Recognition (HAR) has emerged as a promising approach for monitoring and analysing human activities in a non-intrusive manner, leveraging WiFi signals for activity classification. Despite advancements, existing WiFi-based HAR research lacks consideration of subject (human) bias. This results in learning models performing well on individuals used in the training samples but failing to generalise to new/unseen subjects, in contrast to known good practices in machine learning. In this paper, we address this oversight directly by systematically examining the evaluation methodology for the WiFi-based HAR context. Specifically, we investigate the impact of Leave-One-Subject-Out Cross-Validation (LOSOCV) in a hybrid architecture combining Convolutional Neural Networks (CNN) and Attention-based Bidirectional Long Short-Term Memory networks (ABiLSTM), designed to capture both spatial and temporal patterns in WiFi signals. However, our emphasis remains on the application of LOSOCV as a method for improving generalization and reducing subject bias, rather than on the architecture itself. The model's effectiveness is evaluated using LOSOCV, and we compare its performance against conventional hold-out validation and k-fold validation. Additionally, we utilize weighted metrics for model evaluation to address class imbalance, ensuring a fair assessment across all activity categories. Our results demonstrate the importance of LOSOCV in providing a realistic assessment of HAR model performance and underscore that addressing subject bias is essential for the deployment of these systems in practical scenarios such as healthcare monitoring, smart homes, and security applications.

Keywords

WiFi, Human Activity Recognition (HAR), Channel State Information (CSI), Deep Learning, Convolutional Neural Network (CNN), Bidirectional Long Short Term Memory (BiLSTM), Subject Bias

1. Introduction

Human Activity Recognition (HAR) has steadily emerged as one of the most prominent research areas using different sensing technologies. HAR is involved in many applications including healthcare [1, 2], fitness tracking [3, 4], elderly people care [5], and security and surveillance [6]. HAR techniques can be separated into three categories based on the type of technology used in the data collection: vision-based, sensor-based (including wearable sensors), and WiFi-based.

Sensor-based HAR uses sensors like accelerometers, gyroscopes, and wearable devices. Wearable devices such as smartphones, and smartwatches are costly, privacy-intrusive, and inconvenient to wear for some people, while HAR accuracy is affected by placement and calibration challenges [7]. Vision-based approaches use static cameras or built-in camera devices, but they have limitations due to privacy intrusion, high energy consumption, lighting changes, camera perspectives, and background clutter [8, 9]. Recent years have seen a significant increase in interest in WiFi sensing applications due to the ubiquitous use of WiFi and the advancement of wireless communication technology [10].

WiFi signals have emerged as a leading technology in HAR applications due to their advantages in privacy-preservation, low cost, and as well their its potential for passive environmental deployment [11]. WiFi signals can be analysed in a number of different ways, one prominent way is through the use of Channel State Information (CSI) [12]. A CSI sample is a 2D matrix that captures the temporal and spatial dynamics of the environment [13]. Each CSI sample captures the amplitude and phase information

---

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland

[*] Corresponding author.

[†] These authors contributed equally.

0000-0002-4378-9270 (A. Elkelany); 0000-0001-7088-273X (R. Ross); 0000-0003-1766-2441 (S. Mckeever)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

across multiple subcarriers and antennas over time, reflecting how the wireless signal changes as it passes through or is obstructed by objects and people in the environment [14]. These variations in the signal provide a rich source of information that can be used to identify different human activities, such as walking, sitting, falling, or running. By analysing these CSI patterns, machine learning (ML) models can be trained to recognize and classify human activities with high accuracy, even in environments where direct visual observation is not possible, making WiFi-based HAR a powerful and non-intrusive method for monitoring human behaviour.

A key challenge for WiFi-based HAR systems is their generalization capability across diverse subjects (individuals). This capability is essential for any HAR system intended for large-scale or real-world applications, as it must accurately recognize activities from new subjects without requiring the collection of labelled data for each new user or retraining. In the context of WiFi signal monitoring, subject bias is particularly significant, given that the physical characteristics of individuals—such as body size, shape, age, gender, and even clothing—can profoundly impact signal propagation in monitored spaces, as reported in [15]. Consequently, learning models trained on CSI measurements collected using a limited number of individuals may struggle to perform well on unseen subjects, as these models often capture subject-specific patterns that contribute to subject bias. This bias can lead to substantial variations in system performance, depending on individual characteristics and movement patterns, which has been highlighted in previous studies [16, 17].

Learning Model evaluation plays a crucial role in assessing the effectiveness of WiFi-based HAR systems and in selecting the most suitable model architecture [17, 16]. WiFi-based HAR studies typically adopt either a single model or a subject-specific model approach as in [18, 19, 20, 21]. The single model approach involves building one model using data from all subjects, whereas the subject-specific approach creates an individual model for each subject. In both scenarios, traditional hold-out or k-fold cross-validation methods are used to evaluate the models. However, a significant limitation of these conventional evaluation methods is that data from the same individual appear in both the training and testing sets.

Previous works like Wi-Motion [18], adaptive antenna elimination-based model [22], Wi-Sense [19], STC-NLSTMNet [23], THAT [24], ViT based HAR [25] demonstrated high accuracies up to 99.88% using various learning models like support vector Machine (SVM), random forest (RF), convolutional neural networks (CNN), spatio-temporal convolution with nested LSTM (STC-NLSTM), convolution augmented transformer and vision transformers architectures (ViT). However, these models are evaluated under conditions where all subjects are part of the training dataset using either traditional validation or k-fold cross-validation, leading to potentially inflated performance metrics as it does not adequately test the models' ability to generalise to new or unseen subjects. This absence of evaluating the learning models on new/unseen subjects raises concerns regarding the generalization capabilities of those models to new/unseen subjects, as it fails to account for variability among subjects and the impact this variability may have on model performance when applied to unseen individuals. Addressing subject bias is crucial for ensuring that WiFi HAR models are robust and accurate across diverse users, making them reliable for real-world applications.

