

Article

A Foundational Edge-AI Sensing Framework for Occupancy-Driven Energy Management in SMOs

Yutong Chen ^{1,*} , Daisuke Sumiyoshi ², Xiangyu Wang ¹, Takahiro Yamamoto ³, Takahiro Ueno ⁴ and Jewon Oh ⁵

¹ Graduate School of Human-Environment Studies, Kyushu University, Fukuoka 819-0395, Japan; wangxiangyu@energy.arch.kyushu-u.ac.jp

² Faculty of Human-Environment Studies, Kyushu University, Fukuoka 819-0395, Japan; sumiyoshi@arch.kyushu-u.ac.jp

³ Faculty of Engineering and Design, Kagawa University, Takamatsu 761-0396, Japan; yamamoto.takahiro@kagawa-u.ac.jp

⁴ Department of Architecture, Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu City 802-8577, Japan; ueno@kitakyu-u.ac.jp

⁵ Department of Architecture, Faculty of Engineering, Sojo University, Kumamoto City 860-0082, Japan; ohjewon@arch.sojo-u.ac.jp

* Correspondence: chen.yutong.262@s.kyushu-u.ac.jp

Abstract

Occupant presence is a primary driver of Heating, Ventilation, and Air Conditioning (HVAC) and lighting energy consumption in office environments. Existing occupancy-sensing solutions often rely on privacy-sensitive modalities or require costly infrastructure, limiting their applicability in Small and Medium Offices (SMOs). To address these limitations, this study proposes a lightweight CSI-based occupancy-sensing framework based on a dual-core ESP32-S3 architecture, enabling concurrent CSI processing, environmental sensing, and cloud communication. A multi-stage signal preprocessing pipeline compresses raw CSI streams into a compact 56×8 statistical feature matrix, achieving 98.86% classification accuracy for multi-level occupancy estimation. Compared with image-based baselines such as DenseNet121, the proposed approach reduces input data size to 24 kB and model parameters to 138 K, yielding over $129\times$ reduction in transmission volume without sacrificing performance. These results demonstrate that the proposed framework provides a practical, privacy-preserving, and edge-deployable solution for occupancy-aware energy management in SMOs.

Keywords: channel state information; device-free occupancy detection; edge computing; mini transformer; dual-core architecture; feature extraction; smart office energy management



Academic Editor: Grigorios E. Koulouras

Received: 10 January 2026

Revised: 1 March 2026

Accepted: 4 March 2026

Published: 5 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Buildings represent a major source of global CO₂ emissions, making energy efficiency in the built environment a critical priority. While advanced automation and optimization techniques have driven significant efficiency gains in residential and large commercial buildings, these approaches are not readily applicable to Small and Medium Offices (SMOs). Unlike large commercial buildings, which increasingly deploy Building Energy Management Systems (BEMS) supported by dedicated facility teams, SMOs typically lack centralized energy-management infrastructure, professional energy engineers, and the financial capacity to adopt advanced sensing or control systems. Consequently, energy-related decisions in SMOs are often made manually or governed by fixed schedules, resulting in

persistent inefficiencies such as unnecessary cooling, suboptimal temperature setpoints, and prolonged lighting operation during low-occupancy periods. Occupant behavior has been widely recognized as a key determinant of building energy consumption [1,2]. Addressing energy inefficiency in SMOs therefore requires solutions that are cost-effective, easy to deploy, and operable without dedicated on-site management.

In such environments, reliable occupancy information is fundamental to effective energy management, supporting both adaptive system operation and the delivery of personalized energy-saving feedback to occupants. However, acquiring reliable occupancy data in SMOs remains challenging: vision-based systems introduce significant privacy risks and are difficult to deploy in shared office environments [3], while device-assisted approaches such as badge tracking or smartphone sensing depend on active user compliance and are therefore unsuitable for continuous passive monitoring [4]. Conventional sensor modalities, such as Passive Infrared (PIR) detectors [5] and CO₂-based inference [6], are constrained by limited spatial resolution, delayed response times, and poor robustness under varying environmental conditions.

Wi-Fi Channel State Information (CSI) has emerged as a promising modality for device-free occupancy sensing, owing to its fine-grained characterization of multipath propagation and its compatibility with commodity wireless infrastructure [7]. By combining fine-grained channel measurements with machine learning techniques, CSI-based methods have demonstrated strong performance in crowd counting and human activity recognition [8–11]. Unlike camera-based systems, CSI sensing operates without capturing visual information, making it inherently suitable for privacy-sensitive deployment environments [12]. For instance, Zou et al. [13] proposed WiFree, achieving 99.1% binary occupancy detection accuracy and 92.8% crowd counting accuracy in controlled settings. Yang et al. [11] demonstrated CSI-based human detection and activity recognition in smart-home environments, reporting accuracies of up to 98% using commodity Wi-Fi routers. However, these studies are predominantly evaluated in controlled laboratory settings and do not address the system-level constraints of practical SMO deployment, including hardware limitations, long-term operational stability, and integration with energy management workflows.

Beyond algorithmic limitations, practical deployment of CSI-based sensing is further constrained by hardware dependencies and the substantial computational overhead of high-dimensional CSI processing, which typically demands PC- or cloud-level computing resources [7–10]. Furthermore, only a small number of studies have established an explicit link between CSI-derived occupancy estimates and practical energy management actions. Natarajan et al. [14] demonstrated an ESP32-based CSI Human Activity Recognition (HAR) system capable of driving smart LED control, reporting estimated energy savings under representative occupancy scenarios. However, such demonstrations remain appliance-level and rule-driven and fall short of addressing SMO-oriented system requirements, including long-term operational robustness, multi-sensor coordination, and feedback mechanisms capable of sustaining behavioral change.

In the absence of centralized BEMS and dedicated facility personnel, energy conservation in SMOs relies predominantly on occupants' voluntary behavioral responses. Research in behavioral science has demonstrated that delivering actionable, occupancy-aware energy-saving recommendations can substantially raise occupants' energy consciousness and promote sustained conservation behaviors [15]. Accordingly, a deployable and reliable occupancy-sensing capability is a prerequisite for generating the context-aware, occupancy-driven feedback that underpins effective behavioral energy management in SMOs.

Despite these advances, translating CSI-based occupancy sensing into a practically deployable system for SMOs remains an open challenge [16]. Beyond computational

constraints, existing CSI systems rarely achieve the end-to-end integration required for long-term unattended operation under the practical constraints of SMO environments. As a result, despite the demonstrated potential of CSI sensing, no integrated system-level solution currently exists that jointly addresses deployable occupancy estimation, robust edge preprocessing, and feedback-oriented behavioral energy management under SMO constraints.

Recent deep learning-based methods have substantially advanced CSI-based sensing accuracy and robustness, yet their computational and memory demands remain prohibitive for low-cost edge deployment in SMOs. In our prior work [17], four image-based deep learning architectures (Vision Transformer, ResNet50, DenseNet121, and EfficientNetB0) were systematically benchmarked for CSI-based occupancy classification, with DenseNet121 identified as the strongest-performing baseline. Nevertheless, even this best-performing candidate depends on high-dimensional time-frequency spectrograms and cloud-level computation, rendering it unsuitable for direct edge deployment in resource-constrained SMO environments. To bridge this gap, Tiny Machine Learning (TinyML) techniques have been increasingly adopted to compress and optimize deep learning models for deployment on microcontroller-class hardware [18]. Armenta-Garcia et al. [19] demonstrated the feasibility of CSI-based inference on ESP32-S3 microcontrollers by leveraging model quantization, Pseudo Static Random Access Memory (PSRAM) utilization, and optimized embedded toolchains, achieving 92.43% accuracy for activity recognition. However, these efforts focus primarily on model executability rather than the system-level requirements of end-to-end deployable sensing, including robust signal preprocessing under non-stationary conditions and stable long-term operation on resource-constrained devices.

Beyond runtime feasibility, CSI-based occupancy sensing is a temporal inference problem by nature, as occupancy states and human activities produce time-varying amplitude patterns in CSI streams. Transformer architectures have proven effective at capturing long-range temporal dependencies in building energy applications, including energy consumption forecasting [20] and HVAC fault detection [21]. These results indicate that attention-based sequence modeling is well suited to the temporal structure of CSI data. Nevertheless, standard Transformer architectures are computationally intensive, precluding their direct deployment on microcontroller-class edge devices for on-device occupancy inference. This gap motivates the design of lightweight Transformer variants that preserve sequence modeling capability while satisfying the memory, latency, and robustness requirements of SMO edge deployments.

