Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079691

ARTICLE

WiFi-Based Indoor Intrusion Detection via Two-Level Gait Feature Fusion Model

Lijun Cui[1], Yongjie Niu[2], Yuxiang Sun[1], Xiaokang Gu[1], Jing Guo[1] and Pengfei Xu[1,*]

[1]College of Computer Science, Northwest University, Xi'an, China
[2]College of Mathematics and Computer Science, Yan'an University, No. 1 North Gongxue Road, Yan'an, China
*Corresponding Author: Pengfei Xu. Email: pfxu@nwu.edu.cn

Received: 26 January 2026; Accepted: 18 March 2026; Published: 08 May 2026

ABSTRACT: Indoor intrusion detection is essential for various applications, including security systems and smart homes. Recently, WiFi-based detection has gained popularity due to its low cost and non-invasive nature. Current Channel State Information (CSI) based frameworks primarily use deep learning to extract gait signatures; however, their performance depends heavily on extensive labeled datasets. These methods struggle to differentiate between unlabeled and labeled data that exhibit similar features. To address this challenge, we propose a novel Two-level Feature Fusion model for Indoor Intrusion Detection (TFF-IID) utilizing commercial WiFi CSI. The model adopts a two-level structure to learn rich feature representations and introduces a Transformer with multi-head self-attention alongside a multi-scale convolution module to process sensor data. Additionally, it incorporates a self-supervised learning module to capture general normality patterns. Based on this architecture, TFF-IID achieves accurate intrusion detection using only CSI. Empirical evaluations on a private gait dataset demonstrate that TFF-IID achieves an intrusion detection accuracy of 73.5% and an F1-score of 76.2% across 10 unauthorized subjects. Moreover, cross-scenario assessments verify that the proposed model maintains high efficiency and robustness in environments characterized by diverse spatial layouts and multipath complexities. Furthermore, TFF-IID outperforms the best baseline by 19.7% and 25.7% in accuracy and F1-score, respectively.

KEYWORDS: Intrusion detection; transformer; feature fusion; WIFI signal; CSI; multi-scale convolution module

1 Introduction

Indoor intrusion detection is defined in this study as the continuous monitoring and dynamic analysis of indoor environments to identify and classify unauthorized behaviors or anomalous activities using integrated sensor networks and intelligent algorithms. This technology is essential for enhancing indoor security. In recent years, both industry and academia have proposed various solutions, which can be broadly categorized into non-Radio Frequency (RF) and RF-based methods. Non-RF approaches typically include camera-based systems [1], ultrasonic sensing [2] and infrared-based methods [3,4]. However, these techniques can offer high accuracy; they often face significant challenges, such as privacy concerns, susceptibility to environmental interference, and the requirement for specialized, high-cost equipment.

Over the years, WiFi-based indoor intrusion detection systems have been proposed as WiFi is widely utilized for its extensive coverage, flexibility, ease of deployment, and low cost. More importantly, the wavelength of WiFi signals is generally larger than the surface roughness of most objects, which inherently helps protect user privacy. Taking advantage of these advantages, the Channel State Information (CSI) of

WiFi signals has enabled various applications, including gesture recognition [5,6], indoor localization [7], respiratory monitoring [8,9], gait recognition [7], and identity recognition [10].

Despite achieving relatively high precision, these methods still face several challenges, primarily their heavy reliance on high-dimensional features and extensively labeled data. For instance, Zhang et al. [11] proposed WiFi-ID, which captures CSI-based gait data for human identification. More recently, DCS-Gait [12] has achieved robust identification across diverse environments by extracting invariant gait features through cross-attention metrics and high-quality pseudo-labeling. Nevertheless, effectively identifying new users in open-set scenarios remains a significant challenge when labeled information is unavailable. The WiDIGR [13] system can recognize individuals through gait analysis, achieving an average precision ranging from 78.28% for six subjects to 92.83% for three subjects. However, it remains unable to effectively recognize new users who are not part of the training set. Therefore, accurately identifying new users by leveraging unique features without relying on labeled data remains a significant challenge. Most current intrusion detection systems rely on closed-set datasets, which limit their ability to identify new users with similar physical characteristics but no labeled information. This limitation poses security risks; for instance, our experiments show that two volunteers with similar postures can induce nearly identical CSI changes (as shown in Fig. 1), potentially leading to detection errors. Consequently, identifying new users and distinguishing between similar individuals in open-set scenarios are critical tasks for enhancing the reliability of security monitoring systems

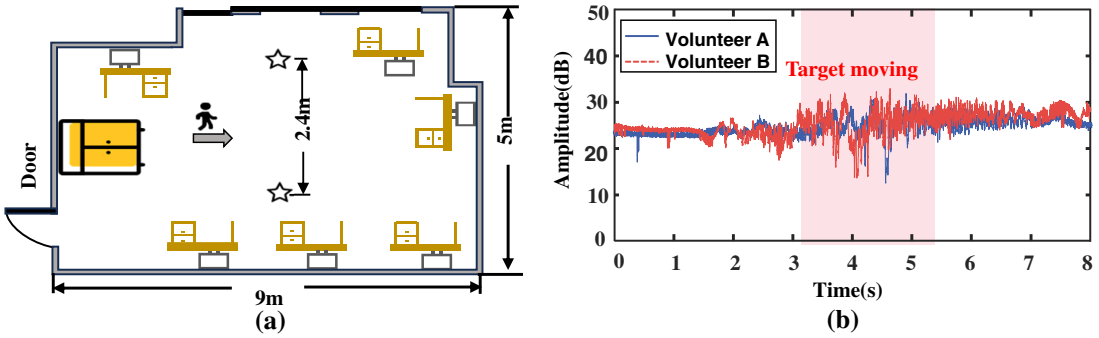


Figure 1: Comparison experiment between two similar volunteers walking along a specified route. (**a**) Experimental layout and the predefined walking trajectory; (**b**) CSI amplitude variations for two volunteers, demonstrating nearly identical signal patterns that pose potential security risks in open-set scenarios.

To address these challenges, we propose a novel Two-Level Feature Fusion based Indoor Intrusion Detection (TFF-IID) model leveraging commercial off-the-shelf (COTS) WiFi devices. The framework employs a two-level fusion structure designed to enhance generalization by mitigating interference from diverse human walking patterns. Specifically, TFF-IID utilizes a Transformer with multi-head self-attention and a multi-scale convolution (CNN) module to process raw sensor data and capture essential gait features. Building upon this, a self-supervised learning paradigm is integrated to characterize latent features, enabling the model to establish robust normality patterns for improved anomaly detection. Furthermore, a feature reconstruction module is implemented to learn comprehensive feature representations. The architecture concludes with a feature fusion module designed for accurate intrusion detection. This entire TFF-IID model is built on a convolutional encoder framework, facilitating efficient end-to-end training.

The main contributions of this paper are summarized as follows:

(1) Introduction of Self-Supervised Learning for CSI Sensing: We integrate a self-supervised learning paradigm into CSI-based intrusion detection, utilizing an encoder-decoder network to extract latent features and reduce the heavy reliance on labeled data.

(2) Novel End-to-End Detection Framework: We propose an end-to-end intrusion detection model that merges a feature reconstruction module with conventional label-based features to effectively distinguish between known and unknown identities.