While the importance of personalized models has been acknowledged [26], the ability of models to generalise to new subjects remains important. One of the most effective ways to evaluate generalization is through Leave-One-Subject-Out Cross-Validation (LOSOCV). LOSOCV rigorously tests the model's ability to generalise by training it on data from all but one subject and then evaluating it on the excluded subject. This process is repeated for each subject in the dataset. By doing so, the model is exposed to the data from every subject but evaluated in a way that simulates real-world scenarios where new, unseen subjects would need to be recognized. This paper explores the significance of LOSOCV in WiFi-based HAR and how it can ensure models perform well across diverse subjects, addressing the pressing need for subject generalization in non-intrusive activity recognition systems. Consequently, our contributions to this work are as follows:

- First, we review WiFi-based HAR research from the last three years. This review reveals that only one paper has evaluated models using LOSOCV, highlighting a significant gap in the adoption of subject-independent evaluation techniques within the field. This raises concerns about the inflated accuracy of the majority of WiFi-based HAR models reported in the literature.

- Second, we propose a WiFi-based HAR model to detect and classify activities, with particular emphasis on generalization across different subjects.
- Third, we perform a comprehensive comparison between LOSOCV and non-LOSO evaluation methods using two public WiFi-based HAR datasets, quantifying the impact of subject bias and demonstrating the advantages of subject-independent evaluation.

2. Related Work

In the context of WiFi-based HAR, subject bias arises when the performance of the recognition system varies significantly based on the specific characteristics of the individuals being monitored. This can include factors such as physical characteristics (e.g., height, weight, body composition), movement patterns, and environmental context.

For example, WiFi-based HAR systems can be used for monitoring elderly patients or people with mobility issues. Subject bias is particularly problematic here because patients may exhibit distinct movement patterns depending on their physical conditions, leading to inaccuracies in activity recognition if the model was trained on younger, healthier individuals.

Another example is WiFi-based HAR systems can be used in fitness centres or at home to track and analyse users' physical activities. Subject bias becomes an issue because individuals have different fitness levels, body types, and workout styles. A model trained on a small subset of users might struggle to accurately track exercises for users with different movement dynamics, potentially providing inaccurate feedback on their performance.

Therefore, addressing subject bias is essential to ensure WiFi HAR models are robust, accurate, and generalised across different users, making them reliable for real-world applications.

We reviewed the literature concerning WiFi-based HAR published in the last three years. We analysed the evaluation methods used across different studies. Our focus is on understanding the variety of validation techniques employed to assess the models' ability to generalise, particularly when new or unseen subjects are involved. We classify these evaluation approaches into four categories [17, 16]: Hold-out Validation (HO), k-fold cross-validation (k-fold CV), Leave-One-Subject-Out (LOSO) validation, and Leave-One-Subject-Out Cross-Validation (LOSOCV). However, not all of them are well-suited for testing generalization across subjects. Notably, only three papers utilize LOSO alongside HO and k-fold CV techniques in their evaluations, while only one paper applies the LOSOCV method in addition to the k-fold CV technique.

Table 1 summarizes the WiFi-based HAR previous studies in the last 3 years, highlighting the publication year, the number of subjects involved, the evaluation methods applied, and the accuracy reported in each study. This table serves as an overview of the state-of-the-art evaluation practices in the WiFi-based HAR domain during the last 3 years. The four evaluation techniques can be explained as follows [17, 16]:

- Hold-out Validation (HO): This method splits the dataset into training and testing sets, but if the subjects in the training and test sets overlap, the model may perform well due to memorizing subject-specific patterns, rather than generalizing to new individuals. The HO technique requires less computational power as it only runs a single time; however, if the data is split again, the model's outcomes are likely to vary. The HO technique was extensively employed in numerous WiFi-based HAR studies [8, 20, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37].
- K-Fold Cross Validation (k-fold CV): In k-fold cross-validation, the dataset is divided into k subsets (or "folds"), and the model is trained k times, each time using k-1 folds for training and the remaining fold for testing. While this ensures that each data point is used for both training and testing, it often does not fully account for subject diversity because subjects can be included in both the training set and testing set. The k-fold CV technique was utilized in numerous WiFi-based HAR studies [22, 21, 34, 38, 39, 40, 41, 42, 43, 44, 45, 46].

- Leave-One-Subject-Out (LOSO): It is also called a subject-specific validation, a single, specific subject is selected for testing, while the model is trained on data from all other subjects. This method aims to measure how well the model can generalise to a specific unseen subject. Unlike methods like k-fold or hold-out validation, subject-specific validation guarantees that the selected subject's data is excluded from the training process, which helps assess the model's true ability to generalise to a new individual. By focusing on just one specific subject for testing, this method does not provide a comprehensive view of how well the model generalises to a broader population. The performance may vary significantly among different subjects, and a single evaluation may not capture these variations. This could lead to misleading conclusions about the model's performance. The LOSO technique is utilized in a limited number of WiFi-based HAR studies [47, 32, 37].
- Leave-One-Subject-Out Cross Validation (LOSOCV): This technique is particularly valuable for evaluating generalization in WiFi-based HAR. In LOSOCV, the model is trained on data from all subjects except one and then tested on the excluded subject. This process is repeated for each subject in the dataset. LOSOCV provides a rigorous evaluation of the model's ability to generalise to unseen subjects, which is crucial for the real-world deployment of HAR systems. Unlike other techniques, LOSOCV prevents data from the same subject from appearing in both the training and test sets, ensuring that the model is not simply learning subject-specific features. LOSOCV technique is only applied in one paper [46].