The BI-Tech (Behavioral Insight × Technology) project was initiated by Japan's Ministry of the Environment in 2019 to bridge the gap between occupancy-sensing technology and occupant-driven energy conservation in office environments. The BI-Tech approach integrates IoT-based environmental sensing with data-driven decision support to promote self-motivated energy-saving behaviors, shifting the focus from passive automated control to active occupant engagement. Prior work within this framework [22] has shown that systems delivering context-aware feedback and personalized energy-saving suggestions can raise occupants' energy awareness and yield measurable reductions in energy consumption. However, the effectiveness of such behavioral interventions depends on reliable, low-cost occupancy sensing integrated with environmental monitoring, a requirement that existing CSI research has not yet fulfilled for SMO environments.

Taken together, these observations identify a critical gap: despite the technical maturity of CSI sensing and the demonstrated effectiveness of behavioral energy management, few existing systems jointly address low-cost occupancy estimation, multi-modal environmental monitoring, and feedback-oriented energy management under SMO constraints. To address

this gap, this study presents a deployable edge-AI occupancy-sensing framework for SMOs within the BI-Tech platform, with the following three contributions:

- A dual-core ESP32-S3 edge platform with PSRAM support, enabling concurrent multi-stage CSI preprocessing and multimodal environmental sensing on a single low-cost node;
- A lightweight Mini Transformer model designed for the compact 56×88 CSI feature space, achieving 98.86% accuracy in multi-level occupancy classification with 138K parameters and a per-window data footprint of 24 kB, corresponding to a data volume reduction exceeding $129\times$ compared with the DenseNet121 baseline;
- Integration of the occupancy-sensing module with multimodal environmental sensors and the BI-Tech behavioral intervention platform, enabling occupancy-aware energy-saving reminders for lighting and HVAC management in SMO environments.

2. System Architecture and Edge Implementation

2.1. Design Constraints in SMOs

Accurate occupancy information has been widely identified as a prerequisite for intelligent building operation and energy-efficient control. Chaudhari et al. [23] demonstrated that occupancy data is central to optimizing lighting, HVAC, and ventilation systems, particularly under occupant-driven control strategies. However, deploying occupant-centric sensing systems in SMOs entails practical constraints that are not present in large commercial buildings. Compared with large commercial buildings, SMOs typically operate under constrained budgets, depend on fragmented facility management services, and lack dedicated energy management personnel. These constraints impede the adoption of conventional BEMS, which typically require substantial upfront investment, specialized maintenance, and professionally trained operators. As a result, occupancy-sensing solutions for SMOs must be low-cost, straightforward to deploy, and capable of autonomous long-term operation with minimal maintenance overhead, while providing reliable estimates to support occupancy-driven control of lighting and HVAC systems.

2.2. Overall BI-Tech Architecture

Building on the design requirements established above, our previous work introduced BI-Tech as an IoT-based platform for energy monitoring and feedback-oriented behavioral intervention in SMOs [22]. In this paper, a four-layer architectural view is adopted to clarify the placement of the proposed occupancy module and its interfaces with downstream processing and decision functions. As shown in Figure 1, the BI-Tech architecture comprises four layers: (i) a Perception layer that collects Wi-Fi CSI from the transmitter node and multimodal environmental signals; (ii) a Processing layer that performs CSI preprocessing and feature extraction, occupancy estimation, and sensor data aggregation; (iii) a Decision layer that integrates occupancy estimates and sensor context to generate occupancy-aware decision cues such as energy-saving reminders; and (iv) an Application layer that delivers user-facing alerts and decision support through the BI-Tech client applications.

This study focuses on the Perception and Processing layers, with particular attention paid to the CSI-based occupancy estimation pipeline and its data interface with the Decision layer. In particular, the study implements CSI-based occupancy estimation on the BI-Tech edge node and defines the structured data outputs required to trigger occupancy-aware reminders in the Decision layer. Although this paper describes the end-to-end workflow and interface design for the sensing-to-feedback pathway, the long-term field validation of occupancy-driven actuation and the quantitative assessment of energy savings remain directions for future work.

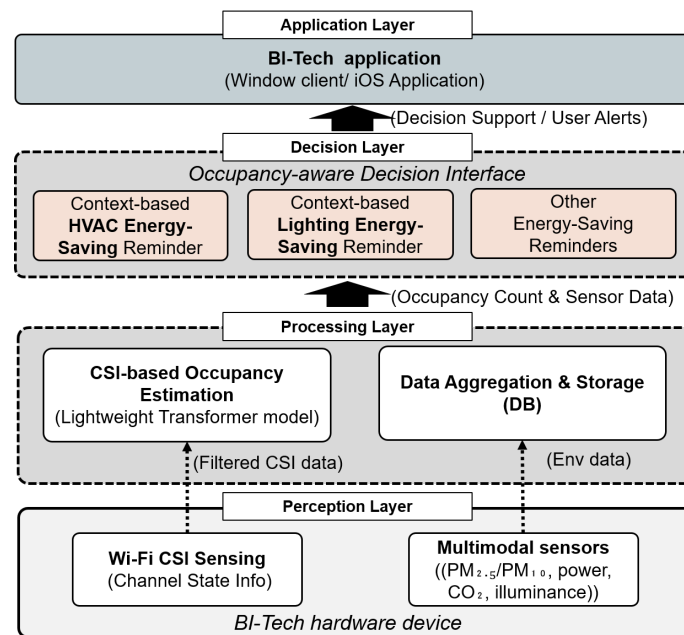


Figure 1. Four-layer BI-Tech architecture and the integration point of the proposed CSI-based occupancy estimation module. The Perception layer collects Wi-Fi CSI and multimodal environmental signals. The Processing layer performs CSI preprocessing/feature extraction and lightweight occupancy inference and aggregates sensor data for storage. The Decision layer consumes the estimated occupancy count and sensor context to generate occupancy-aware decision cues (e.g., HVAC/lighting reminders). The Application layer delivers user-facing alerts through BI-Tech client applications (Windows/iOS).

2.3. Edge Hardware Platform and Dual-Core Design

The edge hardware platform is built around two ESP32-S3 microcontrollers, providing a low-cost foundation for concurrent CSI acquisition and multimodal environmental sensing. One device serves as a dedicated Wi-Fi transmitter and environmental sensing node, while the other operates as the CSI receiver, responsible for signal acquisition, local preprocessing, and data uplink. The ESP32-S3 supports IEEE 802.11n [24] Wi-Fi and CSI extraction through the ESP-IDF framework, meeting the requirements of device-free sensing on microcontroller-class hardware.

The receiver node is configured to support concurrent CSI acquisition and indoor environmental monitoring. In addition to CSI reception, the node integrates a suite of indoor environmental sensors: CO₂ concentration (SCD40), particulate matter PM_{2.5}/PM₁₀ (SDS011 via UART), temperature and humidity (SHT35, expandable via an I²C multiplexer), and horizontal illuminance (TSL2561). This sensor selection is informed by practical deployment experience in SMO environments, where long-term reliability and low maintenance burden are primary considerations. Sensors relevant to indoor air quality are treated as primary, while peripheral sensors can be omitted without compromising the core CSI-based occupancy estimation functionality. An external omnidirectional antenna is employed to improve wireless robustness in multipath-rich indoor environments, benefiting both CSI stability and Message Queuing Telemetry Transport (MQTT) communication reliability. The omnidirectional configuration was selected to ensure consistent spatial coverage regardless of device orientation, which is important for plug-and-play deployment in SMOs where devices may be repositioned by occupants.

Since CSI acquisition requires the receiver to operate on a fixed Wi-Fi channel, a time-division communication strategy is adopted. The receiver alternates between two operating modes within each 5-min cycle: a communication phase and a CSI acquisition

phase. During the communication phase, the receiver connects to the local network, processes environmental sensor readings, and uploads the results to the MQTT broker. During the CSI acquisition phase, the receiver switches to the transmitter's dedicated channel and captures a 6 s CSI burst for occupancy inference. Within each cycle, environmental sensing and MQTT uplink operations are repeated at regular intervals, whereas CSI acquisition is performed once at the end of the cycle. This scheduling design preserves stable CSI measurement conditions on the dedicated channel without disrupting periodic environmental data reporting.