(3) Comprehensive Dataset and Performance Validation: We establish a large-scale indoor gait dataset featuring WiFi CSI data for eight distinct actions performed by 20 individuals. Our algorithm achieves a detection accuracy of 73.5% in the presence of 10 unknown identities, significantly outperforming existing CSI-based methods.

2 Relate Work

In contemporary smart environments, indoor intrusion detection focuses on identity verification to authenticate legitimate users. This process utilizes diverse technologies, which are broadly classified into non-Radio Frequency (RF) and RF-based methods, encompassing visual, acoustic, and wireless signal-based recognition systems.

Traditional sensing technologies, such as visual [14], acoustic [15,16], and infrared-based systems [17,18], extensively employed in monitoring applications. However, visual and acoustic-based methods often raise significant privacy concerns, which can limit their suitability for sensitive indoor environments. While infrared systems provide better privacy and accuracy, they are frequently constrained by narrow detection angles and limited range, resulting in "blind spots" that hinder comprehensive coverage. In contrast, WiFi-based methods offer enhanced privacy protection and high accuracy, making them a robust alternative for indoor intrusion detection.

WiFi-based indoor intrusion detection primarily employs two categories of techniques: Received Signal Strength Indication (RSSI) [19,20] and Channel State Information (CSI) [21,22]. RSSI-based detection is widely adopted because the data can be easily collected using standard wireless devices without requiring specialized hardware [23,24]. However, RSSI is inherently a coarse-grained metric and is highly sensitive to multipath interference, which leads to significant signal fluctuations even in static environments [25]. These factors degrade the overall performance of RSSI-based systems, reducing their accuracy and making it difficult to maintain reliable detection.

In contrast, CSI-based detection methods provide fine-grained and high-fidelity measurements. Leveraging high frequency-domain and temporal resolution, CSI can effectively distinguish individual multipath components and exhibits superior sensitivity to subtle human movements compared to RSSI [26]. Consequently, it has become a cornerstone for diverse sensing applications. In recent years, extensive research has focused on utilizing commercial off-the-shelf (COTS) WiFi devices for intrusion detection tasks. For instance, Tian et al. [27] utilized CSI to achieve adaptive, device-free human intrusion detection in dynamic indoor environments. WIDD [28] integrates the Angle of Arrival (AOA) with multipath signal amplitudes to perform Jarque-Bera (JB) testing for identifying the direction of intrusion. Additionally, FreeSense [29] leverages the phase difference between amplitude waveforms across multiple receiving antennas to accurately detect human motion.

Recent advancements in deep learning have effectively addressed noise-related challenges in traditional CSI extraction by learning latent features and significantly improving detection accuracy. For instance, WiWho [30] extracts gait signatures for identity authentication; however, it struggles to differentiate

unknown identities when relying on single gait features, particularly when new users closely resemble known individuals. To capture complex temporal dependencies, Mohd Noor et al. [31] designed for activity recognition. Similarly, Chen et al. [32] introduced an Attention-Based LSTM (ABLSTM) method to enhance the precision of WiFi-based activity sensing. Furthermore, Jobanputra et al. [33] combined LSTM with Recurrent Neural Networks (RNNs) to extract discriminative features for various human interactions.

3 System Overview

The proposed intrusion detection framework, which integrates supervised and self-supervised learning strategies, is illustrated in Fig. 2. The model architecture consists of four primary modules.

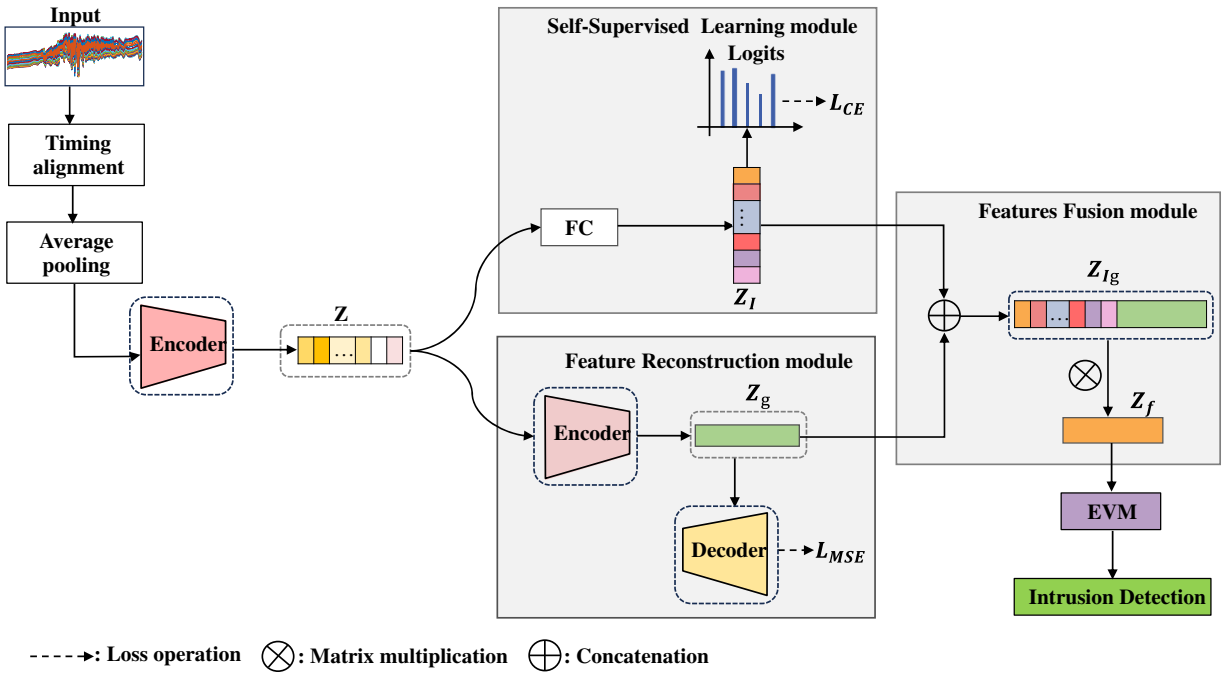


Figure 2: The structure of the proposed TFF-IID. It consists of four components: Encoder module, self-supervised learning module, Feature Reconstruction module and Feature Fusion module.

(1) Encoder module: The preprocessed CSI data is encoded via a hierarchical self-attention mechanism within a Transformer network to extract identity-specific features, denoted as Z_I. By capturing complex temporal dependencies and latent patterns, the model effectively learns representations closely tied to individual identity, ensuring high feature richness. A linear layer is subsequently employed to recognize known identities based on these representations.

(2) Self-Supervised Learning Module: This module characterizes user patterns through a self-supervised approach, eliminating the requirement for manually labeled data.

(3) Feature Reconstruction Module: An autoencoder architecture is utilized to reconstruct the features Z, focusing on identity-centric representations Z_g. By filtering redundant information through the encoding-decoding process, the autoencoder retains essential identity-related features, thereby enhancing the model's overall robustness.

(4) Feature Fusion and Intrusion Detection: The extracted features Z are fused with the reconstructed features Z_I to produce a unified feature vector Z_{Ig}. This vector is then fed into an Extreme Value Machine (EVM)for the detection of unknown identities. Leveraging its rapid learning and strong generalization capabilities,

the EVM enables the system to effectively handle unknown intruders and produce final recognition results, improving the model's applicability.