3. Methodology

In this section, we present our proposed approach for HAR using WiFi signals which we used to investigate the impact of different evaluation techniques and the influence of subject bias. Our approach involves using two distinct public datasets to comprehensively assess model performance. The approach includes a data collection phase, data pre-processing methods, the development of an activity classifier and the evaluation. By evaluating the activity classifier performance with HO, k-fold CV and LOSOCV, we aim to determine the extent of subject bias present and evaluate the model's ability to generalise to new users. This approach provides a rationale for understanding how different evaluation techniques affect the robustness and reliability of HAR systems.

3.1. Data Collection

In our experimental evaluation, we used two distinct public datasets that have multiple rooms or environments: GJWiFi[48] and OPERAnet [29]. The GJWiFi is a dataset for WiFi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments. This dataset was collected at the German Jordanian University. So, We refer to this dataset as GJWiFi [2]. GJWiFi dataset was gathered across three distinct spatial environments: a Laboratory (denoted as E1), a Hallway (denoted as E2), and a Hybrid environment (denoted as E3) that combines the Laboratory and Hallway environments with an 8 cm thick barrier in between. Environments E1 and E2 are configured for Line-of-Sight (LOS) conditions, while E3 is set up for Non-Line-of-Sight (NLOS) conditions.

The authors of [48, 49], who are part of the dataset authors' group, identified 12 activity classes across the five sessions and consolidated them into six labels for HAR. In this work, we adopted this six-class labelling approach, following the data providers' method. The six activities classes are no movement, sitting down / standing up, walking, turning and picking up a pen from the ground. In each environment, 10 subjects voluntarily participated in the data collection, performing each activity 20 times. Each received packet contains 90 complex CSI values (1 transmit antennas x 3 receive antennas x 30 subcarriers). GJWiFi has been used widely in previous works [49, 36, 22, 20].

---

[1] WiFi-based HAR publicly available datasets with links are available in this URL: https://github.com/amakelany/Public-WiFi-Based-HAR-datasets.

[2] The GJWiFi dataset directories are provided by the original authors in this URL: https://data.mendeley.com/datasets/v38wjmz6f6/1

Table 1

Summary of WiFi-based HAR Studies in the recent 3 years, highlighting the lack of subject bias considerations. SC is an abbreviation for a self-collected dataset by the authors of the paper cited, which is private. StanWiFi, Wiar, GJWiFi, ARIL, Widar3.0, NTU-FI, UT-HAR, SignFi, SAR, CSI HA, CSI HAR, and 5G-HAR are names of WiFi-based HAR publicly available datasets[1].

Ref.	Publication Year	Number of Subjects	Evaluation Method	Performance (Accuracy %)	Addresses Subject Bias?
[43]	April 2021	6 in StanWiFi, 7 in SC dataset	10-fold CV	StanWiFi → 97.34, SC dataset → 98.95	No
[24]	May 2021	6	HO	98.55	No
[44]	Oct. 2021	3	10-fold CV	96.55	No
[35]	Oct. 2021	6	HO	85	No
[8]	Oct. 2021	3	HO	95	No
[45]	Nov. 2021	6 in StanWiFi, 5 in SignFi	10-fold CV → StanWiFi, 5-fold CV → SignFi	ARIL → 98.20, StanWiFi → 98, SignFi→ 95.42	No
[46]	**Jan. 2022**	**20**	**10-fold CV and LOSOCV**	**10-fold CV → 94.00, LOSOCV → 91.27**	**Yes**
[34]	Jan 2022	6	10-fold CV, and HO	10-fold CV → 99.33, HO → 100	No
[33]	Feb 2022	9	HO	90	No
[47]	Feb. 2022	5 in SignFi 4 in SC dataset	HO and LOSO	SignFi →92.80, SC dataset → 93.92	No
[28]	March 2022	Not stated	HO	98.10	No
[42]	April 2022	6	10-fold CV	92	No
[27]	Aug. 2022	20	HO	98	No
[29]	Aug. 2022	6	HO	93.50	No
[37]	Sept. 2022	5	HO and LOSO	HO → 93.60, LOSO → 92.88	No
[30]	Feb. 2023	30	HO	LOS → 96.39, NLOS → 95.09	No
[36]	Apr 2023	10 in Wiar 30 in GJWiFi	HO	Wiar → 96, GJWiFi → 94.33	No
[40]	July 2023	3 in CSI-HAR, 6 in StanWiFi	5-fold CV	CSI-HAR → 99.62 , StanWiFi → 97.88	No
[31]	July 2023	10 in Wiar, 9 in SAR, 5 in Widar3.0	HO	Wiar → 99.40, SAR → 99.30, Widar 3.0 →99.30	No
[21]	July 2023	3 in CSI HAR, 1 in CSI HA, 4 in 5G-HAR	5-fold CV	CSI HAR → 97.90 CSI HA → 98.30 5G-HAR → 98.60	No
[41]	Aug. 2023	2	10-fold CV	83.39	No
[32]	Aug. 2023	5 in SignFi, 1 in ARIL, 3 in CSI-HAR	HO and LOSO	SignFi → 93.50 , ARIL → 97.50 , CSI-HAR → 99.50	No
[22]	Sept. 2023	6 in StanWiFi, 30 in GJWiFi	10-fold CV	StanWiFi → 99.84 , GJWiFi(LOS) → 97.65, GJWiFi(NLOS)→ 93.33	No
[38]	Jan 2024	6	10-fold CV	98.44	No
[25]	March 2024	6 in UT-HAR, 20 in NTU-FI,	HO	UT-HAR → 98.78, NTU-Fi → 98.20	No
[20]	June 2024	20 in NTU-FI, 10 in Wiar, 30 in GJWiFi	HO	NTU-Fi→ 99.82, Wiar → 99.56, GJWiFi → 99.10	No
[39]	July 2024	64	10-fold	99.42	No