Figure 2 illustrates the dual-core task allocation of the CSI receiver node. To prevent contention between computation-intensive CSI preprocessing and network operations, tasks are statically assigned to dedicated CPU cores under FreeRTOS. Core 0 manages all CSI-related processing, including packet reception, segmentation into fixed-length windows, multi-stage signal filtering, and window-level feature extraction. In parallel, Core 1 handles network connectivity and peripheral I/O operations, including Wi-Fi association, MQTT/TLS session management, structured payload publishing, and periodic environmental sensor sampling. This core-level role separation ensures deterministic timing for CSI processing on Core 0 and uninterrupted network operation on Core 1, both of which are essential for reliable long-term deployment in SMOs.

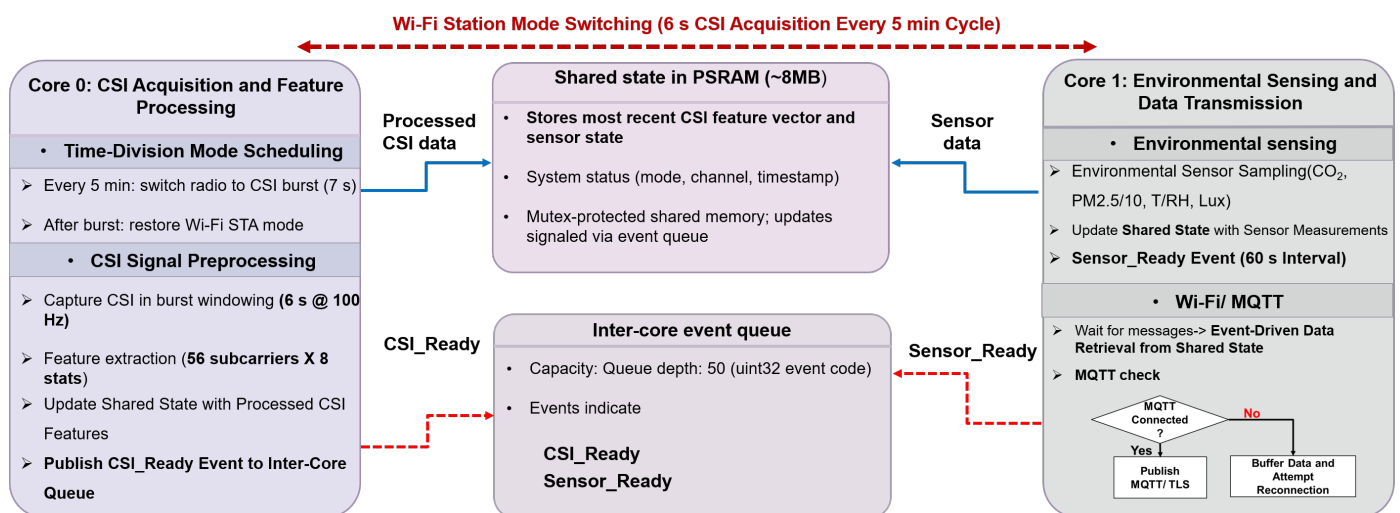


Figure 2. Dual-core workflow of the BI-Tech edge device (ESP32-S3). Core 0 performs real-time CSI preprocessing (Hampel filtering, interpolation, Kalman smoothing, and wavelet transform) and extracts compact features from selected subcarriers. In parallel, Core 1 samples multimodal environmental sensors (e.g., power, CO₂, PM_{2.5}/PM₁₀, illuminance) and logs/transmits the measurements for system integration. Solid arrows denote data transfer between modules, while dashed arrows represent inter-core event signaling.

To support continuous CSI streaming and window-based preprocessing on a resource-constrained microcontroller, external PSRAM is enabled in octal mode (~8 MB in the current hardware configuration). A PSRAM-backed shared state is used to store the latest processed CSI features, environmental sensor snapshots, timestamps, and control variables. Access to shared data is protected by a mutex to ensure consistency. In addition, lightweight inter-core communication is implemented using FreeRTOS queues for event signaling, such as “CSI data ready” and “sensor data ready”. Upon receiving an event, the publishing task retrieves the latest snapshot from the shared state, formats it into a structured payload, and transmits it via MQTT/TLS. MQTT publishing is serialized to prevent conflicts between CSI-related transmissions and periodic sensor uploads. This combined hardware–software design enables reliable CSI acquisition and multimodal

sensing under long-term unattended operation, which is essential for deployment in SMOs. Based on this organization, CSI data and environmental measurements are processed locally on the edge node through a window-based signal processing pipeline. The detailed CSI preprocessing and feature extraction methodology is described in Section 3.

3. CSI Feature Extraction

To enable real-time occupancy inference under the hardware and communication constraints of SMOs, we propose a lightweight CSI feature extraction pipeline that performs local compression and noise suppression on edge devices before cloud transmission. Unlike conventional approaches that transmit raw CSI streams or apply computationally intensive deep feature learning on cloud servers [17,25], our method extracts compact statistical descriptors directly on the ESP32-S3 device, achieving a $129\times$ compression ratio while preserving occupancy-discriminative patterns.

As illustrated in Figure 3, the BI-Tech system adopts a unified edge-side processing architecture in which raw CSI amplitude streams undergo multi-stage denoising, subcarrier selection, and statistical feature construction on the device. The resulting feature snapshots are encoded into structured JSON payloads and transmitted to the cloud via MQTT over TLS. This design ensures distributional consistency between offline training and online deployment, as both stages execute identical feature extraction operations.

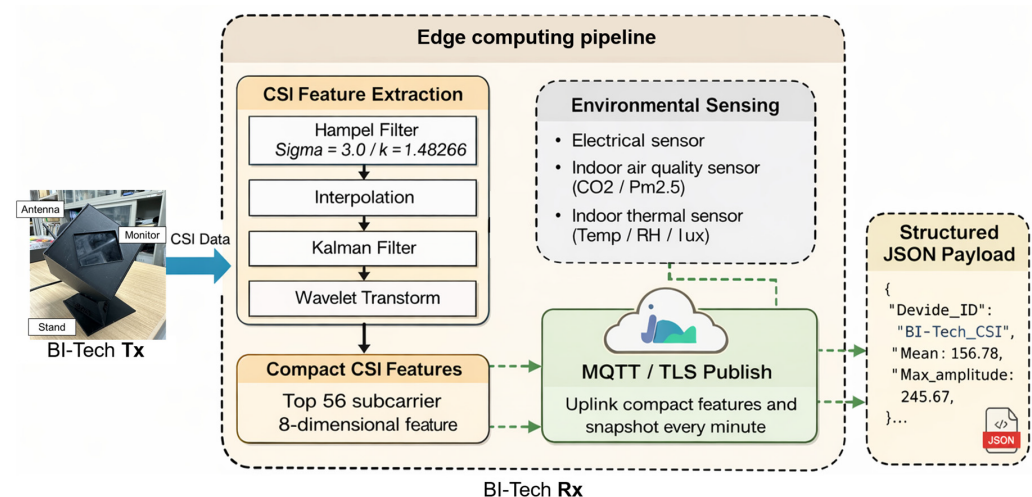


Figure 3. Unified edge-side processing pipeline of the BI-Tech system. Raw CSI streams captured by the BI-Tech receiver undergo four-stage denoising (Hampel filtering, interpolation, Kalman filtering, and wavelet decomposition), followed by subcarrier selection and statistical feature extraction. The compact 56×8 feature matrix is combined with environmental sensing data (temperature, humidity, CO_2) and transmitted to the cloud via MQTT, reducing per-window data volume from 450 KB to 1.75 KB.

