Annals of Telecommunications
https://doi.org/10.1007/s12243-026-01207-3

A transformer-based framework for device-free human activity recognition using Wi-Fi CSI

Allan Costa Nascimento dos Santos[1,3] · Pamella Soares[2,3] · Iandra Galdino[1,3] · Taiane C. Ramos[1] · Célio Albuquerque[1] · Raphael Guerra[1] · Natalia C. Fernandes[1] · Débora C. Muchaluat-Saade[1] · Gheorghita Ghinea[3]

Received: 2 March 2026 / Accepted: 4 August 2026
© The Author(s) 2026

Abstract

Human activity recognition (HAR) plays a pivotal role in ambient assisted living, particularly for monitoring the elderly and patients with chronic conditions. However, traditional approaches relying on wearable sensors or video cameras face significant challenges regarding user compliance and privacy intrusion. To mitigate these issues, this paper proposes a device-free sensing (DFS) (https://github.com/mestrelan/MDA-CSI) framework utilizing Wi-Fi channel state information (CSI), named `DFS-CSI`. We introduce a robust Transformer-based architecture designed to capture long-range temporal dependencies in wireless signals. `DFS-CSI` was validated using a comprehensive dataset from 86 volunteers, ensuring high generalization capabilities across diverse human motion patterns.

Keywords Transformer model · Channel state information · Wi-Fi · Wireless sensing system · Healthcare · Human activity recognition

1 Introduction

Human activity recognition (HAR) has become a cornerstone of the emerging Internet of Things (IoT) ecosystem, enabling critical applications ranging from elderly fall detection and healthcare monitoring to smart home automation and intrusion detection [1]. As the global population ages, the demand for automated, non-intrusive monitoring systems that can ensure safety without compromising independence is growing rapidly. HAR has become essential for monitoring older adults, patients with chronic diseases, and individuals with disabilities [2]. Its impact on healthcare and well-being is profound, as monitoring physical patterns helps identify sedentary behaviors and promotes active habits [3–5]. This mechanism is crucial for preventing chronic conditions such as type-2 diabetes, heart disease, and obesity, in addition to improving mental health [6, 7]. For patients recovering from injuries or surgeries, continuous monitoring provides valuable insights into rehabilitation, enabling healthcare professionals to tailor interventions based on accurate mobility data [8].

✉ Allan Costa Nascimento dos Santos
allans@midiacom.uff.br; Allan.Santos@brunel.ac.uk

Pamella Soares
pamella.soares@aluno.uece.br;
Pamella.SoaresDeSousa@brunel.ac.uk

Iandra Galdino
igar@midiacom.uff.br

Taiane C. Ramos
taiane@midiacom.uff.br

Célio Albuquerque
celio@midiacom.uff.br

Raphael Guerra
rguerra@midiacom.uff.br

Natalia C. Fernandes
natalia@midiacom.uff.br

Débora C. Muchaluat-Saade
debora@midiacom.uff.br

Gheorghita Ghinea
george.ghinea@brunel.ac.uk

[1] MidiaCom Lab, Universidade Federal Fluminense, Niterói, RJ, Brazil

[2] Graduate Program in Computer Science, State University of Ceara, Ceara, Brazil

[3] Department of Computer Science, Brunel University London, London, UK

Published online: 15 August 2026

Safety is another major concern, particularly regarding falls and medical emergencies [3, 9]. Automated systems can detect extended periods of inactivity or sudden falls, triggering alerts to caregivers [3]. Traditionally, HAR systems have relied on two main approaches: vision-based methods (cameras) and sensor-based methods (wearables) [10]. However, both face significant adoption barriers. Cameras raise severe privacy concerns and suffer from performance degradation in low-light conditions [1, 5], limiting their acceptance in sensitive areas like bedrooms or bathrooms [11, 12]. Wearable devices (smartwatches or accelerometers) are often considered intrusive, require regular charging, and are prone to user non-compliance, especially among elderly patients with cognitive impairments who may forget to wear them [5, 8, 13]. Additionally, the cost of high-quality wearable devices and monitoring infrastructure can be a barrier to widespread implementation, particularly among lower-income populations [14].

To overcome these limitations, this paper proposes `DFS-CSI`, a device-free sensing (DFS) framework utilizing Wi-Fi channel state information (CSI). By analyzing how human movement modulates wireless signal propagation (via reflection, scattering, and diffraction) [15], `DFS-CSI` enables activity detection using standard Wi-Fi infrastructure without requiring users to carry devices or stand before cameras [7, 16]. As individuals move, they induce unique distortions in the wireless signal. By capturing and interpreting these variations in CSI data, our `DFS-CSI` can detect physical activities and daily movements, allowing caregivers and healthcare professionals to monitor patients' activities non-invasively [2, 7].

Despite its potential, CSI-based recognition presents significant challenges. Wireless signals are inherently noisy and highly sensitive to environmental changes [1]. Furthermore, distinguishing between complex dynamic activities (e.g., running vs. sweeping) or subtle static activities (e.g., sitting vs. lying) requires a model capable of capturing intricate temporal dependencies [11]. While convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely explored, they often struggle with long-range sequence modeling and parallelization [11, 13].

In this context, this paper proposes a robust Transformer-based architecture for HAR using CSI data. Adopting the self-attention mechanism, the proposed model effectively learns global temporal features from CSI streams. Unlike previous studies that are often limited to small datasets or theoretical accuracy, this work focuses on real-world applicability and efficiency.

`DFS-CSI` uses a Transformer model [17, 18] designed to process time-series data. It incorporates positional encoding, multi-head attention, and a fully connected layer, enabling sophisticated learning of temporal relationships and generating accurate predictions based on patterns identified in the input sequence [2, 8]. We introduced a temporal windowing strategy to reduce response time and enable real-time detection.

Unlike other machine learning models in the literature, `DFS-CSI` features a generalized approach capable of identifying the activities of individuals who did not participate in the model training phase [13, 19]. To ensure high reliability, we developed a robust set of specialist binary models, each optimized to detect a specific activity with maximum precision. The `DFS-CSI` detects the presence of a person in a room and determines whether the person is moving. If the person is moving, the `DFS-CSI` further classifies the activity as walking or running. Otherwise, the `DFS-CSI` classifies the activity as lying down or sitting. To evaluate the performance of `DFS-CSI`, we constructed a comprehensive dataset based on CSI collected in a controlled indoor environment. The experimental protocol was designed to capture the variations in Wi-Fi signals caused by human interference, utilizing a standard 5 GHz network infrastructure. `DFS-CSI` was trained, validated, and tested using CSI data from 86 volunteers of various genders, ages, and physical characteristics, collected in an indoor controlled environment [20]. The collection protocol encompassed a diverse set of scenarios to ensure the model's robustness. Volunteers performed 17 distinct activities, ranging from static activities (such as sitting and lying down with different orientations) to dynamic movements (including walking, running, and sweeping). Additionally, data from an empty room was collected to establish a baseline for the absence of human presence. This data is acquired in real-time, allowing the `DFS-CSI` to monitor the subject's movements as they occur within the environment.

This article represents a significant extension of our preliminary research presented in [19]. This work presents a comprehensive evolution of the methodology, addressing scalability and operational efficiency. While our previous study established the feasibility of the architecture with a limited cohort, this work expands the scope and depth of the analysis in critical dimensions:

- **Dataset expansion and generalization:** We increased the participant pool from 59 to 86 volunteers (a 45.8% increase), ensuring that the model generalizes effectively across diverse body types and physical traits.
- **Feasibility analysis:** We shifted the evaluation focus from pure classification accuracy to operational viability. This study includes a comprehensive assessment of computational resources, covering inference latency, power consumption, and thermal stability, which were not addressed in the original publication.
- **Granular activity analysis:** Moving beyond broad activity groups (e.g., general static activities), we implemented

specific binary classifiers to distinguish fine-grained behaviors, such as differentiating "sweeping" from "running" with high precision.
- **Expansion of detection capabilities:** Unlike our initial research [19], which focused on fewer categories, this study developed and validated a wider range of models to recognize specific activities that were not previously addressed. The `DFS-CSI` can now accurately detect distinct actions—such as specific types of sweeping, running, and other complex movements—making the overall solution much more complete and versatile than previous approaches.
- **Temporal window optimization:** We conducted a rigorous analysis of input window sizes, identifying that a reduced 0.5-s window ($W = 16$ packets) maximizes real-time responsiveness without compromising accuracy.
- **Faster movement detection:** Unlike traditional approaches that rely on rigid temporal resolutions, this study introduces a comprehensive multi-scale analysis using window sizes ranging from 0.5 to 30 s, significantly enhancing the model's ability to capture complex CSI patterns.
- **Green AI and efficiency metrics:** Unlike previous basic resource logging, we provide a holistic "Green AI" evaluation, mapping the trade-offs between inference latency, power consumption, and thermal stability to validate the `DFS-CSI`'s suitability for edge deployment.

Our previous work [19] focused on broad categories, grouping activities into general classes such as "walking/running" (80% accuracy) or "sitting/lying" (75% accuracy).

The current manuscript improves this by decomposing these broad groups into specific binary problems, yielding higher precision.

- **Specific distinctions:** The new model achieves high accuracy in distinguishing new activities.
- **Improved motion analysis:** While the initial model grouped walking and running, the new study successfully separates "walking vs. running" and identifies "sweeping detection."
- **Transitional movements:** The new work successfully analyzes transitional activities (e.g., "sit-to-stand detection" for sitting/standing transitions), a level of detail not present in the initial study.

A major limitation of the previous work was the lack of temporal optimization. The input size was fixed at 1 min of collected data, and its impact was not analyzed.

The current manuscript introduces a detailed study on window sizes. By testing windows ranging from 0.5 to 30 s, we observed that a short window of 0.5 s ($W = 16$ packets) is optimal. This configuration achieved:

- **High accuracy:** 92.32% for movement detection.
- **Real-time response:** The inference time dropped significantly, allowing the `DFS-CSI` to detect movements almost instantly.

This finding proves that the current proposal is not only accurate but also faster and more responsive than the version presented in [19].

The remainder of this manuscript is organized as follows. Section 2 describes related work, while Section 3 presents the proposed methodology. Section 4 discusses the results obtained. Finally, conclusions and suggestions for future work are presented in Section 5.

2 Related work

Wi-Fi-based sensing has gained increasing attention in recent years as a promising approach for device-free monitoring, leveraging the widespread deployment of wireless infrastructure to perceive environmental changes without requiring users to carry dedicated devices [20, 21]. The radio signals exchanged between Wi-Fi transceivers are inherently affected by surrounding objects, as well as by human presence, activity, and movement, leading to measurable variations in signal propagation characteristics [20, 21]. While early studies primarily relied on coarse-grained metrics such as received signal strength (RSS), more recent work has demonstrated that channel state information (CSI) offers a much richer representation of the wireless channel by exposing fine-grained physical-layer (PHY) details, including amplitude and phase information across multiple subcarriers [3, 22, 23]. This increased granularity enables CSI-based systems to better capture subtle human-induced disturbances and multipath effects, resulting in improved robustness and sensing accuracy compared to RSS-based approaches. Owing to these advantages, CSI has been widely explored for various applications, ranging from indoor localization and activity recognition to health-related monitoring. In particular, CSI-based sensing shows strong potential for elderly care scenarios [6, 21], as it enables unobtrusive and continuous monitoring in home or hospital environments while avoiding privacy-invasive sensors and minimizing the need for additional sensing hardware or costly infrastructure upgrades [20].

Sai et al. [24] provide a comprehensive and systematic review of Machine Learning techniques applied to Wi-Fi channel state information (CSI)-based sensing and recognition, analyzing over 65 representative models for applications such as HAR, gesture detection, and indoor localization.

The work establishes a hierarchical taxonomy categorizing approaches into classical models, deep learning, and hybrid strategies while detailing the complete implementation pipeline, from hardware and software tool selection to crucial signal processing steps such as noise reduction and feature engineering. In addition to comparing the performance of architectures in healthcare, smart home, and security scenarios, the authors identify persistent challenges related to generalization in dynamic environments and multi-user interference, proposing future directions involving the use of fundamental wireless models, federated learning for privacy preservation, and efficient architectures such as state space models to enable robust and scalable deployment on edge devices.

Xiao et al. [25] proposed a semi-supervised generative adversarial network (GAN) for CSI-based activity recognition, called CsiGAN. The proposed method uses a complement generator, which can leverage limited unlabeled data to generate diverse fake samples for training a robust discriminator. For the discriminator, they modify the number of probability outputs, which helps establish a more accurate decision boundary. Additionally, they introduced manifold regularization to stabilize the learning process. The model was designed to address performance degradation in leave-one-subject-out validation for CSI-based activity recognition. However, the data used to train the model contained only three candidates (3), with limited activities, resulting in low accuracy.

Xiao et al. [26] present Pilot, a passive, device-free indoor localization system based on channel state information (CSI) in WLAN networks, which detects and locates human entities without them carrying equipment. Initially, the system builds a normal profile from the temporal correlation of CSI in a static environment, showing that this data does not follow classical distributions such as the Gaussian distribution, which motivates the use of non-parametric methods. Next, Pilot defines the concepts of sensing zone and dead spot to characterize where an AP-DP link can or cannot detect the presence of an entity, reducing the search space and improving location accuracy. Anomaly detection is performed by exploiting the temporal stability of CSI in normal states and using kernel density estimation to identify deviations caused by movement, while location is determined by a MAP algorithm that compares anomalous measurements with a passive fingerprint map built offline. The system also integrates data fusion from multiple links to increase accuracy. Experimental evaluations in a laboratory and in a lobby show that Pilot outperforms traditional RSS-based approaches in both anomaly detection and location accuracy, highlighting the potential of using the frequency diversity and temporal stability of CSI for passive indoor localization applications. Although the original article reports good numbers in a controlled environment, comparative studies (e.g., in the WiDetect article table) show that PILOT suffers from approximately 10% false negatives and approximately 10% false positives. This places the actual accuracy at around 80 to 85% in common home environments, as it very easily confuses slow environmental changes with movement.

To detect slow human movement, the sensitivity decreases. The article reports a "true positive rate" that can fluctuate between 85 and 90% depending on the target's speed. If the person moves very slowly, the variance of the eigenvalues is not sufficient to trigger the alarm without generating many false alarms.

Caballero et al. [4] proposed an approach for HAR using commercial Wi-Fi devices. With their method, it is possible to infer the activity of a monitored person in an indoor environment. To achieve this, they clean and process the amplitude of the CSI data collected from the Wi-Fi channel. For the scenario and dataset considered in this study, the results showed that the random forest (RF) algorithm performed best in all tests, achieving an average accuracy of 93.03%. However, the RF model must be trained individually for each person to detect their activity. Hence, the model must be trained on a unique individual's CSI data to function properly.