We also used the public dataset called OPERAnet[3] which is described in detail in [29] to train our models. While OPERAnet includes different RF signals other than the CSI, in this work, we will use only the CSI measurements extracted from the WiFi signals. OPERAnet includes samples for six activities: walking, sitting on a chair, standing from a chair, lying down on the floor, standing up from the floor, and rotating the upper half of the body. Six subjects of different ages conducted these activities. OPERAnet dataset includes CSI measurements for two different furnished rooms, with desks, chairs, screens, and other office objects lying in the surroundings. Additionally, It is used by other researchers for HAR [50, 29]. The WiFi signals were collected from the two rooms using two receivers: the LOS receiver (NUC1), and the NLOS receiver (NUC2) placed in a bi-static configuration (90°) with respect to the transmitter. Each received packet contains 270 complex CSI values (3 transmit antennas x 3 receive antennas x 30 subcarriers).

3.2. Data Preprocessing

Preprocessing WiFi signals is a crucial step performed on raw data before feeding it into the training model, as highlighted in [13]. This data preprocessing consists of four main stages: CSI extraction, data denoising, normalization and windowing.

The first stage is CSI extraction from WiFi signal packets. Channel State Information (CSI) values are the complex values which represent amplitude attenuation and phase shift. For the training process, only the amplitude is considered, as noted in [51, 52], leading to the conversion of these complex values into real values.

The second stage is data denoising, which involves applying a Hampel filter [53] to both datasets for outlier detection and removal in each CSI sequence, resulting in denoised sequences, as described in [18]. The GJWiFi dataset has a sampling rate of 320 packets per second, while the OPERAnet dataset has a significantly higher sampling rate of 1600 packets per second. Since learning models trained on high-frequency data are more prone to capturing noise rather than meaningful patterns, we downsampled the OPERAnet dataset to 320 packets per second as in [8]. This downsampling not only reduces noise but also mitigates the risk of overfitting by simplifying the CSI measurements, allowing the model to focus on the underlying patterns.

In the third stage, the denoised sequences from both datasets are normalized using Min-Max normalization to ensure that all CSI measurements have a consistent scale. By doing so, the normalization process eliminates any discrepancies caused by differing value ranges across datasets, thus preventing these variations from impacting the recognition accuracy of the model [54].

Lastly, in the windowing stage (data transformation or segmentation), the resultant data samples are divided into windows, similar to HAR studies [17, 55, 51]. Each window has a size of one second and is labelled with the most frequently occurring label within its samples. Additionally, a 10% overlap between consecutive windows is utilized to ensure that each row in the transformed vector incorporates information from the preceding window, thereby capturing more continuous and detailed temporal dependencies [56].

3.3. Activity Classification

Inspired by the deep learning model's architecture presented in [51, 8, 57, 58], we apply a model which combines the Attention-Based Bidirectional Long Short-Term Memory (ABiLSTM) network and Convolutional Neural Networks (CNN) to capture both local spatial patterns and long-term dependencies in the data. We refer to this model as CNN-ABiLSTM. The CNN-ABiLSTM model Leverage the strengths of attention and sequential learning to improve performance in activity classification. The CNN component excels at capturing local spatial patterns and features within the data, making it particularly effective for tasks that involve analysing structured inputs such as time series or image data. The addition of the BiLSTM layer enables the model to maintain and leverage temporal information from

---

[3] The OPERAnet dataset directories are provided by the original authors in this URL https://doi.org/10.6084/m9.figshare.c.5551209.v1.

the sequence, while the attention mechanism further refines this process by emphasizing key parts of the input. Therefore, CNN-ABiLSTM efficiently learns both local features and long-term dependencies, all while focusing on the most relevant segments of the input data. To address the issue of overfitting, key techniques such as dropout layers and early stopping are incorporated into the CNN-ABiLSTM.

3.4. Data Splitting and Experimental Setup

We will compare the performance of the LOSOCV approach with that of HO validation and 10-fold cross-validation to evaluate the robustness of the model to generalise to new subjects. We evaluated the CNN-ABiLSTM model in each environment in the GJWiFi dataset and OPERAnet dataset independently. The Models were implemented using TensorFlow and trained using the Adam optimizer with a batch size of 64 and an initial learning rate of 10^{-3} to minimize the loss function. To ensure a more balanced evaluation, we employ weighted precision, recall and F1-score metrics for model assessment. These metrics address class imbalance by assigning weights to each class according to its frequency in the dataset.

4. Results

The results presented in Table 2 showcase the performance of the CNN-ABiLSTM model on the GJWiFi dataset under three different evaluation methods: HO with an 80% training and 20% testing split, 10-fold CV, and LOSOCV. The performance metrics for the 10-fold CV are averaged across the folds, while the LOSOCV results are averaged across the number of subjects. This comparative analysis provides insight into how different validation techniques influence the model's precision, recall, and F1-score, ultimately highlighting the strengths and limitations of each approach in the context of WiFi-based HAR.