3.1. CSI Acquisition and Windowing

CSI provides fine-grained characterization of wireless channels by capturing multipath-induced amplitude and phase variations. In this study, CSI measurements are collected using a single-input single-output (SISO) configuration, with one ESP32-S3 device as transmitter (BI-Tech-Tx) and another as receiver (BI-Tech-Rx), each equipped with a single omnidirectional antenna. CSI amplitude streams are continuously sampled at 100 Hz over a 2.4 GHz Wi-Fi channel with a 20 MHz bandwidth (IEEE 802.11n). Each CSI sample contains 384 bytes of raw I/Q measurements provided by the ESP-IDF CSI API under our packet configuration, comprising the Legacy Long Training Field (LLTF), High-Throughput Long Training Field (HT-LTF), and Space-Time Block Coding High-Throughput Long Training Field (STBC-HT-LTF) [26]. After removing null entries and converting I/Q pairs to ampli-

tudes, 166 valid amplitude channels are obtained. To reduce on-device computation and memory usage, we retain the 56 highest-energy channels for subsequent feature extraction.

The continuous CSI stream is segmented into fixed-length temporal windows of 6 s, corresponding to $T = 600$ samples per window. Adjacent windows overlap by 50% (3 s stride), yielding approximately nine segments per 30-s recording. This windowing scheme is adopted throughout the study to generate consistent input segments for subsequent preprocessing and feature construction.

Let the complex CSI of the i -th subcarrier at time t be expressed as

$$H_i(t) = |H_i(t)|e^{j\varphi_i(t)}, \quad (1)$$

where $|H_i(t)|$ denotes the CSI amplitude, $\varphi_i(t)$ denotes the phase, i is the subcarrier index, and t denotes the discrete time index within the window. Due to phase synchronization challenges in low-cost ESP32-S3 devices, this study focuses exclusively on amplitude information. We define $A_i(t) = |H_i(t)|$ as the amplitude sequence within each window ($t = 1, \dots, T$), where $T = 600$ is the total number of samples per 6-second window.

3.2. Multi-Stage Denoising Pipeline

Raw CSI amplitude sequences captured by the ESP32-S3 are subject to impulsive outliers, missing or corrupted samples, and short-term measurement noise. Following the signal processing workflow in Figure 3, a four-stage preprocessing pipeline is applied to each subcarrier sequence, including Hampel outlier removal, linear interpolation, Kalman smoothing, and wavelet-based denoising, in order to suppress noise components while retaining meaningful temporal variations in the CSI stream.

First, a Hampel filter with window size of $w = 6$ and threshold factor of $\gamma = 3$ is applied to suppress impulsive outliers. The filtered output is defined as

$$A_i^H(t) = \begin{cases} A_i(t), & |A_i(t) - \text{med}(A_i(t-w:t+w))| < \gamma \cdot \text{MAD}_i(t), \\ \text{med}(A_i(t-w:t+w)), & \text{otherwise,} \end{cases} \quad (2)$$

where $A_i^H(t)$ is the Hampel-filtered amplitude of the i -th subcarrier at time t ; $\text{med}(\cdot)$ denotes the median operator computed over the local window $A_i(t-w:t+w)$; $\text{MAD}_i(t)$ is the median absolute deviation of the same local window, computed with normalization constant $k = 1.4826$; w is the Hampel window half-width; and γ is the outlier threshold factor. Samples exceeding $3 \times \text{MAD}$ from the local median are replaced by the median value. The window size $w = 6$ (60 ms at 100 Hz sampling) was chosen to balance outlier detection sensitivity and computational overhead on the ESP32-S3 platform.

Second, missing or corrupted samples flagged during CSI packet reception are reconstructed via piecewise linear interpolation using adjacent valid samples, preserving temporal continuity of the amplitude stream.

Third, a discrete Kalman filter is applied to attenuate Gaussian measurement noise while preserving signal dynamics. The observation noise variance R and process noise variance Q are adaptively set proportional to the variance of the input sequence:

$$R = 0.1 \times \text{Var}(A_i^H), \quad Q = 0.01 \times \text{Var}(A_i^H), \quad (3)$$

where R is the observation noise variance representing measurement uncertainty; Q is the process noise variance representing the expected rate of signal variation; and $\text{Var}(A_i^H)$ denotes the variance of the Hampel-filtered amplitude sequence A_i^H within the current window. The ratio $R/Q = 10$ was empirically tuned to prioritize smoothness over rapid tracking, as occupancy changes occur at timescales longer than individual samples.

Finally, a four-level Daubechies-4 (db4) discrete wavelet decomposition is applied to further suppress high-frequency components (>10 Hz) beyond the Nyquist frequency of human motion. The db4 wavelet was selected for its compact support and good time-frequency localization properties. High-frequency detail coefficients are thresholded using soft thresholding with universal threshold

$$\lambda = \hat{\sigma} \sqrt{2 \log T}, \quad (4)$$

where λ is the soft-thresholding threshold applied to the wavelet detail coefficients; $\hat{\sigma}$ is the noise standard deviation estimated via the median absolute deviation of the finest-scale wavelet coefficients; and T is the number of samples in the current window.

3.3. Subcarrier Selection and Feature Construction

After preprocessing, null subcarrier entries are removed according to the ESP32-S3 specification, yielding 166 valid amplitude channels across the three LTF fields. These channels serve as candidates for subsequent energy-based selection. For each valid subcarrier i , temporal energy is computed as

$$E_i = \sum_{t=1}^T \left[A_i^{\text{denoise}}(t) \right]^2, \quad (5)$$

where $A_i^{\text{denoise}}(t)$ denotes the amplitude of the i -th subcarrier at time t after the four-stage denoising pipeline; T is the number of samples in a 6-second window ($T = 600$); and E_i is the total temporal energy of subcarrier i within the window, used to rank subcarriers for selection. Subcarriers are ranked by E_i in descending order, and the top 56 are retained. The energy criterion is adopted because it preferentially retains channels with higher signal-to-noise ratio, where occupancy-induced perturbations in variance, rate-of-change, and zero-crossing rate are most detectable, and its $O(T)$ computational complexity satisfies the real-time processing constraint of the ESP32-S3 without introducing the overhead of transform-based or supervised dimensionality reduction methods. The value $k = 56$ was chosen to retain sufficient spectral diversity for cross-subcarrier feature discrimination while keeping the feature matrix compact for edge inference; a systematic k -value ablation is identified as a direction for future work.

For each selected subcarrier i , we compute eight statistical descriptors that characterize amplitude distribution, temporal variation, and oscillatory behavior. These features are designed to capture occupancy-discriminative patterns while maintaining computational efficiency for real-time edge processing.

Mean amplitude quantifies the average channel response:

$$\mu_i = \frac{1}{T} \sum_{t=1}^T A_i(t), \quad (6)$$

where μ_i is the mean amplitude of the i -th selected subcarrier within the current window and T is the number of samples per window.

Standard deviation characterizes temporal fluctuation intensity:

$$\sigma_i = \sqrt{\frac{1}{T} \sum_{t=1}^T (A_i(t) - \mu_i)^2}, \quad (7)$$

where σ_i is the standard deviation of the amplitude sequence of the i -th selected subcarrier within the window, and μ_i is the within-window mean defined above. Variance is computed as σ_i^2 to provide an alternative dispersion metric.

Maximum and minimum amplitudes capture the range of channel responses within a window:

$$A_{\max,i} = \max_{t \in [1,T]} A_i(t), \quad A_{\min,i} = \min_{t \in [1,T]} A_i(t), \quad (8)$$

where $A_{\max,i}$ and $A_{\min,i}$ are the maximum and minimum amplitudes of the i -th selected subcarrier within the window, respectively.

Energy is defined as the sum of squared amplitudes:

$$E_i^{\text{feat}} = \sum_{t=1}^T A_i^2(t), \quad (9)$$

where E_i^{feat} denotes the total signal energy of the i -th selected subcarrier within the current window.

Rate of change quantifies short-term amplitude variations between adjacent samples:

$$\Delta A_i = \frac{1}{T-1} \sum_{t=1}^{T-1} |A_i(t+1) - A_i(t)|, \quad (10)$$

where ΔA_i is the mean absolute difference between adjacent samples of the amplitude sequence of the i -th selected subcarrier within the window. Larger ΔA_i values indicate more rapid temporal fluctuations, which are typically associated with human movement.

