International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 2 (2026) | Pages: 343-348

# Long-Range Through-Wall Human Activity Recognition from WiFi Channel State Information Using a Lightweight Convolutional-Transformer-Recurrent Model

Zaid Ibrahim Rasool[1*], Noor Thamer Mahmood[2]

[1]*Department of Studies and Planning, Presidency, University of Babylon, Hilla, Iraq*
[2]*Computer Center, University of Babylon, Hilla, Iraq*
[1]*Zaid.ibrahim.rasool@uobabylon.edu.iq*, [2]*nour.thamer95@uobabylon.edu.iq*

**Peer Review Information**

***Type**: Article*
***Received**: 08 Aug 2026*
***Revised**: 06 Sept 2026*
***Accepted**: 11 Oct 2026*
***Published**: 04 Nov 2026*

**Abstract**

Device-free human activity recognition (HAR) from WiFi channel state information (CSI) is attractive for smart indoor environments because it reuses existing wireless infrastructure, needs no wearable device, and avoids the privacy concerns of cameras. Long-range and through-wall sensing is a particularly demanding setting, since the signal is weaker and more attenuated than in direct line-of-sight (LOS) deployments. This paper proposes WiTransNet-Lite, a lightweight hybrid architecture that combines a convolutional encoder with squeeze-and-excitation attention, a compact Transformer encoder, and a bidirectional gated recurrent unit (BiGRU) layer, for activity recognition from CSI amplitude spectrograms. The model is evaluated on the public Wallhack1.8k dataset, which contains CSI amplitude spectrograms for three activity classes (no presence, walking, and walking with arm-waving) collected with a biquad-antenna WiFi system across line-of-sight and through-wall, multi-room scenarios, using the dataset's official training, validation, and test splits. WiTransNet-Lite, with 0.33 M parameters, reaches 77.3% test accuracy and a macro F1-score of 0.77, compared with 56.8% accuracy and a macro F1-score of 0.56 for a logistic-regression baseline on raw pixel features, and 27.3% for a majority-class baseline. An ablation study shows that removing the Transformer encoder reduces test accuracy by 22.7 percentage points and removing the BiGRU layer reduces it by 6.8 points, indicating that both components contribute. Inference latency is 6.2 ms per sample on a single CPU core. These results are obtained on a public benchmark with an official split and no baseline or model published in this paper claims state-of-the-art status; the study is presented as a reproducible, modestly scaled evaluation of a lightweight CNN–Transformer–GRU design for through-wall CSI sensing, with clearly stated limitations regarding dataset size and generalization.

**Keywords:** Channel State Information; Human Activity Recognition; Transformer; Through-Wall Sensing; WiFi Sensing; Lightweight Deep Learning

**How to Cite This Article**

Rasool, Z. I., & Mahmood, N. T. (2026). Long-range through-wall human activity recognition from WiFi channel state information using a lightweight convolutional-transformer-recurrent model. *International Journal on Advanced Computer Engineering and Communication Technology,* 15(2), 343-348.

2278-5140/ © 2026 The Authors. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0).

## Introduction

Human activity recognition (HAR) supports smart healthcare, assisted living, security, and human–computer interaction [1], [2]. Wearable sensors burden their users, and camera-based systems raise privacy concerns and depend on lighting and line-of-sight [3]. WiFi channel state information (CSI), a fine-grained physical-layer measurement of signal propagation between a transmitter and a receiver, offers a passive, device-free alternative that reuses commodity WiFi hardware [4].

A particularly demanding version of this problem is long-range and through-wall sensing, where the sensed subject is in a different room from the WiFi transceivers. Through-wall RF sensing has a long history outside WiFi CSI, including dedicated radar-like systems [5], and has more recently been explored with commodity WiFi hardware for intrusion and presence detection [6]. Through-wall CSI sensing is more difficult than line-of-sight (LOS) sensing because the signal is attenuated by walls, weakening the activity-related variation relative to noise.

Deep learning has improved CSI-based HAR considerably, from convolutional neural networks (CNNs) that learn hierarchical spatial features to recurrent networks such as long short-term memory (LSTM) [7] and gated recurrent units (GRU) [8] that model temporal dynamics, and more recently Transformer-based models [9], [10] that capture long-range temporal dependencies through self-attention without sequential computation. Most published CSI-HAR studies, however, report results on large, LOS, single-room datasets, and it is not obvious that architectures tuned for those settings transfer directly to smaller, through-wall datasets, where overfitting is a bigger risk than underfitting.