Wang et al. [27] proposed a multimodal channel state information-based activity recognition (MCBAR) system that leverages existing Wi-Fi infrastructure to monitor human activities using CSI measurements. MCBAR employs a multimodal generator to approximate the CSI data distribution across different environmental settings with limited measured CSI data. The generated CSI data from the multimodal generator provides diversity for knowledge transfer. A 2080ti GPU was used in their experiments. However, the model's training data were limited in both participant diversity and activity variety, restricting its applicability to new users.

Li et al. [28] proposed a HAR scheme based on the collaboration between vision and Wi-Fi. They collected CSI data from Wi-Fi signals and human skeletal points from video. They developed a long-short-term Transformer network to combine the CSI data and skeleton points. The proposed method uses the Transformer neural network to form the long-short-term Transformer (LSTT) network, enabling it to derive the human skeleton points from the CSI data. Their method achieves 96% accuracy; however, it was tested using CSI data from only one person, limiting its ability to generalize to a broader population. Additionally, their method requires video data, which introduces significant limitations, including cost, data security concerns, and privacy issues.

Recent studies have also explored the use of commercial Wi-Fi infrastructure combined with traditional signal processing for human activity monitoring. For instance, Boudlal et al. [29] proposed a low-cost CSI-based system that uses filtering techniques and principal component analysis (PCA) to detect basic human movements. Although their work successfully demonstrates the feasibility of distinguishing

between sitting and standing states in both line-of-sight and non-line-of-sight scenarios, their experimental validation was limited to a single volunteer and only two basic activities. Furthermore, their approach relies heavily on manual feature extraction and qualitative validation. In contrast, our proposed DFS-CSI provides a significant advancement by leveraging a Transformer-based deep learning architecture. Instead of being limited to two states, our model successfully classifies 17 distinct, fine-grained activities. Moreover, while Boudlal et al. evaluated their system on one individual, our framework was validated on a highly diverse dataset, proving its person-independent generalization. Finally, our automated approach achieves over 92% accuracy with near real-time processing, demonstrating the superiority of attention mechanisms in handling the complex variability of CSI data in practical environments.

Similarly, Boudlal et al. [30] explored a CSI-based monitoring system specifically designed for elderly care, utilizing traditional signal processing techniques such as the discrete wavelet transform (DWT) and PCA to reduce noise and extract features. While their study successfully demonstrates the visual differences in CSI amplitude when a person transitions between sitting and standing, their experimental setup was highly restricted. They evaluated the system using data from only a single volunteer performing just two basic activities, and their validation relied primarily on qualitative visual analysis using a camera. In contrast, our proposed DFS-CSI leverages a robust Transformer-based deep learning architecture to automatically learn complex temporal patterns. Instead of being limited to two activities, our model successfully classifies 17 fine-grained activities. Furthermore, our approach moves beyond qualitative observations by providing a rigorous quantitative evaluation, achieving accuracy on a highly diverse dataset of 86 volunteers. This demonstrates the superior scalability and real-world applicability of our person-independent model compared to traditional filtering methods.

In a related study, Boudlal et al. [31] proposed a device-free activity recognition system operating specifically in the 5 GHz WiFi band. They employed traditional signal processing techniques, including the Hampel filter, DWT, and PCA, to extract features from four basic movements (sitting, standing, raising a hand, and raising a leg). While their work successfully demonstrates the sensitivity of the 5 GHz frequency to human motion, it suffers from critical scalability limitations. The experimental validation was restricted to a single volunteer, and the evaluation was primarily qualitative, relying on visual interpretations of PCA heatmaps rather than rigorous classification metrics. In contrast, our proposed DFS-CSI also effectively utilizes the 5 GHz infrastructure but completely automates the recognition process using an advanced Transformer architecture.

In another preliminary study, Boudlal et al. [32] explored a multi-environment DF-HAR system using WiFi CSI, applying techniques such as the short-time Fourier transform (STFT) and PCA for feature extraction. However, similar to their other early works, this study was highly constrained in its scope. The experiments involved only a single volunteer performing three basic activities (sitting, standing, and raising a hand). More importantly, the system's evaluation was predominantly qualitative, relying on the visual interpretation of PCA heatmaps rather than employing a robust machine learning classifier to provide concrete accuracy metrics.

In another recent conference proceeding, Boudlal and Serrhini introduced CSHA-CSI [33], a contactless sensing architecture designed for smart home and healthcare applications. While this framework further supports the feasibility of leveraging WiFi CSI for tracking daily activities, it is primarily presented as an early-stage proof-of-concept. The evaluation is confined to small-scale data and basic human motions, lacking the comprehensive testing required to prove broad generalization.

In another recent investigation, Boudlal et al. [34] evaluated the impact of different spatial configurations, with and without line of sight, on activity detection using CSI. Although the study documents in detail how the environment affects signal distribution during actions such as sitting, standing, and walking, the approach has significant methodological limitations. The system relies on traditional techniques, such as PCA, for manual feature extraction. Consequently, its validation is purely qualitative, depending on the visual separation of data into graphs, without providing exact classification accuracy. Additionally, the experiments involved only one volunteer, making it impossible to verify the generalization of the model to new users. To overcome these shortcomings, our DFS-CSI employs an advanced Transformers architecture that automatically extracts temporal patterns without relying on manual observation. Unlike the study focused on only three activities and one individual, our research expands the scope to 17 detailed activities.

Recent advancements have also attempted to merge Wi-Fi CSI with computer vision techniques. For example, Boudlal et al. [35] proposed a framework that integrates CSI with YOLOv8 and MediaPipe to perform skeleton-based pose estimation. While this method provides detailed posture analysis, it relies on camera-based technologies. This reintroduces severe privacy concerns, which defeat the primary advantage of using Wi-Fi sensing in sensitive environments like bedrooms or bathrooms. Furthermore, running heavy visual algorithms like YOLOv8 requires substantial computational resources. In contrast, our proposed DFS-CSI is strictly vision-free and relies exclusively on 1D Wi-Fi CSI signals. By using a highly optimized Transformer model,

we achieve high accuracy without ever capturing a single image; thus, we guaranty complete user privacy and low computational overhead. It proves that complex computer vision frameworks are not required to achieve robust, person-independent human activity recognition.

Most recently, Boudlal et al. [36] advanced their CSI-based framework by incorporating deep learning architectures, such as BLSTM and CNNs, to classify indoor activities. While their system achieved high accuracy and reinforced the privacy-preserving nature of WiFi sensing, the experimental scope remains fundamentally limited. Their deep learning models were trained and evaluated on data collected from a single volunteer performing only three basic activities (sitting, standing, and raising a hand). This lack of participant diversity severely restricts the model's ability to achieve true person-independent generalization in real-world scenarios.

Table 1 compares our proposal with other studies that use machine learning methods to monitor human activities through CSI. `DFS-CSI` stands out with 86 volunteers, a significantly higher number than that of most compared studies. This is important because a larger number of volunteers can increase data variability, leading to a more robust and generalizable model. The proposal is also notable for the quantity and variety of monitored activities. While the compared studies include between 5 and 6 activities, `DFS-CSI` covers 17 different activities and includes data collected in an empty room. This diversity enables the model to distinguish a broader range of behaviors and situations. The proposed `DFS-CSI` uses Transformer-based models, which are particularly effective in learning complex temporal patterns and may be better suited to capture the nuances of postural activities [37]. This gives it an advantage over traditional methods, such as GAN and random forest, used in related work. Unlike traditional multiclass frameworks presented in the literature, we propose a modular system of binary experts, allowing each model to focus exclusively on distinguishing specific activity patterns. Additionally, the proposed `DFS-CSI` is the only one that uses only real CSI data, independent of the person. This approach is advantageous for practical applications,

Table 1 Related work comparison

Ref.	Volunteers	Activities	ML method	Uses only CSI real data independent of the person	Accuracy
[4]	125	Sitting, standing, lying, walking, running, sweeping	Random forest	No	93.03%
[25]	3	Fall, walk, jump, pickup, sit down, stand up	GAN	No	86.27%
[27]	10	Running, walking, falling down, boxing, circling arms, cleaning floor	GAN	No	92.90%
[28]	3	Walking, waving hands, picking up, jumping, raising hands, squatting	Long-short-term Transformer	No	96%
[29]	1	Sitting, standing	Signal processing (PCA)	No	Qualitative visual validation
[30]	1	Sitting, standing	Signal processing (DWT + PCA)	No	Qualitative visual validation
[31]	1	Sitting, standing, hand up, leg up	Signal processing (PCA)	No	Qualitative visual validation
[32]	1	Sitting, standing, raising hand	Signal processing (STFT + PCA)	No	Qualitative visual validation
[33]	1	Basic motions	Signal processing	No	Preliminary validation
[34]	1	Sitting, standing, walking	Signal processing (PCA)	No	Qualitative visual validation
[35]	–	Pose estimation (activity recognition)	YOLOv8 + MediaPipe (computer vision)	No	–
[36]	1	Sitting, standing, raising hand	Deep learning (BLSTM/CNN)	No	Up to 100%
`DFS-CSI`	**86**	**17 distinct activities + empty room**	**Transformer**	**Yes**	**92.32**% movement detection in around 0.5s

Table 2 Comparison between the previous work [19] and the proposed approach

Metric/feature	Previous work [19]	Proposed approach
Dataset size	59 volunteers	86 volunteers
Activity granularity	Broad groups (e.g., Movement)	Fine-grained (e.g., sweeping)
Model architecture	Baseline classification	Transformer + windowing
Resource analysis	Basic (power/time)	**Advanced (thermal/latency)**
Real-time feasibility	Not evaluated	Validated (< 0.5s latency)

as it eliminates the need for system calibration for specific individuals, thereby increasing its applicability in real-world scenarios. With an accuracy of 92.32%, the DFS-CSI is competitive with other methods. This combination of high accuracy and broad applicability demonstrates the efficiency and adaptability of the proposed DFS-CSI.

Table 2 presents a comparison between the results obtained in our previous study [19] and the advancements achieved in this work. While the previous research established a solid baseline using 59 volunteers, the current study significantly expands the dataset to 86 volunteers, ensuring greater robustness against data variability.

A major evolution lies in the activity granularity. The former study focused on broad groups, essentially classifying generic states such as "movement detection" without identifying the specific type of action performed. In contrast, this work addresses fine-grained activities, requiring the model to distinguish complex, specific motion patterns like "sweeping" or "running." Despite this increased challenge, the proposed approach achieved high accuracy in tasks like *lying vs. sweeping*, confirming its precision. Finally, a comprehensive resource analysis validates the DFS-CSI's thermal stability for real-time Edge AI deployment.

3 Methodology

This section presents the proposed methodology for HAR. DFS-CSI implements a training and evaluation flow of interactions to optimize accuracy, adjusting hyperparameters based on losses and performance indicators such as accuracy, precision, recall, F1-score, as well as computational efficiency metrics (execution time and power consumption). The approach involves the development of binary classification models, each addressing a specific aspect of human activity: presence, movement, activity, and locomotion. Figure 1 shows the overall processing flow of the DFS-CSI structure.

3.1 CSI data collection

Data were collected in a dedicated room at Fluminense Federal University's Computer Science Institute using a Raspberry Pi 4B equipped with a bcm43455c0 chipset. The Raspberry Pi 4B has only one standard internal antenna. The experiments were conducted with one router, and the Raspberry Pi was positioned 1 m from the volunteers, as illustrated in Fig. 2.

The room layout, depicted in Fig. 2, shows the experimental setup enabling the capture of Wi-Fi CSI. The experiments were conducted in a 3m $\times$ 4m room containing three tables and a bed. The network consisted of a Wi-Fi router (transmitter), a laptop, and a Raspberry Pi 4B running NEXMON firmware to collect the CSI, which is sensitive to human activities within the monitored space.

When a person is positioned between the transmitter and receiver, they act as an obstacle, affecting electromagnetic signal propagation. By analyzing these variations over time, human presence and body movements can be detected. The participant sat at the center of the room 1 m away from each device. The router and laptop were positioned on opposite sides, and the Raspberry Pi was 1 m away from the participant and in front of him. When in use, the bed was positioned 1 m from the Raspberry Pi and aligned with the tables.

During the collecting sessions, volunteers were allowed to wear their usual clothing. A total of 17 distinct activities were performed, as depicted in Table 3. These activities include stationary activities (e.g., standing, sitting, lying), dynamic movements (e.g., walking, running), and other specific actions. During the collection, the CSI stream was captured and recorded in real-time.

Table 3 details the experimental protocol, which encompasses 17 distinct activities designed to evaluate the DFS-CSI's sensitivity to varying degrees of motion and activity. The activities are organized by ID and grouped into five logical categories: sitting, standing, lying, transitional, and dynamic. The protocol includes specific variations in breathing patterns (IDs 2, 5, 7, 9, 11, 13). In these scenarios, volunteers were instructed to alternate their breathing intensity. Furthermore, the inclusion of transitional activities (IDs 3 and 14) and dynamic actions such as walking, running, and sweeping (IDs 15–17) allows for a comprehensive assessment of the classifier's performance across a wide spectrum of physical intensities.