Zero-crossing rate measures the frequency of sign changes in the zero-mean amplitude sequence, capturing oscillatory patterns:

$$Z_i = \frac{1}{2(T-1)} \sum_{t=1}^{T-1} |\text{sgn}(\tilde{A}_i(t+1)) - \text{sgn}(\tilde{A}_i(t))|, \quad (11)$$

where $\tilde{A}_i(t) = A_i(t) - \mu_i$ denotes the zero-mean amplitude sequence of the i -th subcarrier, obtained by subtracting the within-window mean μ_i ; $\text{sgn}(\cdot)$ is the sign function returning +1 for positive values and −1 for negative values; and Z_i is the normalized count of sign changes per sample pair within the window.

Each 6-s CSI window is thus represented as a compact feature matrix:

$$\mathbf{X} \in \mathbb{R}^{56 \times 8}, \quad (12)$$

where each row corresponds to one selected subcarrier and each column corresponds to one of the eight statistical descriptors defined above. This matrix serves as the direct input to the downstream Mini Transformer encoder.

3.4. Edge-Oriented Compression and Transmission

The proposed statistical feature representation significantly reduces the amount of data transmitted compared to raw CSI. Each 6-second window contains $T = 600$ CSI packets, with 384 bytes per packet under the ESP-IDF configuration, resulting in approximately 225 KB of raw data. After feature extraction, the compact 56×8 matrix requires only 1.75 KB (448 float32 values), corresponding to a compression ratio of approximately $129 \times$.

All feature extraction operations including the four-stage denoising pipeline, subcarrier selection, and statistical computation are executed locally on the ESP32-S3 edge device using an optimized ESP-IDF implementation. The resulting feature matrices are encoded into structured JSON payloads and transmitted to the cloud server via MQTT

over TLS. This edge-first design reduces network bandwidth requirements, minimizes transmission latency, and ensures consistent feature distributions between offline training and online deployment. Table 1 summarizes the data volume reduction achieved at each processing stage.

Table 1. Data Compression Achieved by Edge-Side Feature Extraction.

Processing Stage	Data Size	Compression Ratio
Raw CSI (384 bytes $\times$ 600 samples)	225 KB	1 $\times$
After subcarrier selection (56 channels $\times$ 600 samples)	66 KB	3.4 $\times$
Statistical features (56 $\times$ 8 matrix)	1.75 KB	129 $\times$

4. Mini Transformer Model

To enable real-time occupancy inference under the hardware and communication constraints of small and medium offices (SMOs), we develop a Mini Transformer that operates directly on the compact CSI feature matrices extracted on the ESP32-S3. The model takes the feature matrix X constructed in the previous subsection and outputs the probability over six occupancy classes (0P to 5P+). The overall architecture is shown in Figure 4.

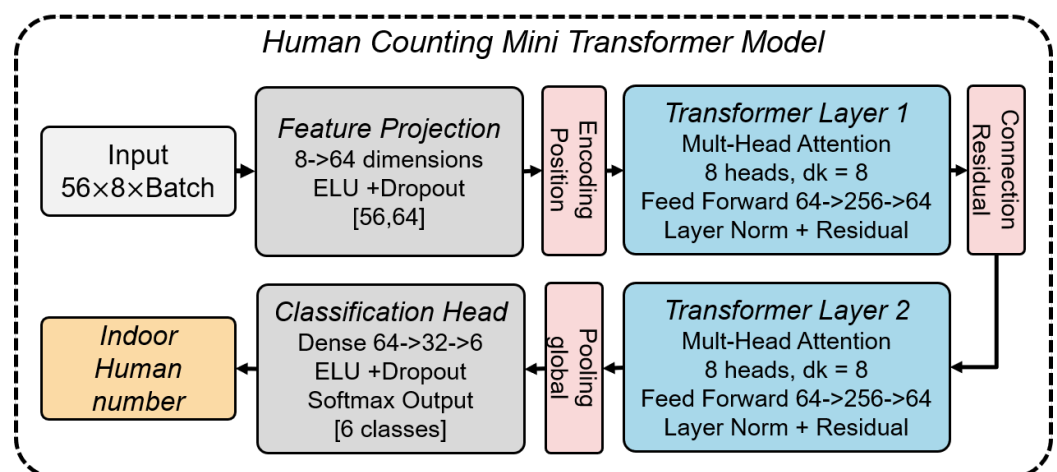


Figure 4. Mini Transformer architecture for occupancy classification using compact CSI feature matrices. The input feature tokens are projected from 8 to 64 dimensions with ELU and dropout, combined with positional encoding, and processed by two Transformer encoder layers (8-head self-attention and 64–256–64 feed-forward). Global pooling and a lightweight classification head (64–32–6) output six occupancy classes (0P to 5P+).

4.1. Input Feature Sequence

The input to the model is the compact CSI feature matrix X constructed from the selected 56 subcarriers and eight statistical descriptors. Each row corresponds to one selected subcarrier and each column corresponds to a statistical descriptor computed from the within-window amplitude sequence. In this way, temporal variations are preserved at the feature level, while the Mini Transformer learns interactions among subcarriers for occupancy discrimination. Although the model does not ingest raw CSI time series, the temporal dynamics within each 6-s window are summarized by the statistical descriptors (e.g., dispersion- and fluctuation-related measures). Therefore, temporal information is preserved implicitly at the feature level, while the Transformer focuses on modeling cross-subcarrier dependencies for occupancy discrimination.

4.2. Feature Projection and Positional Encoding

As shown in Figure 4, the input matrix is treated as a 56-token sequence, where each token is an 8-dimensional descriptor vector of one selected subcarrier. Each token is projected to the embedding dimension $d_{\text{model}} = 64$ through a lightweight projection layer followed by ELU and dropout. ELU was selected based on a controlled activation function comparison conducted in ablation study:

$$\mathbf{H}^{(0)} = \text{Dropout}(\text{ELU}(\mathbf{X}\mathbf{W}_e + \mathbf{b}_e)), \quad \mathbf{H}^{(0)} \in \mathbb{R}^{56 \times 64}. \quad (13)$$

A positional encoding term is then added to provide each token with a consistent index and to break the permutation symmetry of the token set before entering the encoder.

4.3. Mini Transformer Encoder

The encoder contains two Transformer layers. In each layer, multi-head self-attention with 8 heads models global interactions among subcarrier tokens, followed by a position-wise feed-forward network (64–256–64). Residual connections and layer normalization are used to stabilize training:

$$\tilde{\mathbf{H}}^{(\ell)} = \text{LN}(\mathbf{H}^{(\ell-1)} + \text{MHA}(\mathbf{H}^{(\ell-1)})), \quad (14)$$

$$\mathbf{H}^{(\ell)} = \text{LN}(\tilde{\mathbf{H}}^{(\ell)} + \text{FFN}(\tilde{\mathbf{H}}^{(\ell)})), \quad (15)$$

for $\ell = 1, 2$, $\text{MHA}(\cdot)$ denotes multi-head self-attention and $\text{FFN}(\cdot)$ denotes the feed-forward network.

4.4. Global Pooling and Classification

After the second Transformer layer, global pooling aggregates the 56-token sequence into a single representation:

$$\mathbf{g} = \frac{1}{56} \sum_{i=1}^{56} \mathbf{H}_i^{(2)} \in \mathbb{R}^{64}. \quad (16)$$

A lightweight classification head then predicts the occupancy level using a two-layer MLP (64–32–6) with ReLU and dropout, followed by a softmax output:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_2 \text{Dropout}(\text{ELU}(\mathbf{W}_1 \mathbf{g} + \mathbf{b}_1)) + \mathbf{b}_2), \quad (17)$$

where $\hat{\mathbf{y}} \in \mathbb{R}^6$ denotes the probability over the six occupancy classes (0P to 5P+).

4.5. Design Rationale and Robustness to Subcarrier Cancellation

Several CSI studies report that, in the presence of human bodies, instantaneous CSI amplitudes may exhibit subcarrier-level cancellation due to multipath superposition [11,27–29]. Such cancellation mainly affects instantaneous amplitudes on individual subcarriers, making snapshot-based representations sensitive to fading sign changes and local destructive interference.