3.1 CSI Data Preprocessing

To ensure the reliability of features extracted in complex indoor environments, preprocessing the CSI data is essential for mitigating noise interference. This stage involves implementing denoising techniques, such as outlier removal and wavelet denoising, to significantly suppress environmental and hardware-induced noise. These steps result in refined data, facilitating more accurate feature extraction and enhancing overall model performance.

3.1.1 Outlier Handling

Hardware limitations and environmental noise introduce outliers into CSI data, which can compromise feature extraction and degrade intrusion detection accuracy. Fig. 3 illustrates the collected CSI data corresponding to a volunteer's gait. For the sake of clarity, only one of the 30 subcarriers is displayed. As observed, the raw CSI data extracted from commercial WiFi devices exhibits significant outliers caused by hardware constraints and complex multipath interference. To mitigate these effects and eliminate outliers, this study employs the Hampel outlier removal method [34]. Under this approach, data points within a specified threshold are classified as normal, while any data falling outside this range is identified as abnormal and replaced with the local mean. The processed data is shown in Fig. 3b. A comparison between Fig. 3a,b confirms that the outliers have been effectively removed. Such processing significantly enhances the accuracy and reliability of the data for subsequent analysis.

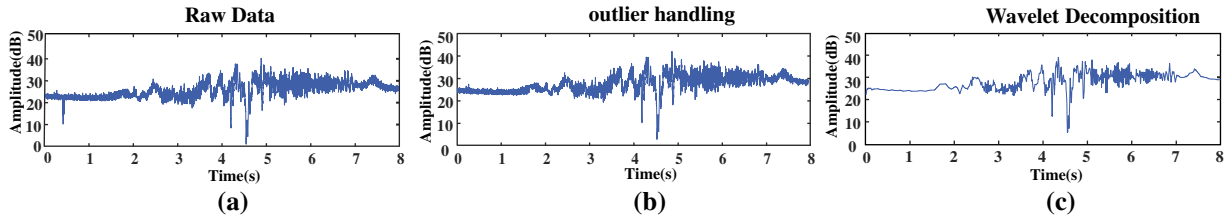


Figure 3: Denoising and decomposition of the CSI signal. (**a**) Original CSI stream of one subcarrier. (**b**) After multipath and outlier removal. (**c**) Effect the CSI signal after subcarrier selection.

3.1.2 Wavelet Decomposition

Even after removing outliers, additional denoising of CSI data is necessary because environmental factors—such as air pressure, electromagnetic interference, and temperature variations—can significantly affect signal propagation and amplitude. Traditional low-pass filters are often inadequate for capturing the transient signal changes associated with intrusion behaviors. Consequently, wavelet analysis is employed to decompose signals into coefficients for multi-scale analysis, which effectively captures short-term variations and local features to enhance signal clarity and recognition accuracy. To suppress environmental noise and hardware-induced impulse interference, we implement a Daubechies (db3) wavelet decomposition. As illustrated in Fig. 3c, this approach preserves the transient features of human gait more effectively than conventional low-pass filtering. Furthermore, it maintains the original waveform characteristics better than the Hampel method alone, ensuring high-fidelity data for subsequent feature learning.

3.2 *Transformer Module*

The inherent long-sequence nature of CSI data poses significant challenges for traditional neural networks in capturing long-term dependencies. To address this, a Transformer-based architecture is employed. Its multi-head self-attention mechanism facilitates the modeling of intricate temporal correlations within gait-induced signal fluctuations. As illustrated in Fig. 4, the Transformer framework consists of three primary components.

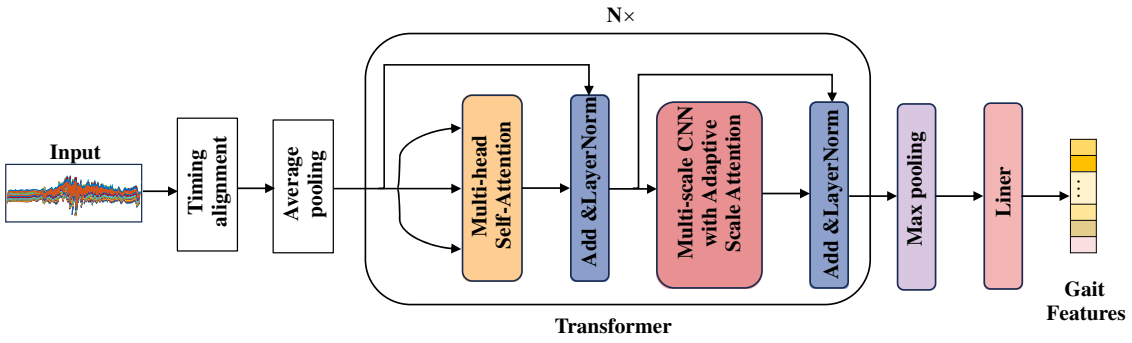


Figure 4: Transformer network architecture.

Input Layer: This layer receives the preprocessed CSI data. To facilitate batch processing, we perform time alignment to ensure all records are of a uniform size. To optimize computational efficiency and reduce memory usage, average pooling is applied to adjacent data points, thereby reducing the temporal dimension.

Multi-Layer Convolution Transformer Layer: This layer is responsible for extracting discriminative features from the input data. Structured as a tower, it consists of stacked multi-head self-attention modules and multi-scale convolution blocks to capture features across different scales [35].

Linear Layer: The final layer computes the probabilities of known identities using a Softmax operation.

Data representation and alignment: CSI data is represented as a set of records $\{r_1, r_2, ..., r_n\}$, where each record is a two-dimensional matrix. In this matrix, each cell contains a real number representing the state value at time t for a specific channel. A primary challenge in CSI-based gait recognition is the variation in walking patterns among individuals. To address this and facilitate batch processing, we perform time alignment on the preprocessed data. Specifically, a fixed 8-s window is used for data analysis. For data exceeding this duration, average pooling is applied to compress the time series by averaging values within each window. This method effectively reduces data volume while preserving key gait information, as defined in Eq. (1).

$$Y_i = \frac{1}{w}\sum_{j=(i-1)\times w+1}^{i\times w} r_j \tag{1}$$

$$r(t) = r(t_i) + \frac{r(t_{i+1}) - r(t_i)}{t_{i+1} - t_i} \times (t_{i+1} - t_i) \tag{2}$$

where $r(t)$ represents the estimated value at time, $r(t_i)$ and $r(t_{i+1})$ denote the known CSI values at time t_i and t_{i+1}, respectively.

After the temporal alignment of the CSI data, the processed sequences are fed into the feature extraction module. This module is inspired by the Multi-scale Convolution Augmented Transformer (MCAT) layer proposed by Lin et al. in [36]. The extraction process is formulated as shown in Eq. (3).

$$f_G = f_{encoder}(r) \tag{3}$$

where f_G represents the extracted gait features, $f_{encoder}(g)$ denotes the MCAT module, which is configured with six layers. The variable $r$represents the input CSI sequence. Following the initial feature extraction, a linear layer F_L is utilized to further extract identity-related features f_I associated with specific classes. The process is defined as $f_I = F_L(f_G)$. where f_I represents the extracted identity features, F_L represents the extracted identity features. This layer maps the high-dimensional gait feature space into a lower-dimensional space corresponding to the number of identity classes for classification. The loss function $Loss_1$ employed during the training process of the Transformer network is formulated as shown in Eq. (4).

$$Loss_1 = -\sum_i^N y_i log(f(\theta; r)) \tag{4}$$

where, y_i represents the ground-truth label, and $f(g; r)$ denotes the predicted probability distribution after the CSI sequence r is processed by the linear layer. The variable r specifically denotes the input CSI sequence.