CSI data used in this research [20] were collected from a Wi-Fi network operating at 5 GHz with an 80 MHz bandwidth. To minimize external interference and ensure data reliability, we performed a spectrum analysis to identify and

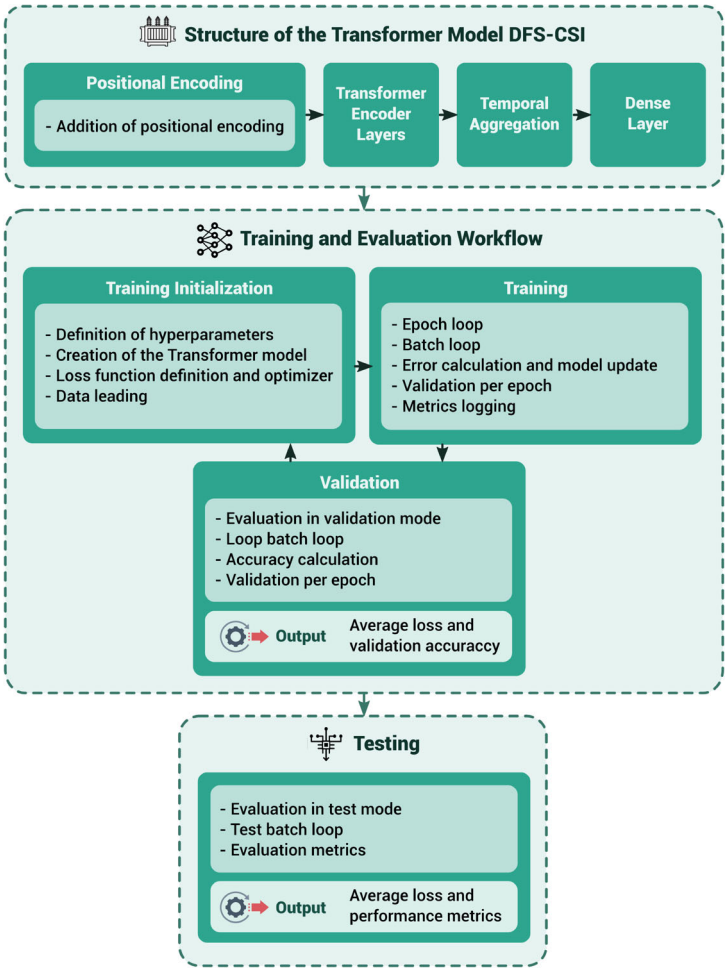


Fig. 1 `DFS-CSI` block diagram

select an unoccupied channel within the ISM band. Additionally, an 80 MHz bandwidth was selected because it provides greater CSI granularity and a larger volume of data compared to a standard 40 MHz bandwidth. To generate the necessary traffic for CSI extraction, the laptop continuously sent ping requests to the router with a transmission interval of 136 ms. This specific interval is crucial because it generates an adequate sampling frequency to capture micro-movements and reconstruct vital signs, such as breathing rates (which typically fall between 0.2 Hz and 3.5 Hz), effectively preventing signal aliasing.

While the volunteer performed the activities, the Raspberry Pi passively monitored the network traffic by filtering the MAC addresses of both the laptop and the router. The CSI stream was captured in real time and saved directly into `.pcap` files. The laptop continuously sent pings to generate traffic, while a Raspberry Pi 4B with a single antenna and NEXMON firmware passively captured CSI at 33–34 samples per second over 60 s, yielding approximately 2000 samples per minute across 256 subcarriers. The dataset comprises 17 distinct activities, plus an empty-room scenario, with 60 s of data collection for each participant. This study was approved by the Ethics Committee of Federal Fluminense University (CAAE 54359221.4.0000.5243). The demographic profile of this cohort is presented in Fig. 3. It exhibited a mean age of 26.8 years, a mean height of 171.9 cm, and a mean weight of 73.1 kg.

The network's transmitter was a TP-Link Archer C7 (AC1750) Wi-Fi router.[1] This specific model was selected because it is an affordable, widely available commercial device. By utilizing standard, low-cost equipment, we highlight a major advantage of our `DFS-CSI`: it is capable of highly accurate human activity recognition without the need for expensive, dedicated, or specialized monitoring hardware.

Regarding the physical transmission setup, the Archer C7 is equipped with three external antennas and supports MU-MIMO technology. However, because the Raspberry Pi 4B relies on a single internal antenna, the CSI data was effec-

[1] https://www.tp-link.com/br/home-networking/wifi-router/archer-c7/

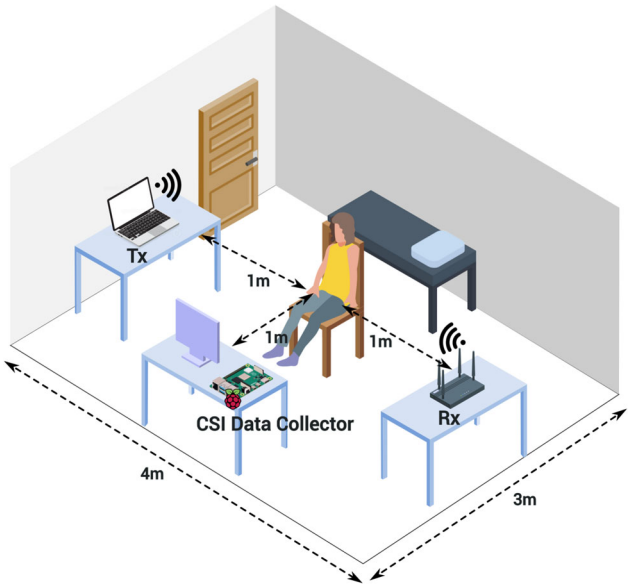


Fig. 2 Layout of the experiment room

tively captured in a 1 × 1 Single-Input Single-Output (SISO) configuration. This setup generated a CSI matrix with a single spatial stream containing 256 subcarriers per packet. Despite being a SISO link, this configuration proved to be highly efficient and provided more than enough data granularity for our classification model.

To maximize the signal's interaction with the volunteers, all devices (router, laptop, and Raspberry Pi) were positioned at a fixed height of 0.75 m from the floor. This specific height ensures a clear line of sight (LoS) across the human body, which is the optimal region for capturing breathing patterns and dynamic body movements. Furthermore, the devices were kept strictly stationary throughout the experiment, ensuring that all captured signal variations were caused exclusively by the users' physical activities and not by hardware instability.

3.2 CSI data preprocessing

Prior to processing the collected data, we performed an analysis over all 256 subcarriers to identify the most informative ones. Figure 4 illustrates the variability of collected data across an 80 MHz channel centered at 5 GHz. The results indicate a significant increase in the variance of the signal within the first 60 subcarriers, suggesting they are more sensitive to human presence, while the remaining frequencies exhibit relatively stable characteristics. Consequently,

Table 3 Detailed description of the 17 activities performed by volunteers, categorized by activity and movement dynamics

ID	Category	Description of activity
Sitting activities		
1	Static	Sitting facing the collector
2	Breathing	Sitting facing the device, alternating breathing
4	Static	Sitting facing away from the collector
5	Breathing	Sitting facing away from the device, alternating breathing
Standing activities		
6	Static	Standing facing the collector
7	Breathing	Standing facing the device, alternating breathing
8	Static	Standing facing away from the collector
9	Breathing	Standing facing away from the device, alternating breathing
Lying activities		
10	Static	Lying on bed (stomach up), sideways to collector
11	Breathing	Lying on bed (stomach up), alternating breathing
12	Static	Lying on bed (stomach down), sideways to collector
13	Breathing	Lying on bed (stomach down), alternating breathing
Transitional activities		
3	Transition	Alternating between sitting (activity 1) and standing (activity 6)
14	Transition	Alternating between standing (activity 6) and lying (activity 10)
Dynamic activities		
15	Motion	Walking activity (walking in place) facing the collector
16	Motion	Running activity (running in place) facing the collector
17	Motion	Sweeping activity in the indicated area

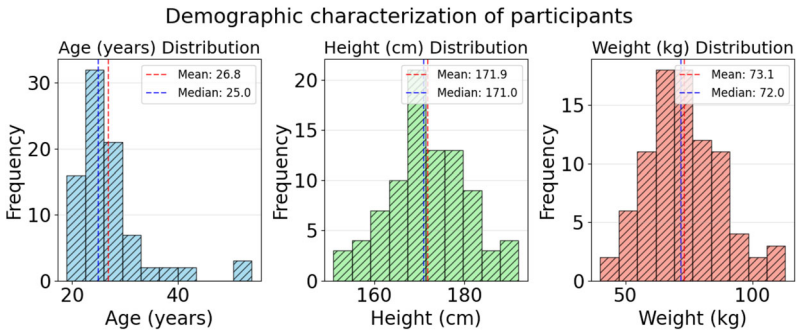


Fig. 3 Demographic characterization of the volunteers

we restricted our analysis to this high-variance region. From these initial 60 subcarriers, we removed 12 null and pilot tones that carry no meaningful data, resulting in a final feature matrix of 2000 samples × 48 active subcarriers for each collected data related to each activity.

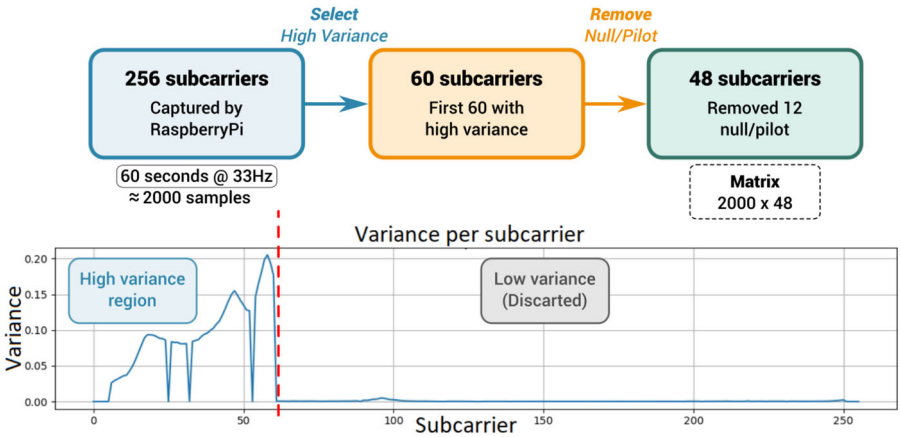
We calculated the magnitude of each sample of the collected signal within each of the 48 selected subcarriers. This calculation was performed using the components of the complex numbers obtained from the collected signals. Since a complex number is defined by its real and imaginary parts, we determined the magnitude based on these components. Thus, we obtained the magnitude for all 2000 samples across the 48 for each data collection. The resulting time series was used as input to the Transformer model [18]. After extracting the magnitude, we applied a Z-score normalization step (standardization) to the data, ensuring a mean of zero and a standard deviation of one. This normalization is a crucial step to stabilize the training process and improve the convergence rate of the Transformer model [18].

Finally, to mitigate bias and ensure dataset variability regarding the volunteer, the `DFS-CSI` applies a randomization step that shuffles the processed sample lists. Subsequently, a balancing mechanism adjusts the dataset size by truncating the larger class to match the number of samples in the smaller class, ensuring an equal distribution of data. Then, the data is labeled to identify each class for supervised learning.

The data structure at each stage of processing is as follows. The input has the shape batch size, sequence length, and number of channels, representing time-series data with multiple samples, a temporal sequence, and channels (subcarrier). During positional encoding, the shape remains batch size, sequence length, and number of channels, but position information is added to the data. The Transformer encoder layer keeps the same shape: batch size, sequence length, and number of channels, but the data undergoes self-attention and feed-forward processing to capture patterns in the sequence. After temporal aggregation, the data is reduced over time, resulting in the shape batch size and number of channels, which combines the information from the entire sequence. Finally, in the fully connected layer, the data is processed to produce the output with the shape batch size and output size, providing the final prediction for each sample in the batch.

To guarantee that the model generalizes well to new, unseen users (a person-independent approach), the dataset was split based on the volunteers rather than shuffling the entire data pool. We divided the participants into three distinct sets: 70% for training, 20% for validation, and 10% for testing. This strict separation ensures that the data of a volunteer present in the test set is never seen by the model during the training phase. Finally, the processed time series were converted into standard PyTorch tensors to serve as input for the neural network. The processed PyTorch tensors are

Fig. 4 Variance per subcarrier



publicly available on GitHub (https://github.com/mestrelan/MDA-CSI) to support the reproducibility of this research.

3.3 Transformer model

This section presents the architecture of the Transformer model used for HAR. Unlike traditional recurrent networks, the Transformer uses an attention mechanism to efficiently capture long-range dependencies within the CSI time-series data. In the following, we detail the specific internal components of the encoder layers, as well as the training strategies designed to maximize classification performance across different activity scenarios.

3.3.1 Design and model strategies

To comprehensively evaluate the capabilities of the Transformer architecture in recognizing human activities via CSI, we have used a set of 17 different models. These models cover a wide range of scenarios, including static activities (sitting, standing, lying), dynamic activities (walking, running, sweeping), and transitional activities (sit-to-stand, lying-to-standing).

Based on this dataset, the training strategy was divided into two strategies: one-vs-all detection models and specific binary comparison models, as Fig. 5 shows.

Strategy A—One-vs-all detection models The objective here is to isolate a specific activity from the entire remaining dataset. The target activity is labeled as `Class 1`, while all other 16 activities are aggregated as `Class 0`. This strategy tests the model's ability to extract unique features of a specific action amidst a diverse background of other movements. The models included in this category are:

- **Sweep detection**: Targets activity 17 versus all others.
- **Presence detection**: Targets the general presence of a human (all activities) versus an empty room.
- **Standing-to-lying detection**: Activity 14 versus all other activities.
- **Sit-to-stand detection**: Activity 3 versus all other activities.
- **Run detection**: Activity 16 versus all other activities.
- **Walk detection**: Targets activity 15 versus all other activities.
- **Sit detection**: All sitting activities 1, 2, 4, and 5 as `Class 1` versus all other non-sitting activities.
- **Lying detection**: Aggregates all lying activities 10 to 13 as `Class 1` versus all other non-lying activities.
- **Stand detection**: All standing activities, 6 to 9 as `Class 1` versus all other non standing activities.

Strategy B: Specific binary comparison models In these experiments, the dataset is filtered to include only the relevant classes, ignoring the rest. This approach evaluates the fine-grained resolution of the Transformer architecture when separating physically similar or logically opposite activities. The models in this category include:

- **Lying/sweeping**: Activities 10 to 13 versus activity 17.
- **Dynamic activities/static activities**: All dynamic activities (3, 14, 15, 16, 17) versus sitting (1, 2, 4, 5), standing (6, 7, 8, 9), lying (10, 11, 12, 13).
- **Lying/not lying**: Lying (10, 11, 12, 13) versus all other upright activities (sitting/standing).
- **Walk/run:** Activity 15 versus activity 16.
- **Running/sweeping:** Activity 16 versus activity 17.
- **Standing still vs. all other stationary activities**: Standing (activities 6–9) versus other static activities (sitting/lying), often referred to as "still" activities.