In contrast, our method does not depend on frame-to-frame subtraction or image differencing. Instead, each 6-s window is summarized by statistical descriptors computed from denoised amplitude sequences on the selected subcarriers. Human presence induces persistent non-stationarity in the channel (e.g., increased fluctuation magnitude and energy redistribution), which is reflected in window-level statistics such as variance, energy, and rate-of-change even when some instantaneous subcarriers partially cancel. Moreover, the Mini Transformer models demonstrate cross-subcarrier interactions over the 56-token sequence, allowing the classifier to integrate information across multiple

subcarriers rather than relying on any single subcarrier response. This design improves robustness to subcarrier-specific fading and cancellation and enables reliable occupancy estimation without explicit reference subtraction.

5. Experimental Setup and Evaluation Protocol

5.1. Experiment Configuration

All experiments were conducted in a typical indoor office environment representative of SMO deployment conditions. A single-input single-output (SISO) Wi-Fi sensing link was established using two ESP32-S3 devices: one configured as the dedicated transmitter (BI-Tech-Tx) and the other as the CSI receiver (BI-Tech-Rx). Both devices were equipped with a single omnidirectional antenna. To evaluate robustness to multipath variations induced by different transceiver placements, we designed four deployment configurations with distinct Tx/Rx locations around a typical office desk arrangement, as shown in Figure 5. In each case, the number of occupants was controlled from 0 to 5+ persons, and participants were instructed to maintain natural sitting/standing behavior within the designated area. These cases reflect practical SMO conditions, where the sensing link may traverse different propagation paths and be affected by furniture and wall reflections.

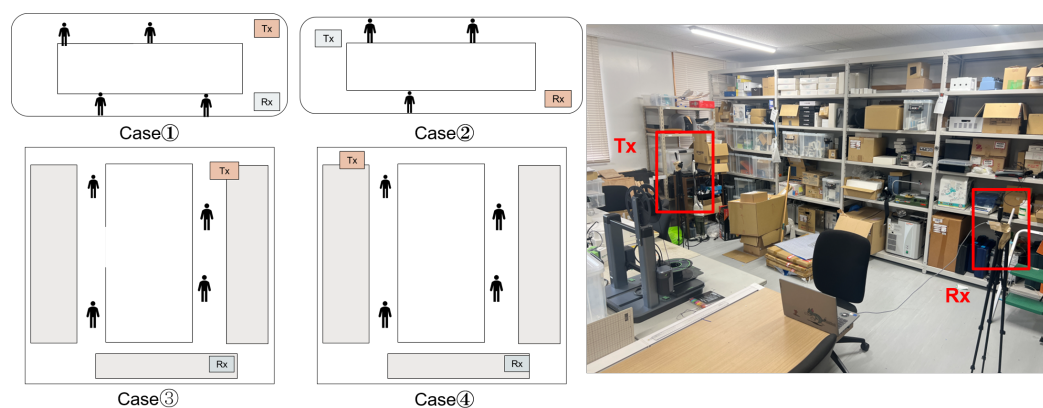


Figure 5. Four deployment cases used for evaluating robustness to transceiver placement and multipath conditions. BI-Tech-Tx and BI-Tech-Rx denote the transmitter and receiver ESP32-S3 nodes, respectively. Each case places Tx/Rx at different sides of the desk layout to induce distinct propagation geometries while keeping the occupied region comparable. Line-of-Sight (LOS) (Case 1, 3) and Non-Line-of-Sight (NLOS) (Case 2, 4) conditions simulating realistic office occupancy scenarios.

5.2. Data Characterization and Input Structuring

5.2.1. Data Acquisition and Window-Level Sample Construction

CSI amplitude streams were continuously recorded under the four deployment cases in Figure 5 with controlled occupancy sessions. Following the edge feature extraction pipeline described in the previous section, each fixed-length window was converted into one compact CSI feature matrix $\mathbf{X} \in \mathbb{R}^{56 \times 8}$, where the 56 rows correspond to the selected valid subcarriers and the eight columns correspond to the statistical descriptors computed from the within-window amplitude sequence. Each window was treated as one training sample, and its ground-truth occupancy label (0P–5P+) was assigned according to the number of participants present in the monitored area during that window.

To reduce label ambiguity while preserving realistic office dynamics, participants were instructed to maintain natural but relatively stable behaviors within the designated area (e.g., sitting, slight posture changes, occasional standing), rather than performing deliberate large movements.

5.2.2. Dataset Overview, Split Protocol, and Leakage Prevention

Overall, the resulting dataset contains 6997 window-level samples with an imbalanced class distribution typical of practical office recordings (0P: 1509; 1P: 1993; 2P: 602; 3P: 582; 4P: 985; 5P+: 1326). For model development, the dataset was split into training, validation, and test subsets using a stratified protocol to preserve the class distribution across splits (Train: 4897; Val: 1049; Test: 1051). All normalization statistics for feature scaling (mean and standard deviation) were computed only on the training set and then applied to the validation and test sets to prevent information leakage. This protocol ensures that reported results reflect true generalization rather than benefiting from test-set statistics.

5.2.3. Imbalance Handling

To address class imbalance while keeping the evaluation unbiased, Synthetic Minority Oversampling Technique (SMOTE)-based oversampling was applied only to the training set in the feature space after standardization using training-set statistics. The validation and test sets were kept unchanged, ensuring that reported performance reflects generalization under realistic multipath conditions and class priors rather than benefiting from synthetic samples during evaluation.

5.2.4. Model Training Environment and Optimization

Model training was implemented using PyTorch (v2.5.1) with CUDA acceleration and executed on an NVIDIA GeForce RTX 3060 GPU (San Tomas Expressway, Santa Clara, CA, USA). We trained the Mini Transformer described in the previous section using a fixed training protocol across experiments to ensure fair comparisons. Unless otherwise specified in the ablation study, the optimizer was Adam with learning rate 10^{-4} and weight decay 5×10^{-4} . Early stopping was applied based on validation performance to avoid over-training.

5.3. Ablation Study for Training Strategy Selection

To identify which training strategies most consistently improve validation performance, we conducted an ablation study starting from the baseline Mini Transformer and incrementally introducing commonly used components. This analysis also serves to isolate the effect of model capacity from data-level and loss-level factors. All ablation settings followed the same split protocol and differed only in the applied training strategy. The compared configurations include: (i) baseline without enhancement, (ii) baseline + SMOTE (training set only), (iii) baseline + class-specific augmentation, (iv) baseline + data augmentation, (v) baseline + focal loss, (vi) baseline + mixup, (vii) baseline + regularization, (viii) a larger-capacity model, and (ix) an all-components setting with the larger model. For each setting, we monitored validation accuracy and summarized the results using Accuracy, Macro F1, and the overfitting gap (defined as the training–validation accuracy difference), as reported in Table 2.

The ablation results indicate that balancing and augmentation influence performance in different ways. SMOTE yields the most consistent gain in both Accuracy and Macro F1, confirming that data-level imbalance is a dominant factor limiting generalization. In contrast, stronger regularization substantially reduces the training–validation gap but may underfit, leading to a noticeable drop in Accuracy and Macro F1. Data augmentation reduces the overfitting gap as well, but its net effect on Accuracy is smaller than SMOTE in this dataset.

Table 2. Ablation study results under the final training configuration. Test Accuracy and Macro F1 are evaluated on the held-out test set. Gap denotes the training–validation accuracy difference estimated from validation curves.

Configuration	Test Acc.	Macro F1	Overfit Gap
Baseline	0.9211	0.9135	0.067
Larger Model	0.9169	0.9037	0.064
Data Augmentation	0.9245	0.9163	0.046
SMOTE	0.9312	0.9250	0.062
Focal Loss	0.9153	0.9046	0.064
Class-Specific Aug.	0.9270	0.9181	0.055
Regularization	0.8901	0.8739	0.013
Mixup	0.9211	0.9105	0.061
All Components (632 K)	0.8800	0.8685	0.047

Bold values indicate the best performance among all configurations.

5.3.1. Effect of Activation Function

To further investigate the effect of architectural choices on model performance, we conducted a controlled comparison of four activation functions applied consistently across both the feature projection layer and the feed-forward network: ReLU, Leaky ReLU, ELU, and GELU. All other settings were held fixed, including the SMOTE training strategy identified above, the dataset split, optimizer configuration, training budget (100 epochs), and random seeds.