By leveraging the Transformer network's capacity to model long-range dependencies and capture intricate temporal correlations within gait signals, the proposed model effectively identifies known identities. This capability significantly enhances the overall reliability and performance of the indoor intrusion detection system.

3.3 Feature Reconstruction Module

The feature reconstruction module employs an AutoEncoder architecture, comprising an encoder and a decoder structured with multi-layer feed-forward neural networks. The corresponding training process is illustrated in Fig. 5. Specifically, the encoder transforms the input feature Z into a latent representation Z_I, which is formulated as $f_Z = F_e(f_G)$. Here, F_e denotes the mapping function of the encoder network, which comprises three feed-forward neural network layers. The term f_Z represents the transformed latent feature representation. Correspondingly, the decoder maps these latent features back to the original data space to generate reconstructed data, which is formulated as $f_R = F_d(f_G)$. F_d denotes the mapping function of the decoder network, which comprises three feed-forward neural layers. The term f_R represents the reconstructed feature vector.

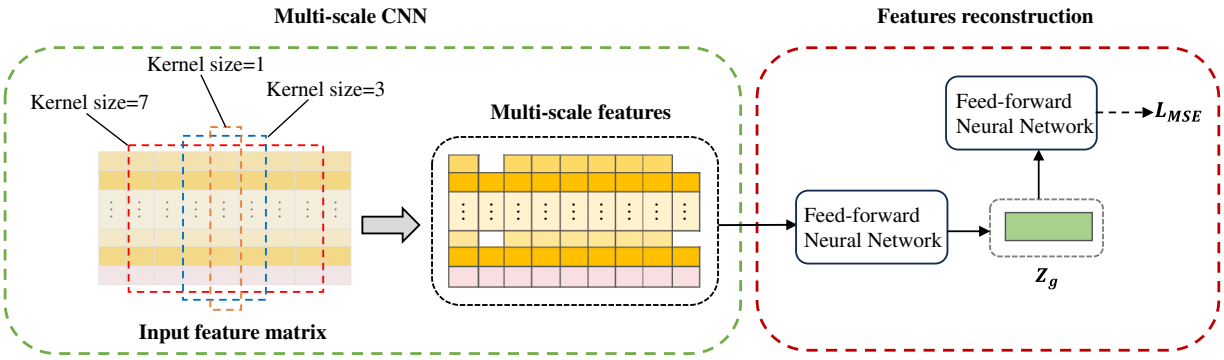


Figure 5: Encoder and decoder configurations.

The primary objective of the AutoEncoder is to minimize the reconstruction error between the original gait features f_G and the reconstructed features f_R. This process ensures that the latent space effectively captures the most salient characteristics of the input signal while filtering out non-essential noise.

To ensure that the learned latent representation exhibits desirable characteristics, such as distribution consistency and robustness, this study utilizes Kullback-Leibler (KL) ddivergence to quantify the disparity between the posterior distribution $q_\theta\left(f_Z|f_R\right)$ and the prior distribution $p_\phi\left(f_Z|f_G\right)$. The KL divergence can be defined as Eq. (5).

$$L_{kl} = D_{kl}\left(q_\phi(f_Z \mid f_R) \parallel p_\theta(f_Z \mid f_G)\right) \tag{5}$$

where $D_{kl}(\cdot \parallel \cdot)$ denotes the KL divergence operator. The training objective of the AutoEncoder is to jointly minimize the reconstruction loss and the KL divergence, as formulated in Eq. (6).

$$L = L_{recon} + \beta L_{kl} \tag{6}$$

where β serves as a weighting factor to balance the reconstruction loss and the KL divergence. In our implementation, β is set to 0.5 to ensure an optimal trade-off between reconstruction fidelity and latent space regularization. By optimizing this joint training objective, the AutoEncoder learns robust latent representations that are essential for subsequent tasks, such as unknown identity detection in open-set scenarios.

To quantify the discrepancy between the original features and the reconstructed data, the Mean Squared Error (MSE) is employed as the reconstruction loss function. It is formulated as Eq. (7).

$$L_{recon} = \frac{1}{N}\sum_{i}^{N}\left\|f_{G_i} - f_{R_i}\right\|^2 \tag{7}$$

where N denotes the total number of training samples. The variables f_{G_i} and f_{R_i} represent the ith original gait feature vector and its corresponding reconstructed counterpart, respectively.

3.4 Feature Fusion Module

While Transformer networks effectively recognize known identities using labeled data, they inherently struggle with unknown identities in open-set scenarios due to the absence of prior labels. To overcome this limitation, we leverage the Autoencoder's ability to capture latent representations f_Z, which provide essential cues for identifying unseen identities.

To bridge this gap, we concatenate categorical identity features f_G with latent representations f_Z to create a more comprehensive feature set. This fused representation incorporates supervised knowledge while retaining the underlying characteristics of unknown individuals, thereby enabling effective discrimination between authorized users and intruders. As illustrated in Fig. 6, the fusion process is formulated $f_T = C\left(f_I, f_Z\right)$. denotes the concatenation operation. To further refine these fused features and ensure they exhibit distinct characteristics for intrusion detection, we implement a feature weighting mechanism. Specifically, the concatenated features are passed through a weighting layer to generate an attention-based weight vector A: $A = Sigmoid\left(C\left(f_I, f_Z\right)\right)$. Subsequently, these weights are applied to the fused features via element-wise multiplication (dot product) to emphasize salient information. Finally, a fully connected (FC) layer is employed to enhance the non-linear interaction between different feature components. The calculation of the final refined feature vector f_T is expressed in Eq. (8).

$$f_T = FC\left(A \odot C\left(f_I, f_Z\right)\right) \tag{8}$$

where $\odot$ denotes the Hadamard product. This refined feature vector f_T serves as a high-fidelity representation that significantly improves the system's ability to distinguish between legitimate users and illegitimate intruders.

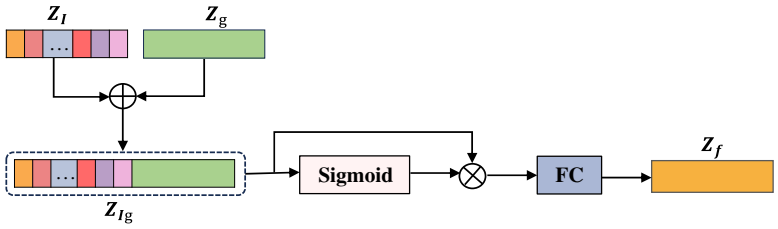


Figure 6: Features Fusion module.

3.5 Open-Set Recognition and Classification

After the feature fusion stage, this study implements an open-set recognition framework that builds upon the principles of the OpenMax algorithm integrated with the Extreme Value Machine (EVM). This hybrid approach is specifically designed to handle the final identity recognition and intrusion detection tasks. Importantly, the EVM operates directly on the refined, fused feature vectors (f_T) generated in the previous module. To facilitate the recognition of authorized user classes, the weibull distribution from Extreme Value Theory (EVT) is introduced.

$$\Psi(f_T, \mu_i) = 1 - \exp\left(-\left(\frac{d(f_T, \mu_i)}{\lambda_i}\right)^{\kappa_i}\right) \tag{9}$$

where $d(f_T, \mu_i)$ denotes the Cosine distance between the input feature f_T and the class mean μ_i. The Weibull parameters, κ_i and λ_i are estimated by fitting the distribution to the 20 most proximate tail samples in the feature space.

$$Identity = \begin{cases} \text{Authorized User } i, & \text{if } \max_i \Psi(f_T, \mu_i) \geq \tau \\ \text{Unknown Intruder}, & \text{otherwise} \end{cases} \tag{10}$$

where the rejection threshold τ is set to 0.65 by grid search in a validation set. The key configurations of the EVM module are summarized in Table 1.