To further expand the depth and value of this study, two additional models were developed by modifying the baseline

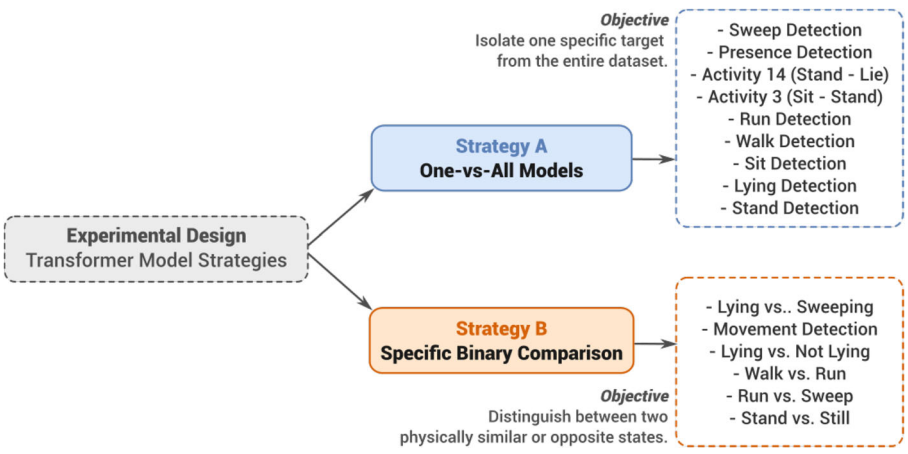


Fig. 5 Strategy A and Strategy B summarized

sweep. detec. and movement detec. models. The first new model, Sweep. detec. Imbalanced, was designed to evaluate the DFS-CSI's performance on natural, unequal data distributions by intentionally bypassing the data balancing step. Normally, the DFS-CSI mitigates bias by shuffling the processed samples and truncating the larger class to match the smaller one, ensuring an equal distribution before labeling. By skipping this truncation, the imbalanced model provides critical insights into how the Transformer architecture handles realistic, skewed datasets.

Following rigorous testing, a window size of 16 was selected for the second model, Mov. detec., because it achieved the highest overall accuracy. Together, these strategic adaptations significantly strengthen the research by demonstrating the model's adaptability to both unadjusted real-world data and highly optimized time windows.

The Transformer model is divided into encoder layers, as depicted in Fig. 6. The encoder consists of several identical stacked layers, where each layer follows a series of steps. First, the multi-head self-attention mechanism [17] enables the model to learn relationships between different moments in the sequence. Each attention head processes a distinct part of the information.

Then, the model calculates three tensors: Q (query), which represents the elements querying other moments in the sequence; K (key), which represents elements that can provide relevant information; and V (value), which contains the actual values of the features to be processed [37]. The similarity between Q and K is computed to identify which moments in time are most relevant to each other. The output is a weighted combination of V, with more important moments receiving higher weights.

The output of the self-attention block is passed through a normalization layer (LayerNorm) to stabilize training. A residual connection is then applied, adding the original input to the self-attention output, helping to maintain more stable gradients. The resulting output is then processed through a feed-forward network (FFN) [17, 18] to learn non-linear patterns. This network consists of a dense layer (Linear), followed by an activation function using ReLU, and another dense layer to adjust the dimension of the output [37]. A second normalization layer stabilizes the output of the feed-forward network, and another residual connection adds the block's input to the final output.

After that, each instant of the sequence is refined, incorporating global information from the entire time series. Now the values represent rich, refined representations of the sequence.

Summarizing the steps of the process: the input of the data with positional encoding, followed by multi-head self-attention to find temporal dependencies, then Normalization and residual connections for stabilization, followed by the feed-forward network to capture more complex patterns, additional normalization and residual connections, and repetition of the process. This structure allows the Transformer model to learn global and local patterns in time series without the limitation of recurrence (as in RNNs), while highlighting temporal relationships [17, 18].

To ensure the highest performance for each model, we conducted a rigorous hyperparameter tuning process and selected, for testing, the configuration that achieved high accuracy–loss balance during validation. The test stage evaluated the model on unseen data using metrics such as loss, accuracy, precision, recall, F1-score, and the confusion matrix. DFS-CSI optimizes a Transformer model for binary classification of CSI time series by exploring hyperparameter combinations through a random search strategy, assessing each configuration on the validation set and selecting the final model based on its test performance. The study analyzes how the number of heads, number of layers, learning rate, number of epochs, and batch size influence performance. The following ranges were explored: **heads**: [3–16]; **layers**: [2–12]; **learning rate**: [0.00001, 0.0001, 0.001, 0.01]; **epochs**: [20, 40, 60, 100]; and **batch size**: [4, 8, 12, 16, 20, 24]. Table 4 summarizes the best configuration identified for each model.

The heads and layers columns in Table 4 specify the number of attention heads and Transformer layers for each task. In general, more complex tasks, such as sit and lying detection, use more heads and intermediate layers, indicating the need for greater attention capacity to capture nuances in activities. On the other hand, simpler tasks, such as presence and motion detection, use a lighter configuration. The learning rate is adjusted for each model.

The wide variation in the number of layers (ranging from 2 to 9) across the binary experts is a direct reflection of the varying physical complexities inherent to each specific classification task. Highly dynamic actions, such as sweeping or running, produce distinct and easily separable CSI energy signatures, allowing shallow networks (2

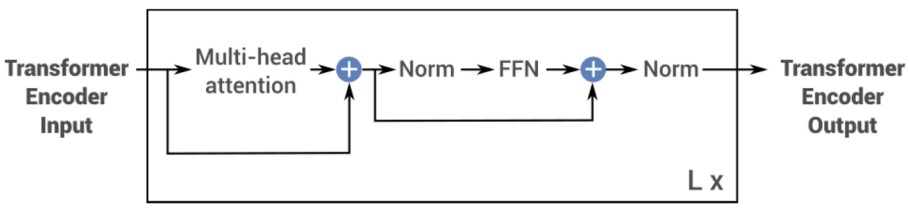


Fig. 6 Transformer encoder layers

Table 4 Best parameters combinations for each model

Activity	Heads	Layers	L. rate	Epochs	Batch size
Sweep. detec. imbalanced	3	8	0.00100	40	4
Mov. detec. window 16	6	9	0.00001	120	12
Sweep. detec	3	2	0.00100	40	24
Movement detec	6	2	0.00100	100	8
Presence detec	6	6	0.00010	20	12
Standing-to-lying detec	6	2	0.00100	20	8
Sit-to-stand detec	12	9	0.00001	40	24
Lying/sweep	3	6	0.00010	40	16
Run. detec	4	2	0.00100	60	8
Walk. detec	6	3	0.00100	100	20
Lying/not lying	4	2	0.00001	20	16
Walk./run	4	2	0.00100	40	16
Sit. detec	4	2	0.00100	100	8
Run./sweep	4	2	0.00010	40	4
Lying detec	6	9	0.00010	60	12
Stand./still	3	9	0.00001	20	4
Stand. detec	3	9	0.00001	100	4

layers) to achieve high accuracy. Conversely, distinguishing subtle static postures or transitional movements requires much deeper feature extraction (up to 9 layers) to capture complex, non-linear relationships. It is essential to emphasize that this variation does not reflect test-split overfitting. The hyperparameter random search and selection were conducted strictly using the training and validation sets. The test set—comprising data from 9 isolated volunteers—remained completely unseen during the tuning phase and was evaluated only once at the final stage. This guarantees that the models genuinely generalized their learning rather than memorizing the test data.

Instead of applying a monolithic architecture, we tailored the Transformer structure, specifically the number of attention heads and encoder layers, to match the complexity of the target activity. This adaptive approach yielded significant insights:

- Architectural depth: Activities with distinct high-energy signatures, such as lying, sweeping, or running, converged effectively with shallower networks, 2 layers. In contrast, more subtle or transitional activities, like activity 3, sit-to-stand, and stand detection, required deeper architectures (up to 9 layers) and a higher number of attention heads (up to 12) to capture the intricate temporal dependencies in the CSI signal.
- Learning dynamics: The learning rate and batch size were also optimized to balance convergence speed with stability. Complex tasks often required lower learning rates (e.g., 10^{-5}) to avoid overshooting the optimal solution during training.

This customization confirms that, although the Transformer framework is broadly applicable to CSI data, its specific configuration plays a crucial role in maximizing both accuracy and resource efficiency.

A crucial design choice in our methodology was the application of 1D attention over the temporal sequence, treating the selected subcarriers as independent feature channels, rather than employing a 2D attention mechanism [38]. 2D attention is often utilized to rigorously capture the underlying physics of wave propagation by exploiting both spatial and frequency correlations simultaneously [38, 39]. However, this is primarily applicable in multi-antenna (MIMO) setups where a true physical spatial matrix exists.

Our collection architecture intentionally relies on a strict 1x1 Single-Input Single-Output (SISO) configuration using the single internal antenna of the Raspberry Pi 4B. Because the physical spatial dimension is strictly limited to a single stream, applying a 2D attention mechanism would result in an artificial spatial-temporal flattening. This would introduce unnecessary computational complexity and inflate the parameter count without effectively modeling true multi-path spatial propagation. By processing the subcarriers as independent channels in a 1D sequence, the model architecture is perfectly aligned with the hardware's physical limitations. This preserves a highly compact architecture, avoids the heavy processing overhead of multi-dimensional attention

maps, and aligns with the strict low-latency requirements of our Green AI edge computing framework.

3.3.2 Temporal windowing strategy for enhanced movement detection

In the initial implementation of DFS-CSI, the models were trained using a fixed input sequence of 2000 samples, representing approximately 60 s of continuous CSI data collection. Despite this approach effectively captures long-term dependencies, it introduces a significant latency constraint for real-time applications, as the DFS-CSI must wait for a full minute of data buffering before performing an inference.

To address this limitation and improve the model's responsiveness, we developed an enhanced movement detection model based on a temporal windowing strategy. In this approach, the continuous CSI stream is segmented into smaller, distinct time windows. This allows the Transformer model to process shorter sequences, thereby reducing the time required for detection and enabling near real-time analysis.

We investigated the impact of sequence length on classification performance by evaluating five distinct window sizes (W). The selected window sizes and their corresponding temporal durations (based on the sampling rate) are defined as follows:

- 0.5 s: $W = 16$ samples;
- 1 s: $W = 33$ samples;
- 5 s: $W = 166$ samples;
- 15 s: $W = 500$ samples;
- 30 s: $W = 1000$ samples.

This segmentation allows us to analyze the trade-off between temporal resolution and classification accuracy. Smaller windows (e.g., 0.5s or 1s) offer lower latency and faster detection of sudden movements, whereas larger windows (e.g., 15s or 30s) provide more contextual information, potentially improving the model's ability to distinguish complex but subtle motion patterns. The modified model architecture adjusts the input dimension of the Transformer Encoder to match the selected window size W, while maintaining the hyperparameters consistent with the new movement detection model.

Figure 7 illustrates the systematic segmentation approach applied to the CSI data stream. The segmentation process strictly employs non-overlapping windows to maintain data integrity and prevent information leakage between samples. This rigorous strategy allows for a detailed analysis of how temporal resolution impacts feature extraction and classification accuracy. By sweeping through sizes ranging from instantaneous snapshots to extended observation periods, the study validates the DFS-CSI's robustness across different dynamic scenarios.

Applying a multi-layer Transformer (up to 9 layers) to a brief sequence of 16 packets may initially seem structurally dense, but it is theoretically justified by the complexity of the Wi-Fi signal. Even within a 0.5-s window, the interactions among the 48 active subcarriers contain intricate frequency variations caused by micro-movements. Deep attention layers allow the model to cross-correlate these subtle subcarrier variations instantaneously. During the design phase, alternative temporal architectures, such as Gated Recurrent Units (GRUs) [40] and Temporal Convolutional Networks (TCNs) [41, 42], can be considered. However, GRUs process data sequentially, which creates a bottleneck for parallel execution on edge hardware [40, 43]. The mathematical foundation of a GRU ($h_t = f(h_{t-1}, x_t)$) [44] strictly forces sequential execution. On edge devices like the Raspberry Pi, this sequential bottleneck prevents modern parallel matrix multiplication, increasing latency.

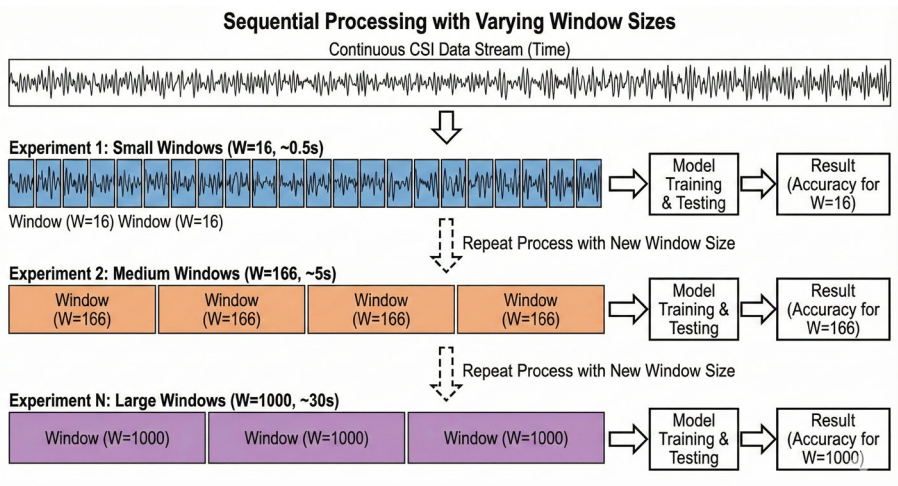


Fig. 7 Sequential processing framework for non-overlapping windowing with different sizes

While TCNs offer parallelization, they require deeply stacked hierarchies to expand their temporal receptive field [41, 42]. Their mathematical operation relies on static convolutional kernels [44]. In contrast, the Transformer's self-attention dynamically computes correlation weights (attention scores) based on the specific input data of that exact instant [2, 17]. This dynamic weighting is superior for capturing the sudden, irregular subcarrier variations caused by micro-movements. The self-attention mechanism of the Transformer provides an immediate global receptive field across the entire 16-packet window [38, 45]. This ensures maximum feature extraction parallelization and ultra-low latency, making it the superior choice for fast response systems. The self-attention mechanism of the Transformer mathematically computes dynamic correlation weights [2], generating an immediate global receptive field across the entire 16-packet window.

Instead of sliding a fixed kernel, the model dynamically correlates every packet with every other packet in the micro-sequence simultaneously. This dynamic feature extraction, combined with full matrix parallelization, ensures that the `DFS-CSI` achieves ultra-low latency processing without the sequential delays of GRUs or the static limitations of TCNs, making it the optimal structural choice for real-time edge computing.

3.3.3 Train, validation, and testing

Our main goal is to recognize human physical activities, for which we developed 17 independent models that take as input the previously described collected and processed CSI data. The dataset was partitioned into three sets: training (70%) with 60 volunteers, validation (20%) with 17 volunteers, and test (10%) with 9 volunteers, using a subject-independent strategy. This means that data from each individual is exclusively assigned to one of the subsets, preventing the same data from being used in both training and testing phases. This rigorous approach ensures that the model learns to generalize human activities across different users rather than memorizing individual-specific patterns, providing a more realistic evaluation of performance on unseen data. The models were developed to detect and classify a variety of human activities within a room, including presence detection, movement detection, gait classification (walking vs. running), and activity classification (sitting vs. lying down). Each of the generated models uses specific activities from the dataset. These models perform binary classification, yielding only two possible outputs.