The results are summarized in Table 3. ELU achieves the highest test accuracy (98.86%) and the smallest overfitting gap (0.0132) among all four candidates. Although Leaky ReLU attains a slightly higher best validation accuracy, ELU demonstrates superior generalization on the held-out test set while maintaining the lowest overfitting gap. ReLU achieves competitive accuracy (98.67%) but shows a slightly larger overfitting gap. Leaky ReLU and GELU both achieve 98.19% test accuracy, with GELU exhibiting the largest overfitting gap (0.0203).

Table 3. Effect of activation function on Mini Transformer performance. All configurations use the SMOTE training strategy and identical hyperparameters. Overfitting gap is defined as the difference between final training accuracy and final validation accuracy.

Activation	Test Acc.	Best Val Acc.	Overfit Gap
ReLU	0.9867	0.9800	0.0157
Leaky ReLU	0.9819	0.9828	0.0149
ELU	0.9886	0.9790	0.0132
GELU	0.9819	0.9800	0.0203

Bold values indicate the best performance among all configurations.

The advantage of ELU can be attributed to its smooth negative saturation behavior. Unlike ReLU, which produces a hard zero for all negative pre-activations, ELU preserves small negative outputs, which tends to push mean activations closer to zero, acting as a form of implicit self-normalization. This property is particularly beneficial in compact Transformer architectures where batch normalization is not applied. GELU, despite its smoother gradient profile, does not provide a clear advantage in the compact 56×8 feature space of this study, and its larger overfitting gap suggests reduced stability under the current training budget. Based on these results, ELU is adopted as the default activation function in the final model configuration.

5.3.2. Model Size Selection and Generalization Performance

The baseline Mini Transformer model contains approximately 138 K trainable parameters, while the larger variant contains 632 K parameters (about $4.6\times$ more). As indicated by the ablation results in Table 2, increasing model size does not improve generalization in our setting. Instead, the larger model exhibits lower Accuracy and Macro F1 than the baseline, together with a comparable or larger training–validation gap, indicating a higher tendency to overfit. These results support selecting the compact baseline architecture as a better accuracy–complexity trade-off.

Although inference is performed on the cloud in this study, lightweightness remains practically important. A smaller model improves throughput and reduces per-request computation and memory footprint under concurrent multi-device deployments, which aligns with the scalability requirements in SMOs.

5.4. Evaluation Metrics and Result Summary

Model performance was evaluated on the held-out test set using overall Accuracy and class-wise Precision, Recall, and F1-score for the six occupancy classes (0P–5P+). In addition to scalar metrics, we report a confusion matrix to visualize class-wise discrimination and typical error patterns, and training/validation accuracy curves to illustrate convergence behavior under the adopted training configuration, as shown in Figure 6.

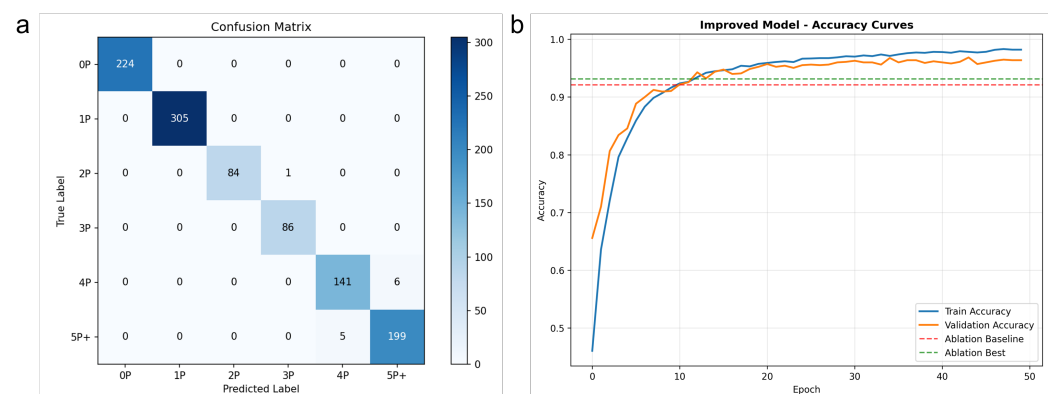


Figure 6. Performance of the Mini transformer model: (a) confusion matrix on the held-out test set; (b) training and validation accuracy curves under the adopted training configuration.

As shown in Figure 6a, the proposed model achieves an overall test accuracy of 1030/1051 (98.86%) across the six occupancy levels. The classification of the low-occupancy regimes is particularly reliable: all samples of 0P and 1P are correctly recognized (224/224 and 305/305, respectively), which is important for practical SMO operation where stable detection of vacancy and single-person presence directly supports downstream control logic.

For medium-to-high occupancy levels, the remaining misclassifications are limited and primarily occur between neighboring classes. For example, “Over 5 People” is occasionally underestimated as 4P, while 4P and 5P+ show minimal bidirectional confusion. Based on the confusion matrix, the per-class recalls are 100.0% (0P), 100.0% (1P), 98.8% (2P), 100.0% (3P), 95.9% (4P), and 97.5% (5P+). The slightly lower recalls for 4P and 5P+ are consistent with boundary ambiguity at higher crowd densities, where small spatial variations can produce similar CSI amplitude patterns. Overall, the confusion structure indicates that most errors occur between adjacent occupancy levels, suggesting that the model captures the ordinal progression of crowd density rather than exhibiting random misclassification behavior.

Figure 6b further confirms stable optimization behavior under the adopted training configuration. Validation accuracy rapidly increases and stays consistently high after the

early epochs, while the training–validation gap remains small, implying that the model generalizes well to unseen samples under the held-out evaluation protocol.

To further illustrate the system-level efficiency gain over the DenseNet121-based configuration used in our previous work [17], Figure 7 compares the output data volume and model parameter size between the two frameworks on a logarithmic scale. As shown in Figure 7(left), replacing high-dimensional time–frequency spectrograms with compact 56×8 statistical CSI feature matrices reduces the per-window data footprint from approximately 128 MB to about 0.02 MB, corresponding to more than a $5000\times$ reduction in transmission volume while preserving key inter-subcarrier dependencies required for occupancy estimation.

In parallel, Figure 7(right) shows that the model size is reduced from approximately 8.0 M parameters in DenseNet121 to 138 K parameters in the proposed Mini Transformer, achieving a $\sim 58\times$ reduction. This substantial decrease in both data volume and model complexity directly translates into lower communication overhead and faster inference, which is critical for scalable edge-cloud deployments in SMOs. Together, these results confirm that the proposed Mini Transformer framework effectively addresses the computational and communication bottlenecks inherent in DenseNet-based end-to-end solutions, while maintaining comparable classification accuracy.

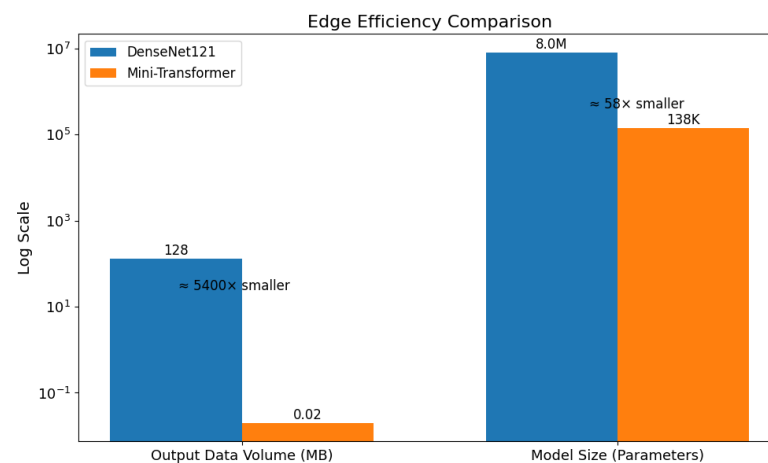


Figure 7. Edge efficiency comparison between the DenseNet121-based framework and the proposed Mini Transformer. **(Left)** Output data volume per window (log scale). **(Right)** Model parameter size (log scale).