Table 1: Detailed configuration and calibration parameters for the EVM module.

Hyperparameter/Protocol	Configuration/Value
Distance Metric	Cosine Similarity
Tail Size (κ)	20
Rejection Threshold (τ)	0.65
Calibration Protocol	Grid search on a small validation set
Optimization Objective	Balanced True Positive Rate (TPR) and FPR

4 Experimental Evaluation

4.1 Experimental Setup

Hardware Configuration: Our experimental platform consists of a Netgear router acting as the Access Point (AP) and a Lenovo laptop equipped with an Intel 5300 Network Interface Card (NIC) as the receiver. The receiver is configured with three omni directional antennas, a setup that ensures versatile deployment across various indoor environments. To capture fine-grained wireless channel information, we utilize the CSI Tool to record CSI measurements for each received packet. The sampling rate is set to 1000 Hz, which

provides sufficient temporal resolution to capture the nuances of human gait. We specifically operate in the 5 GHz frequency band; its shorter wavelength compared to the 2.4 GHz band allows the signal to capture more detailed variations induced by body movements.

Model Parameters: For the proposed Transformer-based architecture, the average pooling size is set to 4 to reduce computational complexity while preserving essential features. The model comprises a stack of 5 encoder modules. The embedding dimensionality is set to 90, and the multi-head attention mechanism is configured with 9 heads. To prevent overfitting, a dropout rate of 0.1 is applied. For the multi-scale convolution blocks, we employ kernel sizes of 1, 3, 5 with a hidden dimension of 360.

4.2 Experimental Environments

The experiments were conducted in a controlled laboratory environment, as illustrated in Fig. 7a. Fig. 7b details the layout and the walking route for the volunteers. To simulate a typical domestic wireless device deployment, the transmitter (Tx) and receiver (Rx) were positioned 2.4 m above the ground with a separation distance of 0.8 m. The devices were precisely aligned to ensure a stable horizontal Line-of-Sight (LOS) link.

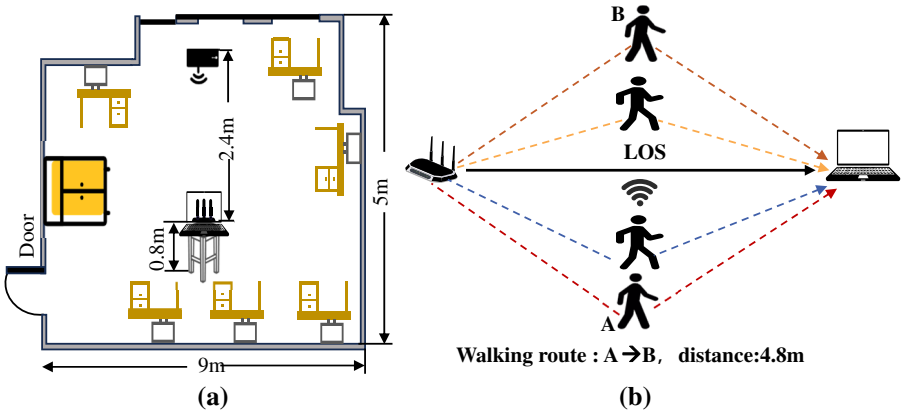


Figure 7: Experiments deployment. (**a**) Schematic layout of the indoor testbed; (**b**) Illustration of the walking trajectory and the Line-of-Sight (LOS) path between the transmitter and receiver.

4.3 Datasets and Evaluation Metrics

Evaluating gait patterns in real-world scenarios is challenging due to the inherent physiological differences among individuals, such as age, gender, height, and weight. To evaluate the proposed TFF-IID model, we curated a comprehensive laboratory dataset comprising CSI samples from 20 volunteers performing eight distinct actions. The subjects represent a diverse demographic: ages ranging from 22 to 28, heights from 150 to 177 cm, and weights from 45 to 80 kg. This diversity ensures the generalizability of the dataset and the robustness of our experimental findings.

We recorded 300 samples for each of the eight actions per volunteer. Participants were instructed to walk along a predefined 4.8-m path under LOS conditions, moving from a designated starting point to an endpoint, as shown in Fig. 7b. Data collection was synchronized with the volunteer's movement: recording commenced at the starting point and concluded at the endpoint, with each one-way traversal producing a single CSI data sample. To simulate realistic intrusion scenarios, we defined eight typical indoor entry behavior patterns, as detailed in Table 2.

Table 2: Enter actions.

Num.	Actions	Samples
01	Empty-handed	300
02	Water bottle	300
03	Suitcase	300
04	Backpack	300
05	Bot. & Suit.	300
06	Back. & Bot.	300
07	Back. & Suit.	300
08	All three	300

Data preprocessing was conducted using MATLAB 2021b. The model training and evaluation were performed on a Linux-based system equipped with 32 GB of RAM and an NVIDIA GeForce RTX 2080 Ti GPU to facilitate efficient deep learning computations. The software ecosystem was built upon Python, leveraging essential libraries such as PyTorch-GPU for neural network implementation, NumPy for numerical analysis, Scikit-learn for metric evaluation, and Tqdm for progress monitoring.

To rigorously evaluate the performance of the TFF-IID model in intrusion detection, we employ four standard classification metrics. These metrics are derived from the elements of the confusion matrix: True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*).

Accuracy: The ratio of correctly predicted samples (both positive and negative) to the total number of samples.

Precision: Also known as positive predictive value, it measures the proportion of correctly identified positive instances among all samples predicted as positive.

Recall: Also known as sensitivity, it quantifies the ability of the model to identify all actual positive samples. In our context, this signifies the probability of correctly detecting an intrusion when it occurs.

F1-Score: The harmonic mean of Precision and Recall, providing a balanced, comprehensive assessment of the classifier's performance, particularly when dealing with uneven class distributions.

4.4 Overall Perform Evaluate

To assess the open-set recognition capabilities of the proposed TFF-IID model, we partitioned the dataset to simulate real-world intrusion scenarios. Initially, data from five randomly selected volunteers were designated as known identities for the training phase. Subsequently, data from four other individuals were introduced as unknown identities (unseen intruders) during the testing phase. To further evaluate the system's robustness, we incrementally increased the number of unknown identities in the test set to observe performance stability across varying levels of task complexity. The results of this open-set evaluation are illustrated in Fig. 8.

As illustrated in Fig. 8a, with a single unknown identity in the dataset, the proposed TFF-IID model achieves a 100% detection rate for the unknown intruder and a 92.86% identification accuracy for known authorized users. This high level of precision underscores the model's reliability for identity verification in practical deployments. As the number of unknown identities increases from two to four, depicted in Fig. 8b–d, the detection rate for unknown individuals experiences a marginal decline of 2.3%, 2.5%, and 5%, respectively. Despite this, the system maintains a robust detection rate exceeding 95%, while the identification

accuracy for known identities remains consistently stable. These results demonstrate the model's high stability and effectiveness in real-world scenarios. The slight decrease in unknown detection performance is attributed to the increased feature similarity within the expanded latent space, yet the overall performance remains well within the requirements for secure intrusion detection.