In the HO evaluation, the model achieves high precision 97.96%, recall 97.83%, and F1-score 97.42%. These values indicate the model performs well on unseen data when a simple train-test split is used.

The results using a 10-fold CV indicate the model's most consistent performance across the three environments with equal average precision, recall, and F1-score all achieving 99.40%. The model benefits from being trained and validated multiple times over different splits, which likely mitigates the risk of overfitting and underfitting. These results suggest that, in WiFi-based HAR, a 10-fold CV is particularly effective for evaluating the model's generalisation capabilities across different activities.

The LOSOCV method, which specifically tests the model's ability to generalise across different subjects, shows a noticeable decrease in performance, with average precision, recall, and F1-score values of 95.69%, 95.39%, and 94.92%, respectively. The reduction in performance metrics underscores that the model may be capturing subject-specific patterns rather than generalised features of the activities. The lower performance of CNN-ABiLSTM across the three evaluation techniques on E3 compared to E1 and E2 can be attributed to E3 being conducted in an NLOS environment, where signal attenuation and multipath effects lead to increased noise and variability in the WiFi signals. This results in reduced model accuracy, as the features used for classification become less consistent and reliable in NLOS conditions.

The results from the OPERAnet dataset presented in Table 3, highlight the performance of the proposed CNN-ABiLSTM model across different evaluation methods: HO, 10-fold CV, and LOSOCV under both LOS and NLOS scenarios. The precision, recall, and F1-score metrics indicate that the model achieves its highest performance using a 10-fold CV, which shows superior results in both Room 1 and Room 2 compared to the other evaluation methods.

In the LOS scenario, the 10-fold CV method achieves the highest F1-scores, with 98.12% in Room1 and 96.02% in Room2. This performance surpasses that of the HO method, which records F1-scores of 97.19% in Room1 and 93.98% in Room2. LOSOCV, designed to mitigate subject bias, shows lower F1-scores of 93.10% in Room1 and 91.78% in Room2.

For the NLOS scenario, the HO method shows F1-scores of 93.89% in Room1 and 93.67% in Room2. The 10-fold CV method again shows superior performance with average F1-scores of 96.32% in Room1 and 95.54% in Room2. On the other hand, LOSOCV records even lower F1-scores of 91.23% in Room1

Table 2
Results of GJWiFi dataset using HO, 10-fold CV and LOSOCV.

Metric	Evaluation Method	E1	E2	E3	Average
Precision%	HO	98.22	98.27	97.38	97.96
	10-fold CV	**99.71**	**99.34**	**99.16**	**99.40**
	LOSOCV	96.87	96.32	93.87	95.69
Recall%	HO	97.90	98.25	97.35	97.83
	10-fold CV	**99.71**	**99.33**	**99.16**	**99.40**
	LOSOCV	96.84	95.90	93.43	95.39
F1-Score%	HO	97.67	98.24	96.36	97.42
	10-fold CV	**99.71**	**99.33**	**99.16**	**99.40**
	LOSOCV	96.24	95.75	92.77	94.92

and 90.04% in Room2. The drop in performance across both rooms using HO, 10-fold CV and LOSOCV in the NLOS scenario compared to the LOS scenario is due to signal variations and obstructions in the NLOS setting.

Table 3
Results of OPERAnet dataset using HO, 10-fold CV and LOSOCV.

Receiver	Metric	Evaluation Method	Room 1	Room 2	Average
LOS	Precision%	HO	97.34	94.10	95.72
		10-fold CV	**98.19**	**96.45**	**97.32**
		LOSOCV	93.67	92.60	93.14
	Recall%	HO	97.20	94.00	95.60
		10-fold CV	**98.12**	**96.12**	**97.12**
		LOSOCV	93.84	92.34	93.09
	F1-Score%	HO	97.19	93.98	95.59
		10-fold CV	**98.12**	**96.02**	**97.07**
		LOSOCV	93.10	91.78	92.44
NLOS	Precision%	HO	94.56	93.78	94.17
		10-fold CV	**96.68**	**95.73**	**96.21**
		LOSOCV	92.23	91.20	91.72
	Recall%	HO	94.40	93.48	93.94
		10-fold CV	**96.98**	**95.88**	**96.43**
		LOSOCV	91.34	90.72	91.03
	F1-Score%	HO	93.89	93.67	93.78
		10-fold CV	**96.32**	**95.54**	**95.93**
		LOSOCV	91.23	90.04	90.64

5. Discussion

The primary goal of this study is to assess how model selection and preprocessing influence an ML model's ability to classify activities for new users. To achieve this, LOSOCV was employed as the evaluation method. A comparison between the results obtained using HO, 10-fold cross-validation and LOSOCV for the same model reveals that the three approaches yield significantly different results.

A critical insight from the evaluation is the implications of subject bias on the model's performance. The fact that 10-fold CV outperforms both HO and LOSOCV is significant. However, HO achieves relatively high F1-scores across different environments in GJWiFi and OPERAnet datasets. It can lead to inflated performance metrics due to its reliance on a limited training set. The 10-fold CV approach benefits from the diversity of the training set, effectively capturing a wide range of activity patterns and reducing the risk of overfitting to specific individuals. While LOSOCV, though designed to mitigate subject bias, still shows lower scores across the metrics. The performance drop observed in LOSOCV indicates that when the model is evaluated on unseen subjects, it struggles to maintain the same level of accuracy achieved during training. This observation underscores the challenges of achieving true generalization in HAR systems. This arises from individual differences in how activities are performed or variations in the WiFi signal received by different users. Consequently, while LOSOCV is a valuable method for assessing generalization, it may not completely eliminate subject bias, particularly if the training data does not adequately represent the diversity of potential users.