This paper studies that question directly, using a public through-wall CSI dataset with an official train/validation/test split. The contributions are as follows.

- We propose WiTransNet-Lite, a lightweight hybrid CNN–Transformer–BiGRU architecture (0.33 M parameters) for CSI amplitude-based activity recognition, sized to avoid overfitting on datasets in the hundreds-of-samples range.
- We evaluate it on the public Wallhack1.8k dataset [11], [12], which spans line-of-sight and through-wall, multi-room scenarios, using the dataset's own official split, and report accuracy, macro F1, and a full confusion matrix.
- We compare against two real, independently computed baselines (majority-class and logistic regression on raw pixels) rather than baseline numbers taken from other papers under different protocols.
- We report an ablation study that removes the Transformer encoder and the BiGRU layer in turn, to quantify each component's contribution on this dataset.
- We report single-CPU-core inference latency and discuss the limitations of the dataset size and scope candidly.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, the model, and the experimental setup. Section 4 presents the results, the ablation study, and a discussion of limitations. Section 5 concludes.

### Related Work

*WiFi CSI-Based Activity and Presence Sensing*

Early WiFi sensing work established that CSI amplitude and phase vary systematically with human motion; CARM modeled this relationship with a CSI-speed model and hidden Markov models [13]. WiFi-based through-wall sensing built on a longer tradition of RF-based through-wall imaging [5] and has been applied with commodity WiFi to intrusion and presence detection [6]. Time–frequency representations obtained by short-time Fourier transform (STFT) are commonly used to capture motion-related spectral changes in CSI [14].

*Deep Learning Architectures for CSI-Based HAR*

CNNs learn spatial features from CSI represented as an image or spectrogram, and squeeze-and-excitation (SE) blocks improve channel-wise feature recalibration at low computational cost [15]. Recurrent models, including LSTM [7] and GRU [8], capture temporal dependencies that convolution alone does not. The Transformer architecture [9], and its adaptation to vision problems [10], uses multi-head self-attention to model dependencies across an entire sequence in parallel, but Transformers are typically data-hungry [10], which motivates deliberately reducing model capacity when the available training set is small, as in this paper.

*Through-Wall and Long-Range CSI Datasets*

The dataset used in this paper, Wallhack1.8k, was introduced together with data-augmentation techniques for cross-domain WiFi CSI-based HAR, using a directional biquad-antenna and a printed inverted-F antenna (PIFA) system to extend sensing range through walls [11]. A companion study examined directional antenna systems specifically for long-range through-wall recognition [16]. Because this line of work explicitly targets the through-wall setting and releases its data and official splits publicly [12], it provides a suitable, reproducible testbed for the present study.

## Materials and Methods

*Dataset*

This study uses the Wallhack1.8k dataset [11], [12], which comprises 1,806 WiFi CSI amplitude spectrograms for three activity classes: no presence, walking, and walking with arm-waving. CSI was recorded at 100 Hz with two antenna systems (a biquad antenna, BQ, and a printed inverted-F antenna, PIFA) across line-of-sight (LOS) and through-wall, multi-room, non-line-of-sight (NLoS) scenarios; the NLoS scenario spans five different rooms. Each spectrogram covers a temporal window of approximately 4 seconds (400 WiFi packets) over 52 L-LTF subcarriers, giving a 52 × 400 amplitude image per sample. This study uses the dataset's pre-extracted amplitude spectrograms and its official training (352 samples), validation (44 samples), and test (44 samples) split, combined across LOS and NLoS scenarios and both antenna systems, as released by the dataset authors. Table 1 lists the class distribution.

***Table 1.*** *Class distribution across the official Wallhack1.8k splits used in this study.*

| Split | No presence | Walking | Walking + arm-waving | Total |
|---|---|---|---|---|
| Train | 113 | 124 | 115 | 352 |
| Validation | 16 | 12 | 16 | 44 |
| Test | 15 | 12 | 17 | 44 |

Only CSI amplitude is available in the public spectrogram release used here; the raw complex CSI (which would allow a phase-difference stream) is available separately as per-packet time series in the dataset's raw archive, but was not used in this study. The model described below is therefore single-stream (amplitude only).