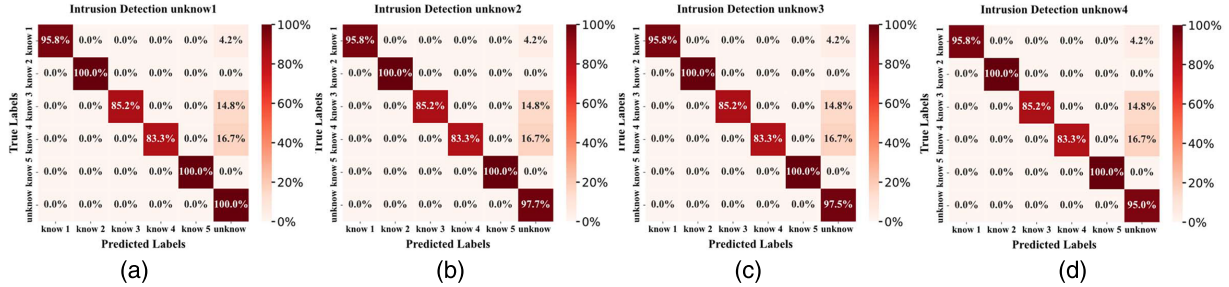


Figure 8: Confusion matrices evaluating the open-set intrusion detection performance of the proposed TFF-IID model under varying task complexities. (**a**–**d**) illustrate the recognition results when incrementally introducing one, two, three, and four unknown identities (unseen intruders) into the test set, respectively.

4.5 Sensitivity Analysis of Parameters

To evaluate the effectiveness of the window setting, we conducted a comparison using different window lengths (T = 4, 8, 12 s), with the results presented in Table 3. The results indicate that the performance of the 4 s window decreased, with accuracy dropping to 67.8%. This reduction is primarily due to the inability of shorter windows to capture 4–6 complete gait cycles, which are essential for extracting stable identity features. Additionally, as shown in Table 3, although the 12 s window slightly improved accuracy from 73.5% to 76.3%, the processing time per sample escalated from 18.5 to 32.3 ms. This 76.3% increase in latency significantly hindered the system's real-time response, which is a critical requirement for effective intrusion detection. Therefore, the 8s window is the optimal choice, effectively balancing detection precision with the low-latency demands of the monitoring system.

Table 3: Sensitivity analysis of window length on detection performance and computational cost.

Window Length (s)	Accuracy (Acc.)	Comp. Cost (ms/Sample)
4 s	67.8%	12.2
8 s (Original)	73.5%	18.5
12 s	76.3%	32.3

4.6 Impact of Training Personal Selection

To evaluate the generalization capability of the TFF-IID model, we employed a randomized partitioning strategy to construct the training and testing sets. From the 20 volunteers (indexed 1–20), five participants (IDs: 8, 14, 15, 16, and 17) were fixed as the test set. The remaining volunteers were randomly organized into four training groups, each comprising five individuals: Group 1 [1, 2, 3, 4, 5], Group 2 [1, 3, 5, 7, 9], Group 3 [2, 4, 6, 7, 10], and Group 4 [6, 9, 11, 12, 13]. The evaluation results across these groups are summarized in Table 4. The performance metrics for all groups are remarkably consistent, with a variance of less than 1%. This indicates that the system's intrusion detection performance is relatively insensitive to the specific composition of the training set. Furthermore, these results validate the efficacy of the feature reconstruction module in capturing intrinsic data correlations. By delving into the underlying patterns of the signal, the

module maintains high stability across diverse training data, ensuring reliable identification of potential threats even when encountering unknown samples.

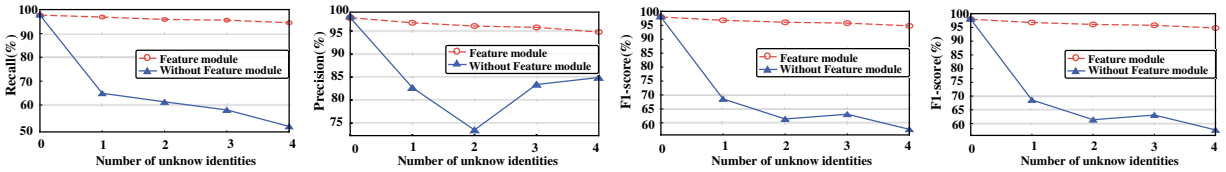
Table 4: Training groups.

Groups	Acc.	Pre.	Rec.	F1
[1, 2, 3, 4, 5]	90.3%	92.0%	90.4%	90.8%
[1, 3, 5, 7, 9]	91.7%	93.2%	91.8%	92.1%
[2, 4, 6, 7, 10]	90.4%	92.2%	90.5%	90.9%
[6, 9, 11, 12, 13]	90.5%	92.5%	90.7%	91.1%

4.7 Ablation Test

4.7.1 Impact of Feature Reconstruction

To verify the effectiveness of the proposed feature reconstruction module in intrusion detection, we conducted a series of ablation experiments. The experimental setup involved 5 known identities for training and 4 unknown identities for testing. The comparative results are illustrated in Fig. 9a–d.



(a) Accuracy **(b) Precision** **(c) Recall** **(d) F1_Score**

Figure 9: Ablation test.

As illustrated in Fig. 9a, while the overall accuracy of both models declines as the number of unknown individuals increases, the baseline model without the feature reconstruction module exhibits a significantly sharper drop; this is primarily because its reliance on supervised learning causes features to be overly coupled with specific labels, limiting its adaptability to unseen data. This constraint is further evidenced in Fig. 9b, where the baseline model's precision initially falls and then shows an anomalous increase, a phenomenon stemming from a conservative prediction strategy adopted when the model's known-identity features prove insufficient for classifying novel samples. Conversely, the recall comparison in Fig. 9c reveals that the baseline model struggles more severely to generalize, frequently misclassifying unknown samples as negative instances and resulting in a steeper decline compared to the proposed model. These trends are synthesized in Fig. 9d, where the F1-score confirms that the feature reconstruction module achieves a more robust balance between precision and recall by effectively decoupling identity, specific information from latent signal characteristics, thereby ensuring reliable and stable intrusion detection in dynamic open-set environments.

4.7.2 Impact of Multi-Scale CNN Module

To verify the effectiveness of the Multi-scale Convolutional Neural Network (MCNN) layer, we conducted an ablation study comparing three architectural variants: (1) the proposed MCNN, (2) a single-scale CNN employing three fixed-size kernels, and (3) a baseline model with the CNN module entirely removed. As shown in Table 5, the MCNN module plays a critical role in both closed-set recognition and open-set intrusion detection. By leveraging kernels of varying sizes, the model is able to capture complementary gait

features across multiple temporal granularities. Removing the MCNN results in a substantial performance degradation; in particular, the detection rate for unknown classes decreases from 73.5% to 65.2%. These findings highlight the importance of multi-scale feature extraction for identifying subtle gait anomalies in open-set scenarios.

Table 5: Ablation study on multi-scale CNN.