It is important to note that LOSOCV is primarily an evaluation technique rather than a solution for subject bias itself. To address subject bias more effectively, a comprehensive set of guidelines for WiFi-based HAR systems can be considered. These guidelines should prioritize the creation of a diverse training set that captures a wide range of user behaviours and environmental conditions, ensuring that the model is exposed to varied examples of activity patterns. From a statistical standpoint, methods such as data re-weighting and stratified sampling should be employed to balance the representation of different user groups within the training data. This ensures that the model is not disproportionately influenced by specific individuals or environmental contexts, ultimately fostering more equitable predictions across all user groups. From an ML perspective, several advanced techniques can be incorporated to reduce bias and improve generalization. Specifically, domain adaptation and transfer learning can be leveraged to minimize performance disparities across different demographic groups. These approaches enable a model to transfer knowledge gained from one dataset to another, effectively reducing bias by allowing the model to adapt to new users or environments more efficiently. In addition, fairness-aware algorithms can be integrated into the model's learning process to address any disparities in performance between different groups. For example, fairness constraints could be applied during training to minimize the performance gap between users with varying WiFi signal strengths, physical attributes, or environmental conditions. This ensures that the model's predictions are equitable and do not favour specific subgroups. Moreover, techniques like data augmentation could be explored to mitigate subject bias further. By generating synthetic data that simulates the behaviour of diverse users in various scenarios, data augmentation enhances the model's ability to generalise to previously unseen subjects. This is especially important in HAR systems, where limited data from diverse user groups may otherwise lead to overfitting or underperformance on new users. These methods collectively contribute to building more robust, fair, and generalizable models that can perform reliably across different user demographics and real-world settings.

6. Conclusion and Future Work

In this study, we investigated the impact of different evaluation techniques on the performance of WiFi-based HAR systems, focusing on addressing the issue of subject bias. We proposed a CNN-ABiLSTM model and evaluated its performance using LOSOCV to measure its generalization capability across unseen users. The results indicate that LOSOCV offers a more realistic assessment of model performance compared to traditional evaluation methods such as hold-out validation and k-fold cross-validation, which often fail to account for subject bias. Our findings emphasize the need for subject-independent evaluation in WiFi-based HAR systems, as conventional methods may lead to inflated performance metrics that do not reflect real-world applicability. This work highlights the critical role of evaluation methodologies in developing robust and generalizable HAR systems for real-world scenarios.

A key area for future investigation is the impact of different preprocessing methods on the ability of the learning model to classify activities for new users. Additionally, exploring more techniques to address subject bias will be crucial. This includes employing domain adaptation and transfer learning strategies to enable the model to generalise better across diverse user populations, as well as utilizing ensemble learning to combine predictions from multiple models. By pursuing these research directions, including both improved preprocessing methods and strategies to mitigate subject bias, we can enhance the robustness and generalizability of WiFi-based HAR systems, ultimately contributing to more reliable applications in diverse real-world settings.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT, SFI Research Centre for AI-Driven Digital Content Technology at Technological University Dublin.

References

[1] J. C. Soto, I. Galdino, E. Caballero, V. Ferreira, D. Muchaluat-Saade, C. Albuquerque, A survey on vital signs monitoring based on Wi-Fi CSI data, Computer Communications 195 (2022) 99–110.

[2] I. Galdino, J. C. Soto, E. Caballero, V. Ferreira, T. C. Ramos, C. Albuquerque, D. C. Muchaluat-Saade, EHealth CSI: A Wi-Fi CSI Dataset of Human Activities, IEEE Access 11 (2023) 71003–71012.

[3] X. Guo, M. Choo, J. Liu, C. Shi, H. Liu, Y. Chen, M. C. Chuah, Device-free Personalized Fitness Assistant Using WiFi, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, ACMPUB27New York, NY, USA, 2018, pp. 1–23.

[4] Y. Zhu, D. Wang, R. Zhao, Q. Zhang, A. Huang, FitAssist: Virtual fitness assistant based on WiFi, in: Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Association for Computing Machinery, 2019, pp. 328–337.

[5] H. Boudlal, M. Serrhini, A. Tahiri, A Monitoring System for Elderly People Using WiFi Sensing with Channel State Information, International Journal of Interactive Mobile Technologies (iJIM) 17 (2023) 112–131.

[6] T. Wang, D. Yang, S. Zhang, Y. Wu, S. Xu, Wi-Alarm: Low-Cost Passive Intrusion Detection Using WiFi, Sensors 19 (2019) 2335.

[7] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, H. Moon, Sensor-based and vision-based human activity recognition: A comprehensive survey, Pattern Recognition 108 (2020) 107561.

[8] P. F. Moshiri, R. Shahbazian, M. Nabati, S. A. Ghorashi, A CSI-Based Human Activity Recognition Using Deep Learning, Sensors 2021, Vol. 21, Page 7225 21 (2021) 7225.

[9] M. Al-Faris, J. Chiverton, D. Ndzi, A. I. Ahmed, A Review on Computer Vision-Based Methods for Human Action Recognition, Journal of Imaging 2020, Vol. 6, Page 46 6 (2020) 46.

[10] K. Pahlavan, P. Krishnamurthy, Evolution and Impact of Wi-Fi Technology and Applications: A Historical Perspective, International Journal of Wireless Information Networks 28 (2021) 3–19.