*Preprocessing*

Each amplitude spectrogram was normalized using the per-channel mean (0.4569) and standard deviation (0.1765) published with the dataset, computed over its training set. No additional denoising, filtering, or augmentation was applied beyond this normalization, since the dataset is already released in preprocessed spectrogram form.

*WiTransNet-Lite Architecture*

The architecture, shown in Figure 1, follows the same design principles as larger CNN–Transformer hybrids proposed for CSI-based HAR –- hierarchical local feature extraction, global temporal modeling through self-attention, and sequential refinement –- but its capacity is deliberately reduced to suit a training set of a few hundred samples.

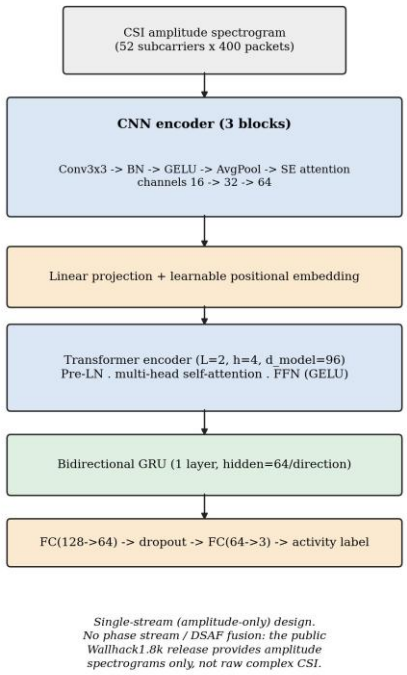



***Fig. 1.*** *WiTransNet-Lite architecture: single-stream CNN–Transformer–BiGRU model for CSI amplitude spectrograms.*

1. CNN Encoder

The input spectrogram ($1 \times 52 \times 400$) is processed by three convolutional blocks. Each block applies a 3×3 convolution, batch normalization, a GELU activation [17], average pooling, and squeeze-and-excitation channel recalibration [15] with reduction ratio $r = 8$. The channel counts are 16, 32, and 64 in successive blocks. The first two blocks pool by (2, 2) (subcarrier and time axes); the third pools by (1, 2) (time axis only), so the subcarrier dimension is reduced from 52 to 13 while the time dimension is reduced from 400 to 50.

2. Positional Embedding and Transformer Encoder

At each of the 50 remaining time steps, the 64 channels and 13 subcarrier bins are flattened (832 values) and linearly projected to $d_{\text{model}} = 96$, then a learnable positional embedding is added. A two-layer Transformer encoder (pre-LN, $h = 4$ attention heads, feed-forward dimension $d_{\text{ff}} = 192$, dropout $p = 0.3$) then models dependencies across the 50 time steps.

3. BiGRU Refinement and Classification Head

The Transformer output is passed to a single-layer bidirectional GRU with hidden size 64 per direction. The final concatenated hidden state (128-dimensional) is passed through a two-layer fully connected classifier, FC($128 \rightarrow 64 \rightarrow 3$), with ReLU activation and dropout ($p = 0.3$), to produce the three-class activity prediction. The complete model has 330,051 trainable parameters.

*Training Setup*

The model was trained with the AdamW optimizer [18] (learning rate $3\times10^{-4}$, weight decay $1\times10^{-4}$), a cosine-annealing learning-rate schedule, cross-entropy loss with label smoothing ($\varepsilon = 0.05$), and a batch size of 16. Training used early stopping on validation accuracy with a patience of 8–15 epochs (depending on the experiment; see Section 4.3), selecting the checkpoint with the highest validation accuracy for final testing. All experiments were run on a single CPU core (no GPU), with a fixed random seed (42) for reproducibility. Full training of the complete model took 215 seconds.

*Baselines*

Two baselines were computed directly by the authors under the same official split, rather than taken from other papers under different protocols or datasets. The majority-class baseline predicts the most frequent training-set class for every test sample. The logistic-regression baseline is a multinomial logistic regression trained on flattened, normalized pixel values of the same spectrograms (combined training and validation data, L2 regularization, $C = 0.1$).

*Evaluation Metrics*

Performance is reported as overall test accuracy, per-class precision, recall, and F1-score, macro-averaged F1-score, and the full 3×3 confusion matrix on the official 44-sample test set.