Model Variant	Known Class Acc.	Unknown Detection (N = 10)
TFF-IID (Full)	92.8%	73.5%
Standard Single-scale CNN	88.4%	68.9%
No CNN (Transformer only)	84.1%	65.2%

4.8 Robustness Analysis

4.8.1 Impact of Unknown Number of Individuals

Building on the previous analysis involving four unknown identities, this section investigates the impact of an increasing number of unknown individuals on intrusion detection performance. Notably, the training set remains fixed, consisting of data from five randomly selected individuals, ensuring that any performance variations are solely attributable to the expanded test set complexity. The evaluation results are summarized in Table 6.

Table 6: Training groups.

Unkonw Num	Acc	Pre	Rec	F1
5	90.3%	92.0%	90.4%	90.8%
6	88.5%	90.4%	88.4%	88.9%
7	85.2%	88.8%	85.1%	85.3%
8	83.2%	86.6%	83.1%	84.1%
9	78.4%	84.2%	78.3%	80.1%
10	73.5%	82.4%	73.6%	76.2%
11	72.6%	79.1%	72.5%	74.3%

4.8.2 Robustness on Different Environments

To evaluate the robustness of the proposed model across different scenarios, experiments were conducted in three environments: laboratory, office, and lobby. The results are presented in Table 7. In the laboratory environment, the model achieved an accuracy of 73.5%. In the office scenario, where severe static multipath interference and signal fading occurred, accuracy slightly decreased to 70.4%, representing a 3.1% drop. In the open-space lobby, where static multipath reflections were minimal, the model achieved an accuracy of 75.7% and an F1-score of 79.2%, indicating that in more open environments, CSI signals more effectively capture dynamic path variations induced by human gait.

Overall, these results demonstrate that the proposed model exhibits strong robustness across scenarios with varying spatial layouts and multipath complexity. It can effectively mitigate interference caused by environmental variations, consistently extract gait features, and maintain high recognition performance.

Table 7: Cross-scenario performance at the 10-unknown.

Scenario	Environment Features	Acc.	Pre.	Rec.	F1
Laboratory	9 m × 5 m	73.5%	82.4%	73.6%	76.2%
Office	6 m × 6 m	69.4%	75.8%	69.1%	74.5%
Hall	15 m × 10 m	75.7%	83.2%	77.6%	79.2%

4.9 Comparative Experiment

We evaluate the proposed method against several representative baseline algorithms. Due to the limited availability of specialized WiFi-based intrusion detection models, we include established open-set recognition methods from the image processing domain to ensure a rigorous and comprehensive assessment. For this comparison, the models are trained on five known identities and evaluated against a test set containing ten unknown labels. The comparative results are summarized in Table 8.

Table 8: Comparison results.

Model	Acc	Pre	Rec	F1
Ours	73.5%	82.4%	73.6%	76.2%
Caution [26]	53.88%	47.96%	53.33%	50.50%
WiAU [36]	45.7%	52.0%	44.35%	47.86%
PRL [37]	42.53%	42.52%	42.63%	42.58%
APRL [38]	43.33%	45.45%	42.83%	44.12%

5 Conclusion

This paper presents TFF-IID, a device-free intrusion detection framework that synergizes supervised and self-supervised learning. By integrating a feature reconstruction module, the model effectively captures the intrinsic characteristics of wireless signals, significantly enhancing its capability to identify unknown intrusion behaviors. Experimental results demonstrate the robustness of TFF-IID in complex open-set scenarios; notably, the model achieves a 73.5% accuracy even when faced with ten simultaneous unknown intruders, validating its effectiveness for reliable indoor security.

Acknowledgement: The authors would like to express their gratitude to all the volunteers who participated in the data collection phase of this study.

Funding Statement: This work was supported by the Shaanxi Province Outstanding Youth Science Foundation Project (2025JC-JCQN-074).

Author Contributions: The authors confirm their contribution to the paper as follows: Conceptualization: Lijun Cui and Pengfei Xu; Methodology: Lijun Cui; Software: Lijun Cui and Yongjie Niu; Validation: Yuxiang Sun, Xiaokang Gu and Jing Guo; Formal analysis: Lijun Cui; Investigation: Lijun Cui and Yuxiang Sun; Resources: Pengfei Xu; Data curation: Lijun Cui and Xiaokang Gu; Writing—original draft preparation: Lijun Cui; Writing—review and editing: Pengfei Xu and Yongjie Niu; Visualization: Lijun Cui and Jing Guo; Supervision: Pengfei Xu; Project administration: Pengfei Xu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and analyzed during the current study are not publicly available due to privacy restrictions related to the biometric gait information of the participants. However, the data are available from the corresponding author upon reasonable request for academic verification purposes.

Ethics Approval: Not applicable. This study focused on the technical analysis of WiFi Channel State Information (CSI) for gait recognition. The data collection process was non-invasive and involved no medical or clinical procedures. All volunteers participated on a fully informed and voluntary basis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nayak R, Behera MM, Pati UC, Das SK. Video-based real-time intrusion detection system using deep-learning for smart city applications. In: Proceedings of the 2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS); 2019 Dec 16–19; Goa, India. p. 1–6. doi:10.1109/ants47819.2019.9117960.
2. Lian J, Du C, Lou J, Chen L, Yuan X. EchoSensor: fine-grained ultrasonic sensing for smart home intrusion detection. ACM Trans Sen Netw. 2024;20(1):1–24. doi:10.1145/3615658.
3. Ang NS, Guerrero PLA, Namoc MEG, Sy WBT, Visda PMC, Magon SA, et al. Simulation and analysis of an infrared-based intrusion detection system. In: Proceedings of the 2024 IEEE 16th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM); 2024 Nov 26–29; Baguio, Philippines. p. 1–4. doi:10.1109/hnicem64917.2024.11258804.
4. Manssor SAF, Sun S, Abdalmajed M, Ali S. Real-time human detection in thermal infrared imaging at night using enhanced Tiny-yolov3 network. J Real Time Image Process. 2022;19(2):261–74. doi:10.1007/s11554-021-01182-z.
5. Han Z, Lu Z, Wen X, Zheng W, Zhao J, Guo L. CentiTrack: toward centimeter-level passive gesture tracking with commodity WiFi. IEEE Internet Things J. 2023;10(14):13012–27. doi:10.1109/jiot.2023.3256520.
6. Moghaddam MG, Shirehjini AAN, Shirmohammadi S. A WiFi-based method for recognizing fine-grained multiple-subject human activities. IEEE Trans Instrum Meas. 2023;72:1–13. doi:10.1109/tim.2023.3289547.
7. Tian Y, Wang Y, Wang Y, Tong X, Liu X, Qu W. Device-free human tracking and gait recognition based on the smart speaker. IEEE Trans Mobile Comput. 2024;23(11):10610–27. doi:10.1109/tmc.2024.3379647.
8. Guo Z, Yuan W, Gui L, Sheng B, Xiao F. BreatheBand: a fine-grained and robust respiration monitor system using WiFi signals. ACM Trans Sen Netw. 2023;19(4):1–18. doi:10.1145/3582079.
9. Ge Y, Wang H, Ho IW. WiRe-breath: a sustainable WiFi-based real-time respiratory monitoring solution. IEEE J Biomed Health Inform. 2025:1–12. doi:10.1109/jbhi.2025.3619664.
10. Liang Y, Wu W, Li H, Han F, Liu Z, Xu P, et al. WiAi-ID: Wi-Fi-based domain adaptation for appearance-independent passive person identification. IEEE Internet Things J. 2024;11(1):1012–27. doi:10.1109/jiot.2023.3288767.
11. Zhang J, Wei B, Hu W, Kanhere SS. WiFi-ID: human identification using WiFi signal. In: Proceedings of the 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS); 2016 May 26–28; Washington, DC, USA. p. 75–82. doi:10.1109/dcoss.2016.30.
12. Liang Y, Wu W, Li H, Chang X, Chen X, Peng J, et al. DCS-gait: a class-level domain adaptation approach for cross-scene and cross-state gait recognition using Wi-Fi CSI. IEEE Trans Inf Forensics Secur. 2024;19(11):2997–3007. doi:10.1109/tifs.2024.3356827.
13. Zhang L, Wang C, Ma M, Zhang D. WiDIGR: direction-independent gait recognition system using commercial Wi-Fi devices. IEEE Internet Things J. 2020;7(2):1178–91. doi:10.1109/jiot.2019.2953488.
14. Castelblanco A, Rivera E, Solano J, Tengana L, López C, Ochoa M. Dynamic face authentication systems: deep learning verification for camera close-Up and head rotation paradigms. Comput Secur. 2022;115(8):102629. doi:10.1016/j.cose.2022.102629.
15. Bhatt R, Singh S, Choudhary P, Saini M. An experimental study of the concept drift challenge in farm intrusion detection using audio. In: Proceedings of the 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2022 Nov 29–Dec 2; Madrid, Spain. p. 1–8. doi:10.1109/avss56176.2022.9959493.
16. Yun S, Chen YC, Qiu L. Turning a mobile device into a mouse in the air. In: Proceedings of the Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services; 2015 May 18–22; Florence, Italy. New York, NY, USA: Association for Computing Machinery; 2015. p. 15–29. doi:10.1145/2742647.2742662.