[11] Z. Hussain, Q. Z. Sheng, W. E. Zhang, A review and categorization of techniques on device-free human activity recognition, Journal of Network and Computer Applications 167 (2020) 102738.

[12] A. Khalili, A.-H. Soliman, M. Asaduzzaman, A. Griffiths, Wi-Fi sensing: applications and challenges, The Journal of Engineering 2020 (2020) 87–97.

[13] Y. Ma, G. Zhou, S. Wang, WiFi Sensing with Channel State Information: Survey, ACM Computing Surveys (CSUR) 52 (2019) 1–36.

[14] J. Liu, G. Teng, F. Hong, Human Activity Sensing with Wireless Signals: A Survey, Sensors 2020, Vol. 20, Page 1210 20 (2020) 1210.

[15] Y. Liang, W. Wu, H. Li, F. Han, Z. Liu, P. Xu, X. Lian, X. Chen, WiAi-ID: Wi-Fi-Based Domain Adaptation for Appearance-Independent Passive Person Identification, IEEE Internet of Things Journal 11 (2024) 1012–1027.

[16] H. Bragança, J. G. Colonna, H. A. Oliveira, E. Souto, How Validation Methodology Influences Human Activity Recognition Mobile Systems, Sensors 2022, Vol. 22, Page 2360 22 (2022) 2360.

[17] D. Gholamiangonabadi, N. Kiselov, K. Grolinger, Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection, IEEE Access 8 (2020) 133982–133994.

[18] H. Li, X. He, X. Chen, Y. Fang, Q. Fang, Wi-Motion: A Robust Human Activity Recognition Using WiFi Signals, IEEE Access 7 (2019) 153287–153299.

[19] M. Muaaz, A. Chelli, M. W. Gerdes, M. Pätzold, Wi-Sense: a passive human activity recognition

system using Wi-Fi and convolutional neural network and its integration in health information systems, Annales des Telecommunications/Annals of Telecommunications 77 (2022) 163–175.

[20] C. Liu, Y. Liu, Y. Hao, X. Zhang, LiteWiHAR: A Lightweight WiFi-based Human Activity Recognition System, 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring) (2024) 1–5.

[21] V. N. G. J. Soares, J. M. L. P. Caldeira, B. B. Zarpelão, J. Galán-Jiménez, H. Shahverdi, M. Nabati, P. F. Moshiri, R. Asvadi, S. A. Ghorashi, Enhancing CSI-Based Human Activity Recognition by Edge Detection Techniques, Information 2023, Vol. 14, Page 404 14 (2023) 404.

[22] M. K. A. Jannat, M. S. Islam, S. H. Yang, H. Liu, Efficient Wi-Fi-Based Human Activity Recognition Using Adaptive Antenna Elimination, IEEE Access 11 (2023) 105440–105454.

[23] M. S. Islam, M. K. A. Jannat, M. N. Hossain, W. S. Kim, S. W. Lee, S. H. Yang, STC-NLSTMNet: An Improved Human Activity Recognition Method Using Convolutional Neural Network with NLSTM from WiFi CSI, Sensors 23 (2022) 356.

[24] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, M. Wu, Two-Stream Convolution Augmented Transformer for Human Activity Recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2021, pp. 286–293.

[25] F. Luo, S. Khan, B. Jiang, K. Wu, Vision Transformers for Human Activity Recognition Using WiFi Channel State Information, IEEE Internet of Things Journal 11 (2024) 28111–28122.

[26] A. Ferrari, D. Micucci, M. Mobilio, P. Napoletano, On the Personalization of Classification Models for Human Activity Recognition, IEEE Access 8 (2020) 32066–32079.

[27] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, L. Xie, EfficientFi: Toward Large-Scale Lightweight WiFi Sensing via CSI Compression, IEEE Internet of Things Journal 9 (2022) 13086–13095.

[28] A. Zhu, Z. Tang, Z. Wang, Y. Zhou, S. Chen, F. Hu, Y. Li, Wi-ATCN: Attentional Temporal Convolutional Network for Human Action Prediction Using WiFi Channel State Information, IEEE Journal on Selected Topics in Signal Processing 16 (2022) 804–816.

[29] M. J. Bocus, W. Li, S. Vishwakarma, R. Kou, C. Tang, K. Woodbridge, I. Craddock, R. McConville, R. Santos-Rodriguez, K. Chetty, R. Piechocki, OPERAnet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors, Scientific Data 2022 9:1 9 (2022) 1–18.

[30] A. Elkelany, R. Ross, S. Mckeever, WiFi-Based Human Activity Recognition Using Attention-Based BiLSTM, in: L. Longo, R. O'Reilly (Eds.), Proceedings of the 30th Irish Conference on Artificial Intelligence and Cognitive Science AICS 2022., volume 1662 CCIS, Springer, Cham, 2023, pp. 121–133.

[31] W. Jiao, C. Zhang, An Efficient Human Activity Recognition System Using WiFi Channel State Information, IEEE Systems Journal 17 (2023) 6687–6690.

[32] F. Deng, E. Jovanov, H. Song, W. Shi, Y. Zhang, W. Xu, WiLDAR: WiFi Signal-Based Lightweight Deep Learning Model for Human Activity Recognition, IEEE Internet of Things Journal 11 (2024) 2899–2908.

[33] T. Nakamura, M. Bouazizi, K. Yamamoto, T. Ohtsuki, Wi-Fi-Based Fall Detection Using Spectrogram Image of Channel State Information, IEEE Internet of Things Journal 9 (2022) 17220–17234.