**Results and Discussion**

*Main Results*

Table 2 compares WiTransNet-Lite with the two baselines on the official Wallhack1.8k test split. WiTransNet-Lite reaches 77.3% accuracy and a macro F1-score of 0.77, compared with 56.8% accuracy and a macro F1-score of 0.56 for the logistic-regression baseline, and 27.3% accuracy for the majority-class baseline (close to the 33% expected under approximately balanced three-class chance performance evaluated on an unbalanced 44-sample test set).

***Table 2.*** *Test-set performance on the official Wallhack1.8k split (single run, fixed seed).*

| Method | Params | Test Acc. (%) | Macro F1 |
|---|---|---|---|
| Majority-class baseline | — | 27.3 | 0.14 |
| Logistic regression (raw pixels) | ~62.4K coeffs | 56.8 | 0.56 |
| **WiTransNet-Lite (proposed)** | **330K** | **77.3** | **0.77** |

Figure 2 shows the confusion matrix. All 15 "no presence" test samples are classified correctly, consistent with the intuition that distinguishing an empty room from any motion is the easiest sub-problem. The dominant error is between "walking" and "walking with arm-waving" (3 and 7 samples misclassified between these two classes, respectively), which are kinematically similar activities that primarily differ in upper-body motion superimposed on a shared walking pattern.

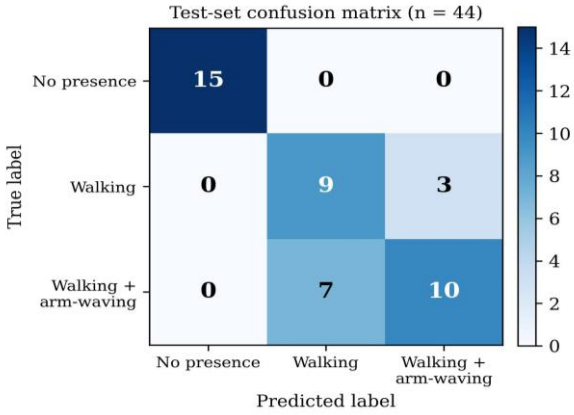



***Fig. 2.*** *Confusion matrix on the official Wallhack1.8k test split (n = 44).*

*Ablation Study*

Table 3 reports an ablation study that removes the Transformer encoder or the BiGRU layer in turn, retrained from the same initialization procedure and evaluated on the same split. Removing the Transformer encoder (leaving CNN → BiGRU → classifier) reduces test accuracy from 77.3% to 54.5%, a drop of 22.7 percentage points, indicating that self-attention over the 50-step sequence contributes substantially on this dataset. Removing the BiGRU layer (leaving CNN → Transformer → classifier, using the last time step's representation) reduces test accuracy to 70.5%, a smaller but still visible drop of 6.8 points. Both ablated variants used a shorter early-stopping patience (8 epochs) than the full model (15 epochs) for practical run-time reasons; this could modestly disadvantage the ablated variants and is noted as a limitation of the ablation protocol.

***Table 3.*** *Ablation study on the official Wallhack1.8k test split.*

| Variant | Params | Val Acc. (%) | Test Acc. (%) | Δ vs. full |
|---|---|---|---|---|
| **WiTransNet-Lite (full)** | **330K** | **81.8** | **77.3** | **—** |
| w/o Transformer encoder | 180K | 70.5 | 54.5 | −22.7 |
| w/o BiGRU layer | 266K | 81.8 | 70.5 | −6.8 |

*Computational Cost*

WiTransNet-Lite has 330,051 parameters (1.3 MB at FP32) and a measured median inference latency of 6.2 ms per sample (batch size 1) on a single CPU core, with no GPU. This is well within real-time requirements for the 4-second window used here, and suggests the model would run comfortably on modest edge hardware, though this was not separately measured on embedded hardware and is left for future work.