17. Hu J, Zhao Y, Zhang X. Application of transfer learning in infrared pedestrian detection. In: Proceedings of the 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC); 2020 Jul 10–12; Beijing, China. p. 1–4. doi:10.1109/icivc50857.2020.9177438.
18. Tan TF, Teoh SS, Fow JE, Yen KS. Embedded human detection system based on thermal and infrared sensors for anti-poaching application. In: Proceedings of the 2016 IEEE Conference on Systems, Process and Control (ICSPC); 2016 Dec 16–18; Bandar Hilir. p. 37–42. doi:10.1109/spc.2016.7920700.
19. Booranawong A, Jindapetch N, Saito H. A system for detection and tracking of human movements using RSSI signals. IEEE Sens J. 2018;18(6):2531–44. doi:10.1109/jsen.2018.2795747.
20. Wounchoum P, Vanichpattarakul T, Dumumpai K, Chaoboworn V, Saito H, Booranawong A. Effects of human presence and movement on received signal strength levels in a 2.4GHz wireless link: an experimental study. J Electr Eng Technol. 2022;17(4):2419–31. doi:10.1007/s42835-022-01070-x.
21. Gui L, Yuan W, Xiao F. CSI-based passive intrusion detection bound estimation in indoor NLoS scenario. Fundam Res. 2023;3(6):988–96. doi:10.1016/j.fmre.2022.05.015.
22. Eom JY, Jang SU, Jeon WS. Wi-Sniffer: WiFi-based intruder detection system using deep learning and decision tree. In: Proceedings of the 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring); 2023 Jun 20–23; Florence, Italy. p. 1–7. doi:10.1109/vtc2023-spring57618.2023.10199724.
23. Chia ZY, Goh PY, Ong LY, Tan SC. The challenge of dynamic environments in regard to RSSI-based indoor Wi-Fi positioning—A systematic review. Future Internet. 2025;17(12):540. doi:10.3390/fi17120540.
24. Tian Z, Shao L, Zhou M, Wang X. A highly-accurate device-free passive motion detection system using cellular network. In: Proceedings of the 2016 IEEE Wireless Communications and Networking Conference; 2016 Apr 3–6; Doha, Qatar. p. 1–6. doi:10.1109/wcnc.2016.7565078.
25. Mendez D, Zennaro M, Altayeb M, Manzoni P. On TinyML WiFi fingerprinting-based indoor localization: comparing RSSI vs. CSI utilization. In: Proceedings of the 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC); 2024 Jan 6–9; Las Vegas, NV, USA. p. 1–6. doi:10.1109/ccnc51664.2024.10454828.
26. Wang D, Yang J, Cui W, Xie L, Sun S. CAUTION: a robust WiFi-based human authentication system via few-shot open-set recognition. IEEE Internet Things J. 2022;9(18):17323–33. doi:10.1109/jiot.2022.3156099.
27. Tian Z, Li Y, Zhou M, Li Z. WiFi-based adaptive indoor passive intrusion detection. In: Proceedings of the 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP); 2018 Nov 19–21; Shanghai, China. p. 1–5. doi:10.1109/icdsp.2018.8631613.
28. Wang J, Tian Z, Zhou M, Wang J, Yang X, Liu X. Leveraging hypothesis testing for CSI based passive human intrusion direction detection. IEEE Trans Veh Technol. 2021;70(8):7749–63. doi:10.1109/tvt.2021.3090800.
29. Xin T, Guo B, Wang Z, Wang P, Lam JCK, Li V, et al. FreeSense: a robust approach for indoor human detection using Wi-Fi signals. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018;2(3):1–23. doi:10.1145/3264953.
30. Zeng Y, Pathak PH, Mohapatra P. WiWho: WiFi-based person identification in smart spaces. In: Proceedings of the 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN); 2016 Apr 11–14; Vienna, Austria. p. 1–12. doi:10.1109/ipsn.2016.7460727.
31. Mohd Noor MH, Tan SY, Ab Wahab MN. Deep temporal conv-LSTM for activity recognition. Neural Process Lett. 2022;54(5):4027–49. doi:10.1007/s11063-022-10799-5.
32. Chen Z, Zhang L, Jiang C, Cao Z, Cui W. WiFi CSI based passive human activity recognition using attention based BLSTM. IEEE Trans Mobile Comput. 2019;18(11):2714–24. doi:10.1109/tmc.2018.2878233.
33. Jobanputra C, Bavishi J, Doshi N. Human activity recognition: a survey. Procedia Comput Sci. 2019;155(1):698–703. doi:10.1016/j.procs.2019.08.100.
34. Law J, Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust statistics-the approach based on influence functions. Statistician. 1986;35(5):565. doi:10.2307/2987975.
35. Li B, Cui W, Wang W, Zhang L, Chen Z, Wu M. Two-stream convolution augmented transformer for human activity recognition. Proc AAAI Conf Artif Intell. 2021;35(1):286–93. doi:10.1609/aaai.v35i1.16103.
36. Lin C, Hu J, Sun Y, Ma F, Wang L, Wu G. WiAU: an accurate device-free authentication system with ResNet. In: Proceedings of the 2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON); 2018 Jun 11–13; Hong Kong, China. p. 1–9. doi:10.1109/sahcn.2018.8397108.

37. Chen G, Qiao L, Shi Y, Peng P, Li J, Huang T, et al. Learning open set network with discriminative reciprocal points. In: Computer vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 507–22. doi:10.1007/978-3-030-58580-8_30.
38. Chen G, Peng P, Wang X, Tian Y. Adversarial reciprocal points learning for open set recognition. IEEE Trans Pattern Anal Mach Intell. 2021;44(11):8065–81. doi:10.1109/tpami.2021.3106743.