The following step involves defining the loss function and optimizer strategies. We employed the binary cross-entropy with logits loss function for the classification tasks, which ensures numerical stability by combining a Sigmoid layer with the cross-entropy loss. For optimization, the adaptive moment estimation (Adam) algorithm was selected due to its efficiency. Additionally, a plateau-based learning rate scheduler was implemented to dynamically reduce the learning rate whenever the validation performance stops improving.

For the generation and training of the presence detection model, data segmentation of the CSI was necessary. This involved selecting data from empty rooms and rooms with human presence. Additional collection was used for empty rooms, while data from all 17 activities were used for rooms with human presence. The data were labeled as: no presence for empty rooms and presence for rooms with human presence.

The experiments were implemented using the Anaconda distribution with Python v3.8.18. Key libraries included *numpy* 1.24.4, *pandas* 2.0.3, and *pytorch* 2.2.0. Device specifications: Processor 13th Gen Intel(R) Core(TM) i7-13700K (3.40 GHz), installed RAM 32.0 GB (31.7 GB usable), Windows 11, NVIDIA GeForce RTX 3080, Driver Version: 581.57, CUDA Version: 13.0.

The `DFS-CSI` does not rely exclusively on GPUs; it was developed to allow smooth execution on conventional CPUs. This ensures that the technology is accessible to a wider range of users, eliminating the need for immediate investments in high-cost hardware. Independence from specific hardware allows the `DFS-CSI` to operate in diverse environments, from high-end servers to local devices. This allows for continuous monitoring even on systems without graphics acceleration.

4 Results

This section presents and discusses the experimental outcomes obtained by evaluating the proposed Transformer models across various activity recognition scenarios.

4.1 Performance analysis

This section presents the performance evaluation of the developed Transformer-based models for HAR using CSI data. Table 5 summarizes the results across five key metrics: accuracy, precision, recall, test loss, and F1-score.

The results indicate that the Transformer architecture is highly effective for distinguishing between distinct dynamic activities, though it faces challenges with static behaviors that share similar CSI signatures. This suggests that the CSI waveforms for specific activities are distinct enough, allowing the model to classify perfectly.

The sweeping detection also yielded robust results. Similarly, movement detection and presence detection maintained high accuracy levels, validating the use of CSI data for occupancy and general activity monitoring.

Figure 8 illustrates the training dynamics for the binary classification between lying and sweeping. This scenario

Table 5 Performance metrics of each model for a test set of 9 participants

Activity	Accuracy	Precision	Recall	Loss	F1-score
Sweep. detec. (imbalanced)	94.64%	94.64%	100.00%	0.6892	0.9725
Mov. detec. window 16	92.32%	95.04%	89.30%	0.2244	0.9208
Sweep. detec	88.89%	81.82%	100.00%	0.4044	0.9000
Movement detec	81.48%	86.96%	74.07%	0.5190	0.8000
Presence detec	87.50%	87.50%	87.50%	0.5756	0.8750
Standing-to-lying D	83.33%	87.50%	77.78%	0.5325	0.8235
Sit-to-stand D	83.33%	75.00%	100.00%	0.6655	0.8571
Lying/sweep	83.33%	80.00%	88.89%	0.9818	0.8421
Run. detec	83.33%	80.00%	88.89%	0.5798	0.8421
Walk. detec	77.78%	85.71%	66.67%	0.6922	0.7500
Lying/not lying	73.61%	70.73%	80.56%	0.6758	0.7532
Walk./run	72.22%	83.33%	55.56%	0.6848	0.6667
Sit. detec	72.22%	70.00%	77.78%	0.6051	0.7368
Run./sweep	72.22%	83.33%	55.56%	0.6554	0.6667
Lying detec	68.06%	84.21%	44.44%	0.6151	0.5818
Stand./still	65.28%	67.74%	58.33%	0.6798	0.6269
Stand. detec	63.89%	62.50%	69.44%	0.6946	0.6579

represents one of the most successful applications of the proposed model, primarily due to the significant physical disparity between the two activities.

Figure 9 presents the results for the movement detection model, which is designed to distinguish between general human activity and static presence.

The training loss curve (third graph) indicates a smooth and steady convergence, demonstrating that the Transformer architecture effectively learns the underlying patterns of human motion from the CSI data. While the validation metrics exhibit a few fluctuations, characteristic of the high variability due to the 86 different participants, the model stabilizes at a high accuracy level.

The confusion matrix indicates that the model maintains a satisfactory level of performance, making it highly suitable for many practical applications. This detector achieves high accuracy for both Class 0 and Class 1. This model proves to be a reliable baseline for presence detection, capable of robustly filtering out static scenarios before more complex activity classification is performed.

4.1.1 Performance analysis of movement detection with windows

To validate the feasibility of real-time monitoring, we analyzed the impact of different window sizes on classification accuracy. Figure 10 illustrates the test accuracy obtained across different temporal window lengths, ranging from 0.5 s (16 samples) to 30 s (1000 samples).

The results reveal a non-linear relationship between observation time and model performance. Contrary to the intuition that more data (larger windows) yields better predictions, the model achieved its peak performance with the smallest window size of 0.5 s ($W = 16$). Specifically, this configuration yielded an accuracy of 92.32%, a precision of 95.04%, and an F1-score of 0.9208.

As shown in Fig. 10, performance drops significantly at $W = 33$ (1s) and $W = 1000$ (30s), while showing a secondary peak at $W = 166$ (5s).

Consequently, the 0.5-s window proves to be the optimal configuration, offering the highest accuracy combined with

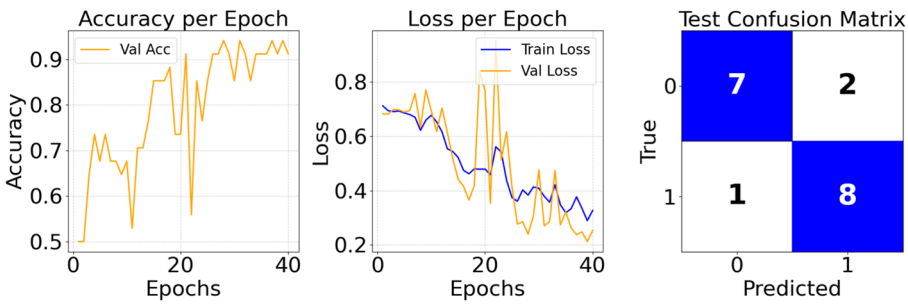


Fig. 8 Classification performance for lying versus sweeping

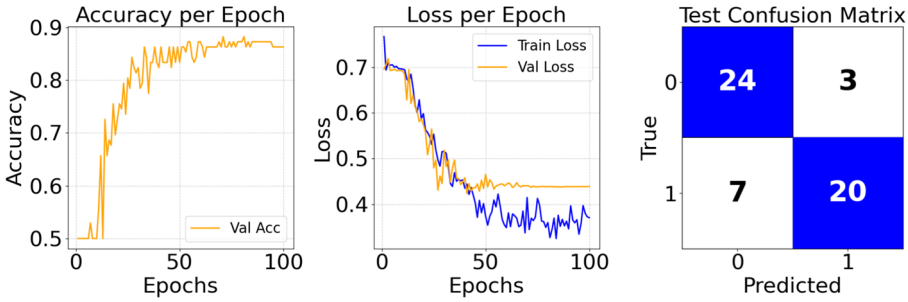


Fig. 9 Training performance and classification results for the movement detection model (movement vs. still)

the lowest possible latency, making it ideal for immediate event detection.

Figure 11 details the learning process and final classification performance of the window-based movement detection model. The training dynamics reveal a highly efficient optimization profile. As observed in the training loss and validation accuracy plots, the model achieves rapid convergence, reaching a high accuracy plateau within the first 20 epochs. This behavior confirms that the reduced temporal window (0.5s) contains sufficient spectral information for the Transformer to quickly distinguish between static and dynamic states.

Although the validation loss exhibits some volatility in the initial epochs, a characteristic expected when generalizing across a diverse dataset of 86, it stabilizes significantly as training progresses, indicating that the model successfully avoids overfitting.

The confusion matrix further corroborates the `DFS-CSI`'s reliability. The model demonstrates a balanced predictive capability, achieving high true positive and true negative rates (correctly classifying 3218 static instances and 3014 movement instances). The low rate of misclassifications, particularly the balanced distribution of false positives and false negatives, underscores the model's robustness and its readiness for deployment in real-world scenarios where consistent detection is critical.

4.1.2 Computational resources analysis

In addition to classification performance, evaluating the computational cost is essential for determining the feasibility of deploying HAR systems in real-world scenarios. Table 6 summarizes the resource consumption for the developed models, specifically focusing on memory usage, execution time, and power draw during both training and testing phases.

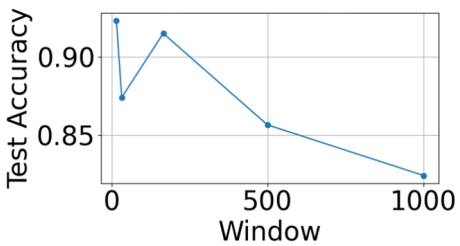


Fig. 10 Test accuracy variation according to the temporal window size (W) for the movement detection model

While classification accuracy is the primary metric for evaluating Deep Learning models, it does not fully determine the practical viability of a system. In the context of HAR using CSI, the models are often intended for deployment on edge devices or embedded systems (such as Wi-Fi routers or IoT gateways), which have strictly limited computational resources.

Therefore, this study goes beyond standard performance metrics to include a comprehensive analysis of computational cost. We evaluate memory usage, training/inference time, and power consumption to ensure that the proposed Transformer architecture is not only accurate but also efficient. This holistic approach is crucial for two main reasons:

1. **Real-time feasibility:** Applications such as fall detection or security monitoring require near-instantaneous response times. Analyzing the inference time allows us to verify whether the model can process signals with low latency.
2. **Energy efficiency:** Continuous monitoring systems operate 24/7. Understanding the power draw is essential for developing sustainable "Green AI" solutions that minimize energy costs and thermal stress on hardware.

The following data present the resource consumption profile of the developed models, demonstrating that high accuracy can be achieved without imposing prohibitive computational overhead.

The results indicate a computationally diverse profile across different activity classification tasks. The **total training time** varied significantly across models. This substantial variation reflects the differences in dataset complexity, temporal window size, and the specific feature extraction challenges posed by each activity type. Models dealing with larger temporal windows or more subtle distinctions between activities naturally required longer training periods to achieve optimal convergence.

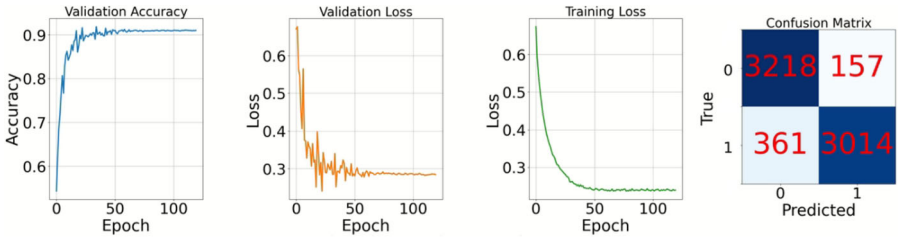


Fig. 11 Training dynamics and evaluation metrics for the movement detection model using the optimal window size ($W = 16$). From left to right: validation accuracy, validation loss, training loss, and confusion matrix

Regarding inference latency (test time), the models demonstrated exceptional speed across all tasks. For instance, the **sweeping detection** model required only 0.16 s to process the test set, while the **standing detection** model completed inference in just 1.12 s. Even the more complex models, such as **movement detection**, completed the inference phase in under 3 s (2.74s). These low-latency values are critical, as they confirm that the DFS-CSI is capable of near real-time activity recognition, a mandatory requirement for applications such as fall detection or security monitoring.

In terms of hardware load, the memory usage varied between 4.5 GB and 9.9 GB, which is within the standard capacity of modern embedded AI accelerators and consumer-grade GPUs. The power draw remained relatively stable during inference (typically between 55W and 224W), and the GPU temperature remained well-controlled (ranging from 52ºC to 68ºC), indicating that the models do not cause excessive thermal stress on the hardware during operation.

A notable observation is the relationship between power draw and execution time. Models such as **running/sweeping** and **running detection** operated at a much lower power range during training, demonstrating that the architecture can adapt to different energy constraints depending on the complexity of the feature space being learned. Dynamic activities with distinct signatures allow for more efficient feature extraction, resulting in lower computational overhead.

Finally, the thermal analysis confirms the safety and stability of the proposed solution. Across all experiments, including those with high power peaks, the GPU temperature never exceeded 70ºC, staying well within safe operating limits (typically rated up to 80–90ºC for modern GPUs) [46]. This thermal stability, combined with memory usage peaking at around 9.9 GB, validates that the developed Transformer models are suitable for deployment on standard commercial hardware without requiring specialized industrial cooling solutions or excessive energy infrastructure.

4.1.3 Streaming latency and model complexity

To fully substantiate the claims regarding real-time processing and Green AI capabilities, it is essential to distinguish between batch inference time and continuous streaming latency. While Table 6 reports the total time to process the entire test set (batch processing), real-world deployment operates on a continuous data stream where latency must be measured per individual window.

Based on our experiments, the DFS-CSI demonstrates exceptional streaming performance. For instance, the *movement detec. window 16* model processed a total of 6232 test samples in 5.93 s. This translates to an individual streaming latency of approximately **0.95 ms per inference**. Because the DFS-CSI takes only $\sim$ 1 ms to process a data window that represents 500 ms of physical time ($W = 16$), the processing speed is orders of magnitude faster than the data generation rate. This absolute lack of computational limitation mathematically guarantees smooth, real-time streaming capabilities without queue delays.

Furthermore, analyzing the model size and computational cost is critical for Edge AI deployment. Unlike massive deep learning architectures that require cloud offloading, our custom Transformer models were designed to be lightweight. The optimal movement detection model possesses a highly compact architecture, resulting in a minimal number of learnable parameters (approximately **0.62 million**) and an exceptionally small storage use of exactly **2.50 MB**.

This minimal model size directly supports the Green AI paradigm, which prioritizes the reduction of computational cost, energy consumption, and carbon footprint alongside accuracy [47]. By drastically reducing the required floating-point operations per second (FLOPs) and minimizing memory bandwidth usage during inference, the DFS-CSI draws less power per classification [48]. Consequently, it can be seamlessly embedded into standard IoT gateways, routers, or low-power edge devices, providing highly accurate, privacy-preserving monitoring with a negligible energy footprint.