[34] E. Shalaby, N. ElShennawy, A. Sarhan, Utilizing deep learning models in CSI-based human activity recognition, Neural Computing and Applications 34 (2022) 5993–6010.

[35] H. Ambalkar, X. Wang, S. Mao, Adversarial Human Activity Recognition Using Wi-Fi CSI, Canadian Conference on Electrical and Computer Engineering 2021-September (2021).

[36] I. A. Showmik, T. F. Sanam, H. Imtiaz, Human Activity Recognition from Wi-Fi CSI data using Principal Component-based Wavelet CNN, Digital Signal Processing 138 (2023) 104056.

[37] S. Zhou, L. Guo, Z. Lu, X. Wen, Z. Han, Wi-Monitor: Daily Activity Monitoring Using Commodity Wi-Fi, IEEE Internet of Things Journal 10 (2023) 1588–1604.

[38] X. Chen, Y. Zou, C. Li, W. Xiao, A Deep Learning Based Lightweight Human Activity Recognition System Using Reconstructed WiFi CSI, IEEE Transactions on Human-Machine Systems 54 (2024) 68–78.

[39] C. Y. Lin, C. Y. Lin, Y. T. Liu, Y. W. Chen, T. K. Shih, WiFi-TCN: Temporal Convolution for Human Interaction Recognition based on WiFi signal, IEEE Access (2024).

[40] S. Mekruksavanich, W. Phaphan, N. Hnoohom, A. Jitpattanakul, Attention-Based Hybrid Deep

Learning Network for Human Activity Recognition Using WiFi Channel State Information, Applied Sciences 2023, Vol. 13, Page 8884 13 (2023) 8884.

[41] A. Natarajan, V. Krishnasamy, M. Singh, Design of a Low-Cost and Device-Free Human Activity Recognition Model for Smart LED Lighting Control, IEEE Internet of Things Journal 11 (2024) 5558–5567.

[42] H. Salehinejad, S. Valaee, LiteHAR: Lightweight Human Activity Recognition From Wifi Signals With Random Convolution Kernels, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2022-May (2022) 4068–4072.

[43] W. Cui, B. Li, L. Zhang, Z. Chen, Device-free single-user activity recognition using diversified deep ensemble learning, Applied Soft Computing 102 (2021) 107066.

[44] Y. Fang, F. Xiao, B. Sheng, L. Sha, L. Sun, Cross-scene passive human activity recognition using commodity WiFi, Frontiers of Computer Science 16 (2022) 1–11.

[45] S. K. Yadav, S. Sai, A. Gundewar, H. Rathore, K. Tiwari, H. M. Pandey, M. Mathur, CSITime: Privacy-preserving human activity recognition using WiFi channel state information, Neural Networks 146 (2022) 11–21.

[46] B. A. Alsaify, M. Almazari, R. Alazrai, S. Alouneh, M. I. Daoud, A CSI-Based Multi-Environment Human Activity Recognition Framework, Applied Sciences 2022, Vol. 12, Page 930 12 (2022) 930.

[47] Y. Zhang, F. He, Y. Wang, D. Wu, G. Yu, CSI-based cross-scene human activity recognition with incremental learning, Neural Computing and Applications 35 (2023) 12415–12432.

[48] B. A. Alsaify, M. M. Almazari, R. Alazrai, M. I. Daoud, A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments, Data in Brief 33 (2020) 106534.

[49] B. A. Alsaify, M. M. Almazari, R. Alazrai, M. I. Daoud, Exploiting Wi-Fi Signals for Human Activity Recognition, 2021 12th International Conference on Information and Communication Systems, ICICS 2021 (2021) 245–250.

[50] M. J. Bocus, H. S. Lau, R. Mcconville, R. J. Piechocki, R. Santos-Rodriguez, Self-Supervised WiFi-Based Activity Recognition, in: 2022 IEEE GLOBECOM Workshops, GC Wkshps 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 552–557.

[51] Z. Chen, L. Zhang, C. Jiang, Z. Cao, W. Cui, WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM, IEEE Transactions on Mobile Computing 18 (2019) 2714–2724.

[52] H. Yan, Y. Zhang, Y. Wang, K. Xu, WiAct: A Passive WiFi-Based Human Activity Recognition System, IEEE Sensors Journal 20 (2020) 296–305.

[53] R. K. Pearson, Y. Neuvo, J. Astola, M. Gabbouj, Generalized Hampel Filters, Eurasip Journal on Advances in Signal Processing 2016 (2016) 1–18.

[54] L. B. de Amorim, G. D. Cavalcanti, R. M. Cruz, The choice of scaling technique matters for classification performance, Applied Soft Computing 133 (2023) 109924.

[55] O. Banos, J. M. Galvez, M. Damas, H. Pomares, I. Rojas, Window Size Impact in Human Activity Recognition, Sensors 2014, Vol. 14, Pages 6474-6499 14 (2014) 6474–6499.

[56] A. Dehghani, O. Sarbishei, T. Glatard, E. Shihab, A Quantitative Comparison of Overlapping and Non-Overlapping Sliding Windows for Human Activity Recognition Using Inertial Sensors, Sensors 2019, Vol. 19, Page 5026 19 (2019) 5026.

[57] X. Yin, Z. Liu, D. Liu, X. Ren, A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data, Scientific Reports 2022 12:1 12 (2022) 1–11.

[58] S. K. Challa, A. Kumar, V. B. Semwal, A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data, Visual Computer 38 (2022) 4095–4109.