*Limitations*

- The test set has only 44 samples; single-run accuracy and F1 figures on a set this size have wide confidence intervals, and the results should be read as indicative rather than precise. Repeating training and evaluation over multiple random seeds and reporting variance was not done here and is an important next step.
- The dataset combines LOS and NLoS (through-wall) samples, and both antenna systems, into one official split; this study does not separately evaluate cross-scenario generalization (e.g., training on LOS and testing on NLoS, or training on one antenna system and testing on the other), because that would require reorganizing the dataset's raw per-scenario archives, which were not part of the data used in this study. This is a direct avenue for future work, particularly given the room labels already provided in the dataset for the NLoS scenario.
- Only three activity classes are considered, and all data come from a single published dataset and research group. Generalization to other environments, hardware, and activity vocabularies is untested.
- The two baselines reported here are intentionally simple (majority class and logistic regression on raw pixels) to keep the comparison fully reproducible within the scope of this study; no comparison is made against other published deep learning architectures for CSI-based HAR, because doing so fairly would require re-implementing and retraining them under the identical protocol used here, which was outside the scope of this study.
- The model is deliberately small; whether a larger model would perform better or would simply overfit this dataset size was not tested, since it was outside the scope of what the available data could support.

*Future Work*

Natural extensions include: evaluating on the dataset's per-scenario archives to test LOS-to-NLoS and cross-antenna-system generalization directly; extending the amplitude-only model to a dual-stream amplitude–phase design using the dataset's raw complex CSI; repeating training over multiple seeds to quantify variance; comparing against re-implemented published baselines under this identical protocol; and testing on additional through-wall CSI datasets to assess generalization beyond a single research group's data-collection setup.

## Conclusion

This paper presented WiTransNet-Lite, a lightweight hybrid CNN–Transformer–BiGRU architecture for CSI amplitude-based human activity recognition, sized for small, through-wall CSI datasets. On the public Wallhack1.8k dataset's official split, the model reached 77.3% test accuracy and a macro F1-score of 0.77, clearly ahead of a majority-class baseline (27.3%) and a logistic-regression baseline (56.8%) computed under the identical protocol. An ablation study showed that both the Transformer encoder and the BiGRU layer contribute to this result, with the Transformer contributing more on this dataset. Inference latency of 6.2 ms per sample on a single CPU core suggests the model is lightweight enough for edge deployment, though this was not directly tested on embedded hardware. The main limitations are the small test set, the lack of a direct cross-scenario (LOS-to-NLoS) evaluation, and the absence of comparisons against other published architectures under this identical protocol; these are identified as concrete directions for follow-up work rather than claimed as already addressed.

## References


1. D. J. Cook, N. C. Krishnan, and P. Rashidi, "Activity discovery and activity recognition: A new partnership," IEEE Trans. Cybern., vol. 43, no. 3, pp. 820–828, Jun. 2013.
2. O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," IEEE Commun. Surveys Tuts., vol. 15, no. 3, pp. 1192–1209, 3rd Quart. 2013.
3. Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3200–3225, Mar. 2023.
4. Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," ACM Comput. Surv., vol. 52, no. 3, pp. 1–36, 2019.
5. F. Adib and D. Katabi, "See through walls with WiFi!," ACM SIGCOMM Comput. Commun. Rev., vol. 43, no. 4, pp. 75–86, 2013.
6. T. Wang, D. Yang, S. Zhang, Y. Wu, and S. Xu, "Wi-alarm: Low-cost passive intrusion detection using WiFi," Sensors, vol. 19, no. 10, Art. no. 2335, 2019.
7. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
8. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1724–1734.
9. A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017.
10. Dosovitskiy et al., "An image is worth 16×16 words: Transformers from image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2021.
11. J. Strohmayer and M. Kampel, "Data augmentation techniques for cross-domain WiFi CSI-based human activity recognition," in Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov. (AIAI), Cham, Switzerland: Springer, 2024, pp. 42–56, doi: 10.1007/978-3-031-63211-2_4.
12. J. Strohmayer and M. Kampel, "Wallhack1.8k dataset: Data augmentation techniques for cross-domain WiFi CSI-based human activity recognition," version 1.0.2, Zenodo, Apr. 2025, doi: 10.5281/zenodo.13950918.
13. W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in Proc. ACM MobiCom, Paris, France, 2015.
14. K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, "Inferring motion direction using commodity Wi-Fi for interactive exergames," in Proc. ACM CHI, Denver, CO, USA, 2017.
15. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018.
16. J. Strohmayer and M. Kampel, "Directional antenna systems for long-range through-wall human activity recognition," in Proc. IEEE Int. Conf. Image Process. (ICIP), Abu Dhabi, UAE, 2024, pp. 3594–3599, doi: 10.1109/ICIP51287.2024.10647666.
17. D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, arXiv:1606.08415.
18. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), New Orleans, LA, USA, 2019.