To fully contextualize the architectural efficiency of the proposed DFS-CSI, it is important to contrast its complexity with other Transformer-based frameworks recently applied to Wi-Fi CSI human activity recognition. A common trend in the literature involves transforming 1D CSI time-series data into 2D pseudo-images (spectrograms) to leverage standard Vision Transformers (ViT) [45, 49]. However, architectures such as the standard ViT-Base require approximately 86 million parameters, demanding immense memory bandwidth and processing power [49–51]. In contrast, our custom 1D Transformer approach processes the CSI sequences directly, requiring only 0.62 million parameters.

Table 6 Consumption results

Activity	GPU memo. used (MB) Train/test	Total time (s)		Power draw (W)		Util. (%)		Fan speed (%)		Temp. (ºC)	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Lying/sweep.	7871	584.97	2.69	86.60	58.87	21	0	30	30	57	57
Sweep. detec. imbalanced	9907	819.02	17.51	127.70	127.18	0	0	34	34	61	60
Mov. detec. window 16	9886	12178.28	5.93	91.57	71.76	0	0	30	30	53	52
Sweep. detec.	6841	77.11	0.16	81.72	76.37	44	29	75	75	59	58
Movement detec.	9954	2562.65	2.74	153.69	106.81	0	0	39	39	62	62
Presence detec.	7598	155.45	2.67	105.36	94.37	0	23	38	38	62	60
Activity 14 detec.	4519	120.22	2.58	98.06	78.00	13	23	30	30	56	55
Activity 3 detec.	9839	513.40	2.87	196.41	131.62	0	36	44	44	63	62
Run. detec.	6362	361.58	2.51	75.03	72.45	0	47	30	30	58	57
Walk. detec.	4505	650.25	2.55	112.53	92.97	12	7	33	33	60	60
Lying/not lying	9945	439.16	2.83	158.58	100.37	0	1	39	39	62	62
Walk./run.	9930	383.36	2.86	90.90	76.25	0	30	32	32	60	60
Sit. detec.	9965	1270.68	3.73	160.90	157.92	43	22	40	40	62	62
Run./sweep.	9912	372.81	2.50	47.77	75.31	16	0	30	30	59	58
Lying detec.	7233	1298.25	3.15	264.04	223.73	0	0	58	58	68	67
Stand./still	8702	312.34	2.91	233.88	139.20	0	52	50	50	65	64
Stand. detec.	5723	3655.67	1.12	63.94	55.31	98	88	63	62	55	55

This means the DFS-CSI operates with less than 1 % of the parameter count of traditional ViT adaptations, proving that high-accuracy CSI classification can be achieved without the extreme computational overhead typically associated with massive attention-based models.

4.1.4 Operational efficiency and trade-off analysis

Figure 12 provides a multi-dimensional perspective on the operational costs of the developed models. By correlating inference latency (x-axis) with power consumption (y-axis), we can identify the optimal deployment scenarios for each activity classifier.

The visualization highlights three distinct operational zones:

1. **High-speed zone (bottom-left):** The *sweep detection* model stands out as the fastest solution, located in the extreme bottom-left corner. With sub-second latency and low-power draw, this model is ideal for battery-powered edge devices where rapid response is critical.
2. **Balanced efficiency zone (center-left):** Models such as *stand detection* occupy the middle ground, offering a robust balance between energy consumption and processing speed. This efficient zone represents the sweet spot for continuous monitoring systems connected to standard power supplies.
3. **High-performance zone (top-right):** The *lying detection* model appears in the upper-right quadrant, characterized as High Load. While this model demands higher power and memory (indicated by the larger red bubble), it is important to note that this resource intensity correlates with the complexity of distinguishing subtle static activities. Crucially, even at this peak load, the thermal signature remains within safe operating limits.

Overall, this analysis confirms that the proposed Transformer architecture is not a simple solution but a versatile framework. It can be tuned to prioritize either speed or energy efficiency, depending on the specific application requirements of the IoT environment.

4.1.5 Thermal stability and hardware safety

Figure 13 presents a critical evaluation of the DFS-CSI's physical impact on the processing hardware. For continuous monitoring applications, such as elderly care or security surveillance, thermal stability is as important as algorithmic accuracy, as overheating can lead to hardware throttling or permanent damage.

The results demonstrate that the proposed Transformer models are thermally efficient. The peak temperature recorded across all experiments was approximately 68°C. This value is significantly lower than the standard safety limit of 80°C usually recommended for commercial GPU operations [46]. Most activities, such as *activity 14* and *stand detection*, maintained the hardware temperature in the mild range. This substantial safety margin implies that the DFS-CSI can be deployed for extended periods.

4.1.6 Energy efficiency and green AI analysis

Figure 14 details the power draw (in Watts) for each activity model during the inference phase. This analysis is

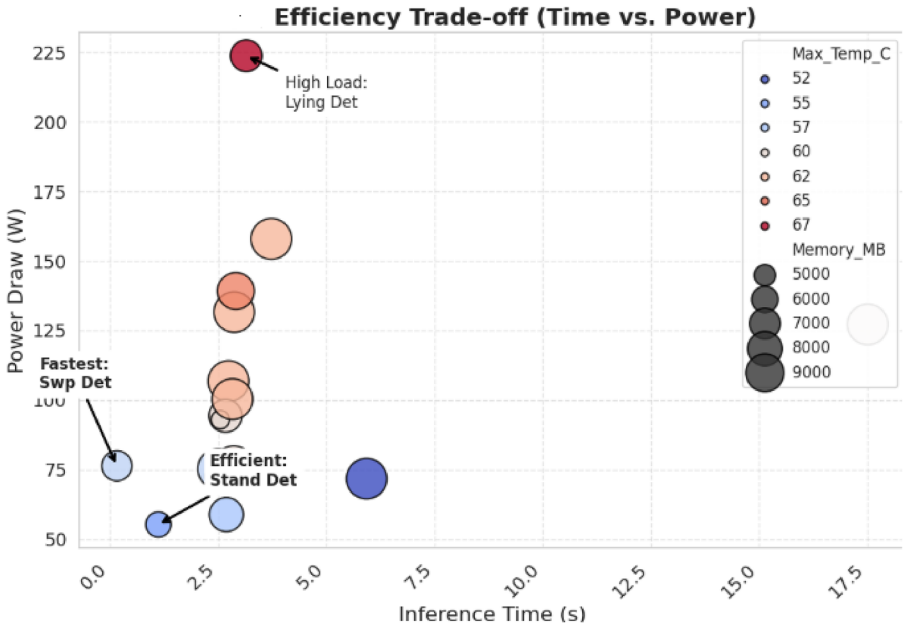


Fig. 12 Efficiency trade-off map (inference time vs. power draw). The bubble size represents memory usage, while the color gradient indicates peak temperature (blue=cooler, red=warmer). Annotations highlight key performance extremes

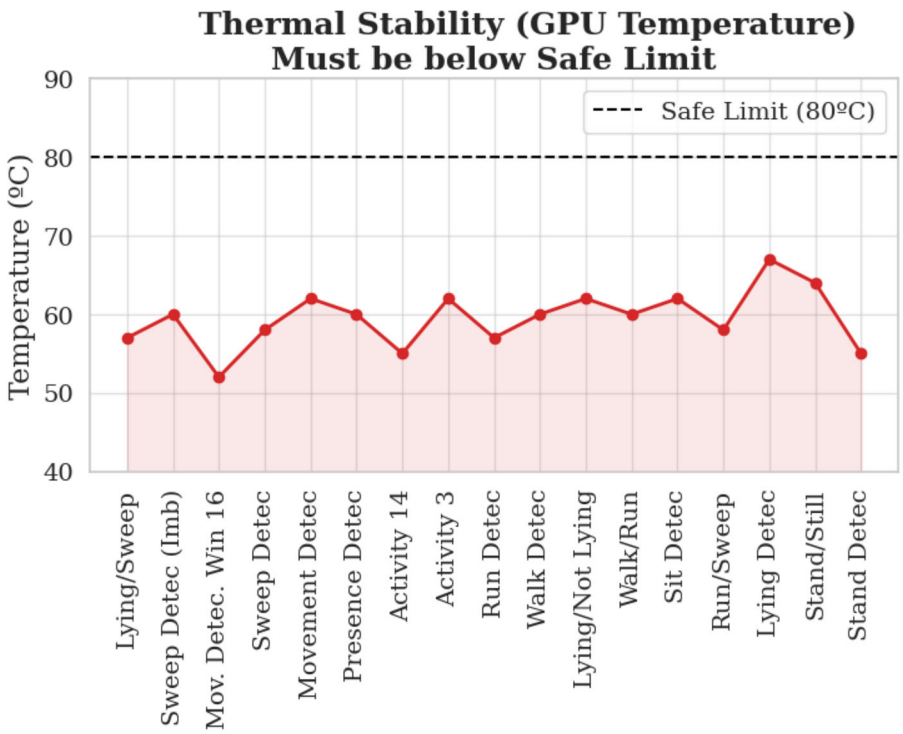


Fig. 13 Thermal stability analysis across all developed models. The red line indicates the peak GPU temperature recorded during operations, remaining consistently below the safety threshold of 80°C

particularly relevant in the context of Green AI, where the objective is to minimize the carbon footprint of artificial intelligence systems without compromising performance.

The data reveal a heterogeneous energy profile. Ideally, simpler detection tasks such as *stand detection* and *lying/sweep* demonstrate exceptional efficiency. These low-power values suggest that these specific models are highly suitable for deployment on energy-constrained edge devices. For the majority of the developed models, the power consumption remains below or near 100W, confirming the overall energy viability of the proposed `DFS-CSI` for continuous monitoring applications.

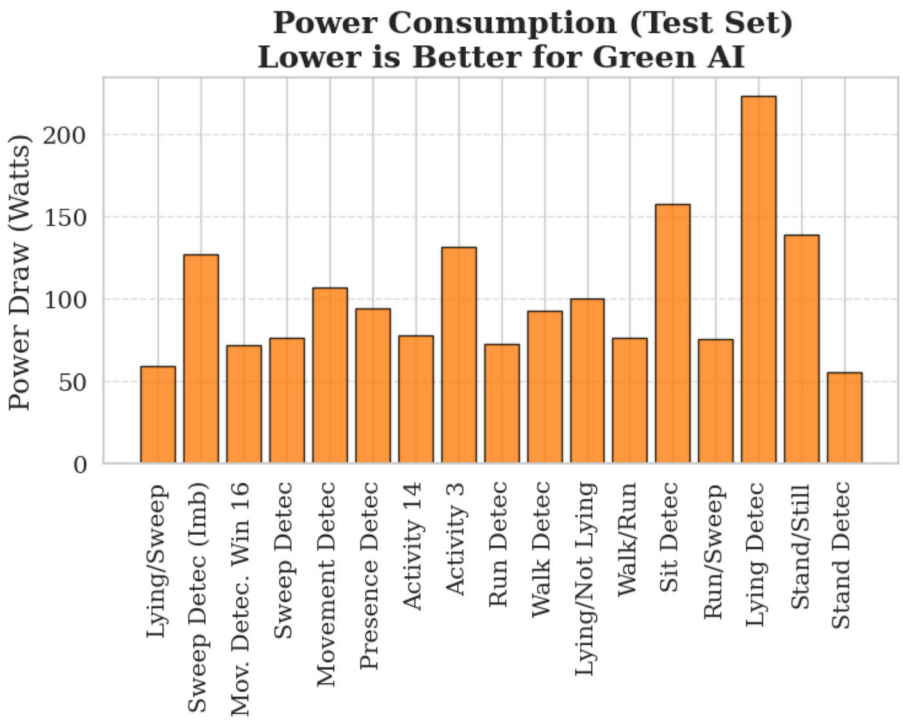


Fig. 14 Power consumption profile during the inference phase. The variability highlights the importance of optimizing models for Green AI, with several classifiers operating below the 100-Watt threshold

Beyond the inference phase, assessing the total energy overhead requires considering the power consumed during the continuous CSI extraction on the Raspberry Pi 4B [52]. Operating at an 80 MHz bandwidth, the continuous packet capture and NEXMON processing draw approximately 4.5 W to 5.0 W [52, 53]. While reducing the channel bandwidth to a standard 40 MHz alternative would slightly lower this extraction power, the 80 MHz configuration was deliberately selected to support our Green AI claims holistically [54]. The 80 MHz channel yields 256 subcarriers, providing the high-resolution frequency granularity essential for capturing micro-movements and vital signs [52]. If a 40 MHz channel were used, the resulting data would be sparser, which would force the neural network to be significantly larger and more computationally expensive to achieve the same accuracy [53, 54]. Therefore, the minor energy increase during data extraction is largely offset by enabling the use of highly efficient, lightweight inference models (requiring as little as 0.62 million parameters), maintaining the `DFS-CSI`'s overall sustainability.

We conducted real-time system resource monitoring on the Raspberry Pi during the 80 MHz data collection process using the vtop system monitor [55] to highlight the importance of quantifying the energy overhead during the continuous CSI extraction phase. As illustrated in Fig. 15 from captured profiling sessions, the computational footprint of the continuous CSI extraction is exceptionally low. Throughout the data collection, the overall system CPU usage remains stable and minimal, generally fluctuating below 15 to 18%, while total memory consumption remains constant at approximately 5% throughout the entire 80 MHz capture session, confirming that the extraction process introduces minimal energy overhead. Most importantly, the core processes responsible for the extraction (such as packet capture via tcpdump and kernel worker threads) demand insignificant processing power.

Because dynamic power consumption in ARM-based edge devices like the Raspberry Pi is strongly correlated with CPU utilization, this minimal computational overhead demonstrates that continuous CSI extraction adds an insignificant energy cost beyond the device's baseline idle power. This low resource profile also justifies our choice of the 80 MHz configuration over a 40 MHz alternative. While a 40 MHz bandwidth might theoretically reduce network interface activity slightly, the 80 MHz setting provides a significantly richer subcarrier resolution. This higher data quality improves the downstream AI model's accuracy and convergence without introducing a computational or energy bottleneck. Therefore, capturing at 80 MHz aligns perfectly with our Green AI claims, as it maximizes data richness for the model while maintaining edge resource usage at near-idle levels.

For full transparency and reproducibility, a complete video recording of the real-time resource profiling session during the CSI data collection is publicly available in our GitHub repository (https://github.com/mestrelan/MDA-CSI/tree/main/DFS-CSI/Vtop).

4.1.7 Real-time processing feasibility

Figure 16 evaluates the inference latency, a critical metric for time-sensitive applications such as fall detection or intrusion alerts. In these scenarios, the `DFS-CSI` must process CSI packets and generate a classification with negligible delay to ensure a prompt response.

The results highlight the exceptional performance of the *sweep detection* model, which stands out as the fastest classifier. It successfully operates, completing the inference phase in approximately 0.16 s. This near-instantaneous processing capability confirms that the Transformer architecture can be optimized for high-speed, real-time edge computing. Other models, such as *stand detection* (1.12s) and *run/sweep* (2.50s), also exhibit strong performance, indicated by the green bars in the chart. It is important to note that this duration corresponds to the processing of the entire test batch. Consequently, the per-sample latency remains sufficiently low to support continuous human activity monitoring without perceptible lag.

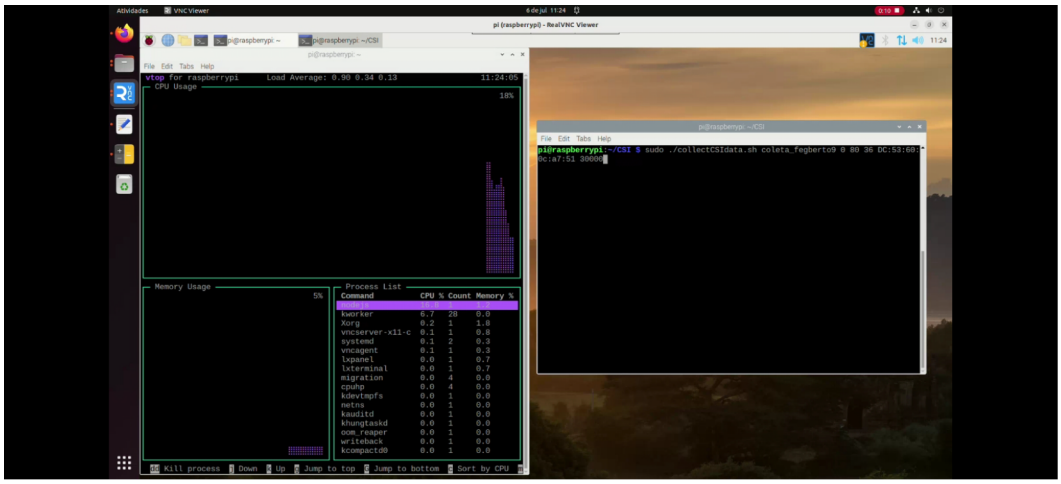
4.1.8 Efficiency of the movement detec. with windows

Beyond classification metrics, the *movement detec. with windows* model ($W = 16$) demonstrated a complex computational profile that warrants careful analysis for Edge AI deployment considerations. Table 7 details the resource consumption during the training and testing phases.

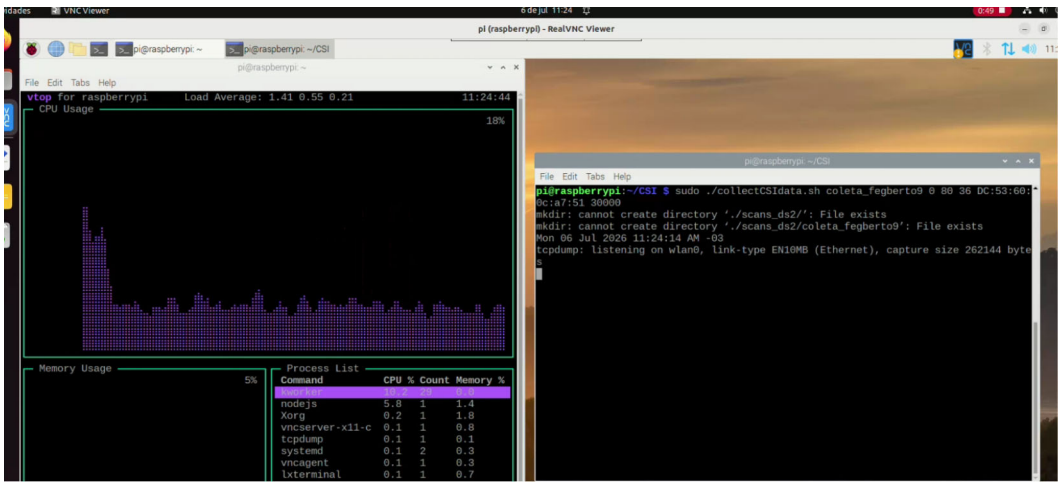
The model required approximately 12, 178 s (roughly 3.4 h) to complete the training process over 120 epochs. This extended training duration, significantly longer than most other models in this study, can be attributed to the large dataset size when processing CSI data with the optimized 0.5-s window. The fine temporal granularity, while beneficial for accuracy, increases the total number of training samples, demanding more iterations for the Transformer to fully converge across all 86 participants.

The test phase was completed in only 5.93 s, demonstrating exceptional real-time processing capability. This stark contrast between training and inference times is a hallmark of well-optimized deep learning systems: while the initial learning phase may be computationally intensive, the deployed model operates with minimal latency.

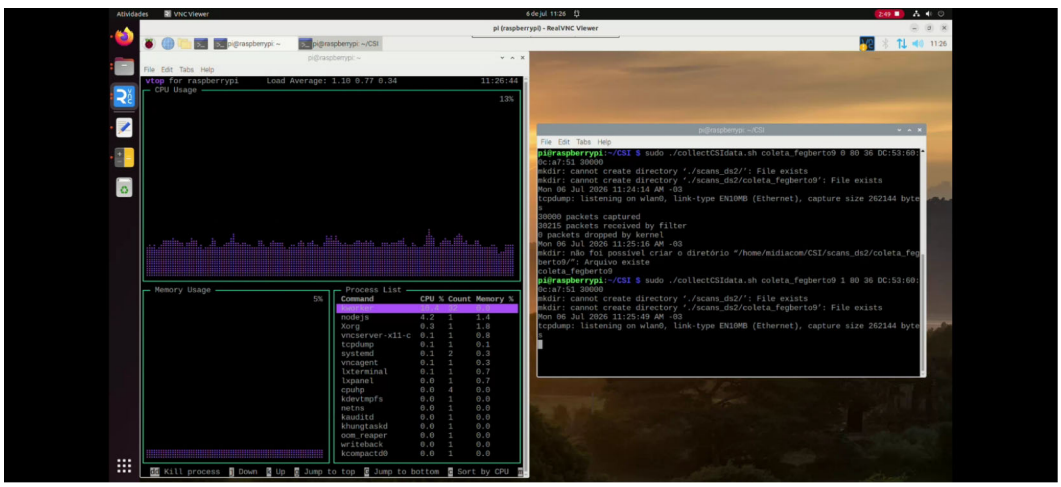
Regarding energy consumption, the inference phase recorded an average power draw of 71.76 W, which is significantly lower than the consumption observed in complex static



(a) Initialization phase.



(b) Mid-collection profiling.



(c) Final phase of collection.

Fig. 15 Real-time Raspberry Pi resource usage monitored via `vtop` during continuous CSI extraction at 80 MHz, showing minimal CPU and memory overhead throughout the process

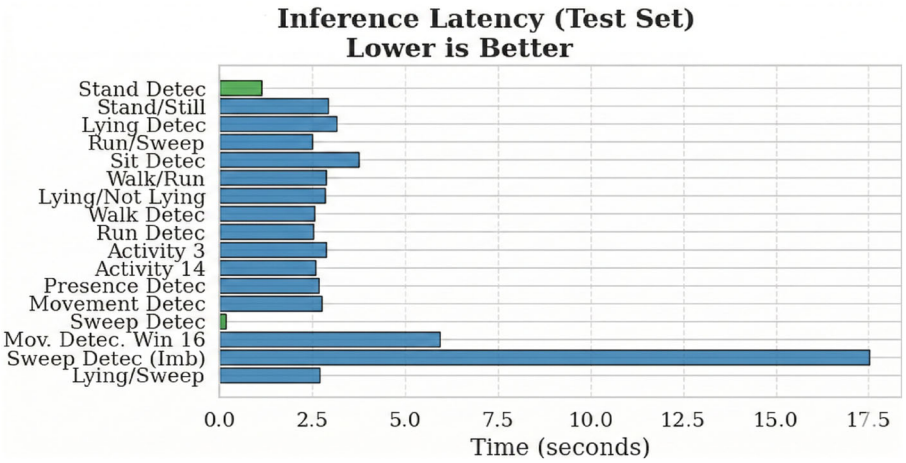


Fig. 16 Inference latency analysis in the test set. The sweep detection model demonstrates exceptional speed, processing the test stream almost instantaneously

activity models. This means that the model firmly meets the Green AI paradigm [56]. The training power draw of 91.57 W remains moderate and reflects the computational effort required to process the extensive windowed dataset.

The GPU memory usage of 9886 MiB (approximately 9.65 GB) is well within the capacity of modern commercial GPUs and even high-end embedded AI accelerators, confirming hardware compatibility for real-world deployment [46]. The combination of ultra-fast inference (< 6 s for the entire test set), high accuracy ($> 92\%$), low energy footprint during operation (< 75 W), and reasonable memory requirements validates that the *movement detec. with windows* model is highly scalable for continuous, real-world monitoring systems. While the training phase demands patience and computational resources, this is a one-time cost. Once deployed, the model delivers efficient, accurate, and energy-conscious performance suitable for 24/7 IoT applications.

4.2 Discussion on evaluation scope and real-world applicability

While the experimental results demonstrate the high accuracy and efficiency of the proposed `DFS-CSI`, it is important to critically discuss the evaluation scope and the steps required for broad real-world deployment.

Table 7 Computational resources for movement detec. with windows ($W = 16$)

Metric	Value
Total training time	12,178.28 s
Total test time	5.93 s
Inference power draw (Avg)	71.76 W
Training power draw (Avg)	91.57 W
GPU memory usage	9886 MiB
Test accuracy	92.32%

Classification strategy The current framework relies on a specialized binary classification approach rather than a single multi-class model. This design choice was intentional to maximize modularity and operational efficiency in edge computing environments. By deploying independent binary experts (e.g., first detecting presence, then detecting movement, and finally classifying specific actions like sweeping or running), the `DFS-CSI` becomes highly adaptable. In a real-world scenario, if a new activity needs to be monitored, a new lightweight binary classifier can be added without the computational burden of retraining a massive multi-class architecture from scratch. However, we acknowledge that a unified multi-class evaluation is a highly relevant approach for certain applications. Developing a hierarchical decision tree to integrate these binary experts, or training a unified multi-class Transformer, is a natural and necessary progression to enhance the `DFS-CSI`'s seamless applicability.

As detailed in our methodology (Strategy A), our primary models employ a "one-vs-all" approach. For instance, our "movement detection" and "sweep detection" models were trained to isolate the target state against a highly complex background comprising all other activities. The "one-vs-all" training strategy incorporates the full dataset's complexity into the binary models, and we highlight the practical versatility and computational benefits of deploying targeted binary experts for specific remote monitoring applications, such as elderly care. Therefore, these binary models successfully learned to distinguish the target action amidst the full variability of the entire dataset.

Additionally, in many real-world remote monitoring scenarios, highly specialized binary experts actually offer superior versatility. For example, in elderly care applications, caregivers often do not need a system that constantly computes a massive multi-class probability matrix to know the exact fine-grained action being performed every second. Instead, it is critically more important to have highly accurate, dedicated binary alarms for specific states (e.g., a highly

reliable "movement vs. still" detector to track daily activity levels, or a dedicated "lying" detector to trigger an alert if a patient is lying on the floor).

Additionally, this modular design allows administrators to deploy only the specific detectors needed for a given patient, minimizing the computational load on edge devices. If a new critical activity needs to be monitored, a single lightweight binary classifier can be added without the burden of retraining a unified multi-class architecture from scratch.

Nevertheless, a unified multi-class evaluation or a hierarchical decision tree is a highly relevant approach for universal applications and is a natural next step for our framework.

Environmental generalization Regarding the physical testing environment, the experiments were conducted in a single, controlled 3 m × 4 m room with a fixed furniture layout and specific router placement. This controlled methodology was crucial for the fundamental goal of this study: to isolate the human variable. By keeping the environment and hardware constant, we successfully proved that the DFS-CSI can achieve true *person-independent* generalization across a large and diverse dataset of 86 volunteers—a major challenge in CSI-based sensing.

Nevertheless, Wi-Fi CSI is by nature sensitive to multipath fading caused by room geometry, furniture displacement, and transceiver locations. Consequently, deploying the currently trained models in a completely different building or using different Wi-Fi chipsets without prior calibration would likely result in performance degradation. To achieve *environment-independent* generalization, future iterations of DFS-CSI must incorporate domain adaptation or transfer learning techniques. These strategies will allow the framework to map the learned human activity features into new environments and varying hardware setups using only a minimal calibration dataset, thereby solidifying its real-world scalability. However, we would like to emphasize that the current framework provides an exceptionally strong foundation for cross-environment deployment. Because our proposed Transformer models are highly compact and computationally lightweight (as demonstrated by our efficiency metrics), retraining or fine-tuning them for a new layout is fast and inexpensive.

Furthermore, since the models have already learned robust, person-independent motion features from 86 individuals, they serve as excellent pre-trained baselines. Deploying them in a new room or with different router placements would likely require only a minimal calibration dataset via transfer learning, drastically reducing the deployment effort compared to training from scratch. Therefore, our lightweight models serve as robust pre-trained baselines, making future environment-independent generalization highly feasible through minimal calibration and transfer learning.

5 Conclusions

This work presents a comprehensive framework for HAR using Wi-Fi channel state information (CSI) and Transformer-based deep learning models. By leveraging a large-scale dataset of 86 participants, we demonstrated that the self-attention mechanism is highly effective at capturing the complex temporal dependencies of human motion without the need for wearable sensors or invasive cameras.

The experimental results highlighted the robustness of the proposed architecture. The DFS-CSI achieved exceptional performance in detecting dynamic activities, particularly with the sweeping detection model, which reached near-perfect recall. Although distinguishing between subtle static activities (such as standing versus still) remains a challenge, the models provided reliable baselines for presence and general movement detection.

For future work, we aim to expand the DFS-CSI's capabilities to handle multi-occupancy scenarios, allowing simultaneous tracking of multiple individuals. Additionally, we plan to investigate techniques to ensure robust performance across different physical environments without the need for extensive retraining. Finally, we intend to explore the development of specialized models for healthcare monitoring, specifically focusing on the detection of sleep apnea through CSI patterns.

To further expand the classification framework in scenarios where computational resources are not a constraint, future versions of the DFS-CSI will explore advanced preprocessing and tokenization techniques. While the current system achieves high accuracy and ultra-low latency using amplitude-only CSI data and direct 1D sequencing, incorporating CSI phase integration presents an opportunity to provide richer spatial-temporal context for cloud-based evaluations.

Furthermore, to enhance the robustness of the proposed framework, recent optimizations were implemented and evaluated on the movement detection model with a window size of 16. First, advanced denoising techniques were integrated by applying a Hampel filter [24, 57] and a discrete wavelet transform (DWT) [24, 57] prior to model injection. This approach effectively removes high-frequency hardware noise while preserving sharp motion transitions. Second, an adaptive tokenization strategy was introduced to the network. Instead of mapping individual raw packets directly to tokens, the model now utilizes a localized temporal

patching mechanism through a 1D stride convolution [2, 49]. This modification feeds more stable temporal contexts into the Transformer's attention mechanism, reducing exposure to point-wise noise. Preliminary results indicate a clear improvement in classification performance. With these new implementations, the movement detection model achieved a test accuracy of 94.25% and an F1-score of 0.9426 (an increase from the previous baseline of 92.32% and 0.9208, respectively). These findings demonstrate that combining advanced signal filtering with temporal patching significantly enhances the model's ability to accurately interpret CSI data. We highlight these advanced techniques as a proven and immediate next step to enhance the framework's capability in future scaled applications.

Acknowledgements The authors would like to express their sincere gratitude to all the volunteers who participated in the data collection sessions. We also extend a special thanks to the students Fábio G. Queirós, Julio C. H. Soto, Dario Monte Vianna Pires Mousinho, and Egberto Caballero Rosillo for their valuable support and dedication during the data acquisition process.

Author contributions All authors contributed to the study conception. The data collection, the DFS-CSI development, and analysis were performed by Allan Costa Nascimento dos Santos (Ph.D Student). The first draft of the manuscript was written by Allan Costa Nascimento dos Santos and Pamella Soares (Ph.D Student). All other authors (Pamella Soares, Iandra Galdino, Taiane C. Ramos, Célio Albuquerque, Raphael Guerra, Natalia C. Fernandes, Débora C. Muchaluat-Saade, and Gheorghita Ghinea) revised previous versions of the manuscript. All authors approved the final manuscript.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CAPES Print, CNPq, FAPERJ, UFF, FINEP, INCT-MACC, and INCT-ICONIoT.

Data availability The raw CSI data and the statistical social data from the dataset are available for research purposes and can be shared upon request.

Code availability The Python code used for `DFS-CSI` is available at https://github.com/mestrelan/MDA-CSI.

Declarations

Ethics approval This study was approved by the Ethics Committee of Federal Fluminense University (CAAE 54359221.4.0000.5243).

Consent to participate Informed consent was obtained from all individuals participating in the study.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ahad MAR, Mahbub U, Rahman T (2021) Contactless human activity analysis, vol 200. Springer
2. Koenigstein N (2025) Transformers in action. Simon and Schuster
3. Soto JC, Galdino I, Caballero E, Ferreira V, Muchaluat-Saade D, Albuquerque C (2022) A survey on vital signs monitoring based on Wi-Fi CSI data. Comput Commun 195:99–110
4. Caballero E, Galdino I, Soto JC, Ramos TC, Guerra R, Muchaluat-Saade D, Albuquerque C (2023) Human activity recognition using Wi-Fi CSI. In: International conference on pervasive computing technologies for healthcare, pp 309–321. Springer
5. dos Santos ACN, de Paula K, Vidal MTL, da Silva JMM, de Sousa C, Fernandes LAF, de Castro TB, Bedo M, Kohwalter TC, Bastos CAM, Seixas FL, Fernandes NC, Muchaluat-Saade DC, Ghinea G (2024) A computer vision model to support individuals with disabilities within university campuses. In: 2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom), pp 1–7. https://doi.org/10.1109/HealthCom60970.2024.10880838
6. Soto JCH, Galdino I, Gouveia BG, Caballero E, Ferreira V, Muchaluat-Saade D, Albuquerque C (2022) Wi-fi CSI-based human presence detection using DTW features and machine learning. In: 2022 IEEE Latin-American Conference on Communications (LATINCOM), p 1
7. Gouveia BG, Galdino I, Caballero E, Soto JCH, Ramos TC, Guerra R, Muchaluat-Saade D, Albuquerque CVN (2024) Parameter tuning for accurate heart rate measurement using Wi-Fi signals. In: 2024 International Conference on Computing, Networking and Communications (ICNC), pp 407–411. https://doi.org/10.1109/ICNC59896.2024.10556168
8. Lakshmi S (2026) Transformative role of transformer models in healthcare: understanding AI and large language models-key innovations powering healthcare. In: Transformative Role of Transformer Models in Healthcare. IGI Global Scientific Publishing, pp 123–160
9. Chen Z, Zhang L, Jiang C, Cao Z, Cui W (2019) WiFi CSI based passive human activity recognition using attention based BLSTM. IEEE Trans Mob Comput 18(11):2714–2724
10. Ahad MAR, Inoue S, Lopez G, Hossain T (2024) Human activity and behavior analysis: advances in computer vision and sensors: volume 1, CRC Press
11. Nag A, Hassan MM, Sil R, Karim A (2026) Transforming global healthcare: the role of AI and IoT in modern medical systems. CRC Press
12. Lopes Y, Fernandes NC, de Castro TB, dos Santos Farias V, Noce JD, Marques JP, Muchaluat-Saade DC (2017) Vulnerabilities and threats in smart grid communication networks. In: Security Solutions and Applied Cryptography in Smart Grid Communications, IGI Global Scientific Publishing, pp 1–28
13. Biswas S (2026) Machine learning and deep learning in human activity recognition and fall detection: algorithms, frameworks, and applications for sustainable healthcare. Springer Nature
14. Agrawal A, Gans J, Goldfarb A (eds) (2024) The Economics of Artificial Intelligence: Health Care Challenges (Chicago, IL, 2024; online edn, Chicago Scholarship Online, 19 Sept. 2024). https://

doi.org/10.7208/chicago/9780226833125.001.0001. Accessed 12 Aug 2026
15. Liu KR, Wang B (2019) Wireless AI: wireless sensing, positioning, IoT, and communications. Cambridge University Press
16. Terry J, Heiskala J (2002) OFDM wireless LANs: a theoretical and practical guide. Sams publishing
17. Rothman D (2021) Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more, Packt Publishing Ltd
18. Joseph M (2022) Modern time series forecasting with Python: explore industry-ready time series forecasting using modern machine learning and deep learning. Packt Publishing Ltd
19. Dos Santos AC, Soares P, Galdino I, Soto JC, de Sousa C, Ramos TC, de Albuquerque CV, Guerra R, Fernandes NC, Muchaluat-Saade DC, Ghinea G (2025) A transformer-based methodology for person-independent human activity recognition using Wi-Fi CSI. In: 2025 13th Wireless Days Conference (WD), pp 1–9. https://doi.org/10.1109/WD67713.2025.11302651
20. Galdino I, Soto JCH, Caballero E, Ferreira V, Ramos TC, Albuquerque C, Muchaluat-Saade DC (2023) eHealth CSI: a Wi-Fi CSI dataset of human activities. IEEE Access 11:71003–71012
21. Wang Y, Wu K, Ni LM (2017) Wifall: device-free fall detection by wireless networks. IEEE Trans Mob Comput 16(2):581
22. Chen C, Zhou G, Lin Y (2023) Cross-domain WiFi sensing with channel state information: a survey. ACM Comput Surv 55(11):1–37
23. Milan AA, Fernandes NC, Medeiros DSV (2022) A Monte Carlo approach for antenna blocking probability estimation in mobile networks. In: 2022 25th Conference on Innovation in Clouds, Internet and Networks (ICIN), pp 146–150. https://doi.org/10.1109/ICIN53892.2022.9758117
24. Sai S, Sharma D, Peelam MS, Chamola V, Guizani M, Niyato D (2026) Machine learning techniques for Wi-Fi CSI-based recognition and sensing: a comprehensive review. IEEE Internet Things J
25. Xiao C, Han D, Ma Y, Qin Z (2019) CsiGAN: robust channel state information-based activity recognition with GANs. IEEE Internet Things J 6(6):10191–10204
26. Xiao J, Wu K, Yi Y, Wang L, Ni LM (2013) Pilot: passive device-free indoor localization using channel state information. In: 2013 IEEE 33rd international conference on distributed computing systems, pp 236–245. https://doi.org/10.1109/ICDCS.2013.49
27. Wang D, Yang J, Cui W, Xie L, Sun S (2021) Multimodal CSI-based human activity recognition using GANs. IEEE Internet Things J 8(24):17345–17355
28. Li S, Ge Y, Shentu M, Zhu S, Imran M, Abbasi Q, Cooper J (2021) Human activity recognition based on collaboration of vision and WiFi signals. In: 2021 International Conference on UK-China Emerging Technologies (UCET), pp 204–208
29. Boudlal H, Serrhini M, Tahiri A (2023) Human activity monitoring system with commodity WiFi infrastructure using channel state information. Indones J Electr Eng Comput Sci 31(2):763
30. Boudlal H, Serrhini M, Tahiri A. A monitoring system for elderly people using WiFi sensing with channel state information. Int J Interact Mobile Technol 17(12)
31. Boudlal H, Serrhini M, Tahiri A (2023) Design and deployment of a practical wireless sensing system for HAR with WiFi CSI in the 5GHz band. In: 2023 International Conference on Networking, Sensing and Control (ICNSC) 1:1–6. IEEE
32. Boudlal H, Serrhini M, Tahiri A (2023) Leveraging WiFi channel state information for efficient and reliable device-free human activity recognition. In: 2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), pp 1–8. IEEE
33. Boudlal H, Serrhini M, Tahiri A (2024) CSHA-CSI: towards contactless sensing of human activities utilizing WiFi channel state information. In: 2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI), pp 1–9. IEEE
34. Boudlal H, Serrhini M, Tahiri A (2024) Exploiting channel state information of WiFi signal for human activity detection: an experimental study. Bulletin of Electrical Engineering and Informatics 13(2):1031–1041
35. Boudlal H, Serrhini M, Tahiri A (2024) A novel approach for simultaneous human activity recognition and pose estimation via skeleton-based leveraging WiFi CSI with yolov8 and mediapipe frameworks. SIViP 18(4):3673–3689
36. Boudlal H, Serrhini M, Tahiri A (2025) Towards a low-cost and privacy-preserving indoor activity recognition system using WiFi channel state information. Multimedia Tools and Applications 84(29):35761–35792
37. Ahmed S, Nielsen IE, Tripathi A, Siddiqui S, Ramachandran RP, Rasool G (2023) Transformers in time-series analysis: a tutorial. Circuits Systems Signal Process 42(12):7433–7466
38. Duan Y, Guo J, Yang C (2025) Learning precoding in multi-user multi-antenna systems: transformer or graph transformer? IEEE Trans Wireless Commun 25:6284–6300
39. Huang Y, Zhang Z, Che J, Yang Z, Yang Q, Wong K-K (2023) Self-attention reinforcement learning for multi-beam combining in mmWave 3D-MIMO systems. SCIENCE CHINA Inf Sci 66(6)
40. Boudlal H, Serrhini M, Tahiri A (2026) An optimized GRU-RNN framework for real-time human interaction recognition using WiFi channel state information with integrated GUI for smart environments. Multimedia Tools and Applications 85(4):321
41. Zhu A, Tang Z, Wang Z, Zhou Y, Chen S, Hu F, Li Y (2022) Wi-ATCN: attentional temporal convolutional network for human action prediction using WiFi channel state information. IEEE Journal of Selected Topics in Signal Processing 16(4):804–816
42. Lin C-Y, Lin C-Y, Liu Y-T, Chen Y-W, Shih TK (2024) WiFi-TCN: temporal convolution for human interaction recognition based on WiFi signal. IEEE Access 12:126970–126982
43. Kang H, Kim D, Toh K-A (2025) Human activity recognition through augmented WiFi CSI signals by lightweight attention-GRU. Sensors 25(5):1547
44. Ye X (2026) Mathematical foundations of deep learning: theory and algorithms. CRC Press
45. Zhang H, Li H Transformer-based EEG decoding: a survey arXiv:2507.02320 arXiv preprint
46. Taneja H, Kim J, Xu JJ, Van Schaik S, Genkin D, Yarom Y (2023) Hot pixels: frequency, power, and temperature attacks on GPUs and arm SoCs. In: 32nd USENIX Security Symposium (USENIX Security 23), pp 6275–6292
47. Schwartz R, Dodge J, Smith NA, Etzioni O (2020) Green AI. Commun ACM 63(12):54–63
48. Wang X, Zhang LL, Wang Y, Yang M (2022) Towards efficient vision transformer inference: a first study of transformers on mobile devices. In: Proceedings of the 23rd annual international workshop on mobile computing systems and applications, pp 1–7
49. Maurício J, Domingues I, Bernardino J (2023) Comparing vision transformers and convolutional neural networks for image classification: a literature review. Appl Sci 13(9):5521
50. Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L: How to train your vit? Data, augmentation, and regularization in vision transformers. arXiv:2106.10270 arXiv preprint
51. Dehghani M, Djolonga J, Mustafa B, Padlewski P, Heek J, Gilmer J, Steiner AP, Caron M, Geirhos R, Alabdulmohsin I et al (2023) Scaling vision transformers to 22 billion parameters. In: International conference on machine learning, pp 7480–7512. PMLR
52. Forbes G, Massie S, Craw S (2020) WiFi-based human activity recognition using Raspberry Pi. In: 2020 IEEE 32nd International

Conference on Tools with Artificial Intelligence (ICTAI), pp 722–730. IEEE
53. Bekaroo G, Santokhee A (2016) Power consumption of the Raspberry Pi: a comparative analysis. In: 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), pp 361–366. IEEE
54. Vujović V, Maksimović M (2014) Raspberry pi as a wireless sensor node: Performances and constraints. In: 2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO) 247(2014):1013–1018. IEEE
55. Hall J (2014) Rethinking the system monitor. Linux Journal 2014(247):4
56. Tabbakh A, Al Amin L, Islam M, Mahmud GI, Chowdhury IK, Mukta MSH (2024) Towards sustainable AI: a comprehensive framework for green AI. Discover Sustainability 5(1):408
57. Gopi E (2016) Gopi, E. Gopi, Digital signal processing for wireless communication using Matlab, Springer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.