6. Discussion

The proposed framework achieves substantial reductions in both communication overhead and model complexity without compromising occupancy classification accuracy. Replacing high-dimensional time–frequency spectrograms with compact 56×8 statistical feature matrices eliminates the MQTT bandwidth bottleneck that renders image-based pipelines unsuitable for on-device deployment in SMOs. At the system level, the dual-core implementation resolves a hardware conflict that algorithm-focused studies tend to overlook: CSI acquisition requires a dedicated Wi-Fi channel that cannot coexist with simultaneous network communication on a single-core device. Assigning CSI preprocessing to Core 0 and network operations to Core 1 under FreeRTOS ensures deterministic timing for both tasks and supports unattended long-term operation, a fundamental requirement in SMOs without dedicated maintenance personnel. At the application level, the integration of reliable occupancy estimates with the BI-Tech behavioral intervention platform closes the sensing-to-feedback loop that prior BI-Tech deployments lacked, enabling occupancy-

aware energy-saving reminders grounded in accurate multi-level counting rather than coarse presence detection.

The ablation results clarify what drives model performance in this feature space. Lightweight attention-based architectures benefit more from data-level balancing via SMOTE than from increasing model capacity or strengthening regularization. Both over-parameterization and aggressive weight decay degraded generalization on the held-out test set, indicating that model size should be matched to the dimensionality of the input feature space rather than scaled up by default.

6.1. Comparison with Alternative Sensing Modalities

To situate the proposed approach within the broader occupancy sensing landscape, Table 4 compares the major sensing modalities against the practical requirements of SMO deployment. Vision-based systems offer high counting accuracy but introduce privacy concerns that are difficult to reconcile with occupant acceptance in shared offices, and they require structured cabling, ceiling-level mounting, and ongoing maintenance [3]. PIR sensors avoid these concerns but are limited to binary presence detection and cannot provide the per-person granularity required for proportional HVAC control [5]. CO₂-based inference is attractive for its dual role in air quality monitoring, but diffusion dynamics in ventilated spaces introduce response delays of several minutes that make real-time count estimation unreliable [6]. Device-assisted approaches such as badge tracking and Wi-Fi probe monitoring require active user cooperation or probabilistic MAC address inference, both of which introduce systematic biases [4].

Table 4. Comparison of occupancy sensing modalities against SMO deployment requirements.

Modality	Count	Privacy	Cost	Maintenance	Device-Free
Camera-based	High	Low	High	High	✓
PIR sensor	Low	High	Low	Low	✓
CO ₂ sensor	Medium	High	Medium	Low	✓
Badge/Smartphone	Medium	Low	Medium	Medium	×
CSI sensing	High	High	Low	Low	✓

Bold entries indicate the sensing method employed in the proposed system (CSI sensing).

As shown in Table 4, CSI-based sensing covers a gap that no single existing modality addresses: it provides multi-level count estimation, requires no body-worn devices, raises no visual privacy concerns, and operates on commodity Wi-Fi hardware that many SMOs already possess. The proposed framework inherits these properties while resolving the computational feasibility barrier that has prevented CSI methods from running on microcontroller-class edge devices. We acknowledge that the comparison in Table 4 is qualitative; a rigorous quantitative benchmark across modalities under identical experimental conditions would require a dedicated multi-sensor deployment study that accounts for room geometry, occupant behavior patterns, and ventilation conditions simultaneously, and is identified as an important direction for future work.

6.2. Comparison with TinyML-Based Edge Occupancy Sensing Approaches

Table 5 compares three representative edge-deployed human sensing systems against the proposed approach. Armenta-Garcia et al. [19] achieved 92.43% accuracy for Wi-Fi CSI-based HAR on the ESP32-S3 using a quantised DenseNet with a 127 kB footprint, demonstrating that CSI inference is feasible on microcontroller-class hardware. However, HAR classifies the activity type of a single subject, whereas multi-level occupancy counting requires aggregating subtle channel responses from multiple static occupants across subcarriers, which represents a distinct inferential setting. Mach et al. [30] reported 99.38%

accuracy for UWB radar-based people counting on STM32 microcontrollers, but the system relies on dedicated radar hardware rather than commodity Wi-Fi infrastructure, and the evaluation involved freely walking participants rather than the seated-office conditions studied here. Pandkar et al. [31] combined CO₂, temperature, humidity, illuminance, and PIR sensors with a Random Forest model on ESP32 devices, achieving $R^2 = 0.923$, but the 1.426 MB model footprint is substantially larger than that of the proposed Mini Transformer, and CO₂-based inference introduces the diffusion delays discussed above.

Table 5. Comparison of TinyML-based edge-deployed human sensing systems.

Method	Sensing	Platform	Task	Model Footprint	Performance	Output
Armenta-Garcia et al. [19]	Wi-Fi CSI	ESP32-S3	HAR (5-class)	~127 kB	92.43%	Activity label
Mach et al. [30]	UWB Radar	STM32	Counting (0–10)	525.8 kB *	99.38% †	Counting
Pandkar et al. [31]	CO ₂ /Temp/PIR	ESP32	Counting (0–5)	1.426 MB	$R^2 = 0.923$	Counting
This Study	Wi-Fi CSI + Multi-Modal Environmental Sensing ‡	ESP32-S3	Counting (0–5+)	138 K params (~441 kB FP32)	98.86%	Counting + Energy-saving reminders

* Optimized ResNet-based CNN with $16\times$ parameter reduction. † 98.22% after 8-bit quantisation. ‡ Environmental signals include CO₂, PM_{2.5}/PM₁₀, temperature, humidity, illuminance, and power consumption.

To the best of our knowledge, few existing works simultaneously combine device-free operation on commodity Wi-Fi CSI, multi-level occupancy counting under realistic seated-office conditions, and microcontroller-compatible model complexity. The proposed framework is designed to satisfy all three constraints jointly.

6.3. Robustness Considerations and Limitations

The 98.86% test accuracy was obtained under controlled conditions and does not fully represent the range of environments encountered in operational SMOs. Several deployment scenarios warrant discussion. Environments with frequent entry and exit events introduce transient CSI disturbances that may affect one or two consecutive 6-second windows. Because the model is trained on steady-state occupancy patterns, these transition windows may be temporarily misclassified. However, because the BI-Tech energy management strategy operates on occupancy session timescales rather than instantaneous presence events, isolated transition-level misclassifications are unlikely to affect the reliability of session-level decisions.

Two representative use cases illustrate this design. For lighting waste detection, sustained illuminance above threshold combined with persistently low occupancy estimates across consecutive 5-min cycles during evening hours triggers a reminder delivered the following morning. For HVAC waste detection, extended system operation during unoccupied periods, such as overnight or across lunch breaks, is identified through aggregated occupancy records and translated into behavioral feedback. These scenarios operate on timescales of tens of minutes to hours, and the 5-min sensing cycle therefore provides sufficient temporal resolution for effective intervention. Prior BI-Tech field deployments [22] have shown that session-level feedback leads to measurable reductions in unnecessary energy use through sustained behavioral adaptation. Accordingly, the present framework focuses on feedback-oriented energy management rather than instantaneous closed-loop actuation. Real-time device-level switching, such as PIR-triggered light-off control, represents a different control paradigm requiring separate architectural and comfort considerations, and is identified as a complementary direction for future integration.

Static human-shaped objects such as mannequins or coats left on chairs modify the multipath background but do not produce time-varying channel fluctuations. The discriminative features used here, including per-subcarrier variance (σ^2), rate-of-change (ΔA), and zero-crossing rate (Z), respond to temporal dynamics rather than static channel structure.

Real occupants generate persistent micro-movements such as postural shifts and respiration that activate these features continuously, whereas static objects do not. The four-stage denoising pipeline further suppresses fixed background contributions to the feature distribution. The same reasoning applies to furniture relocation: repositioned objects alter the static multipath background but not its temporal fluctuations, so the system's reliance on temporal statistics provides inherent robustness to layout changes. Substantial rearrangements involving large reflective surfaces may nevertheless require a brief recalibration to realign the feature distribution.

Moving non-human objects such as equipment trolleys or pushed chairs introduce non-stationary channel perturbations not covered by the current training data. Human body movement tends to produce distributed fluctuations spanning many subcarriers simultaneously due to the large and irregular reflecting surface of the body, while moving objects typically affect a smaller and more spatially coherent subset of subcarriers. The cross-subcarrier attention mechanism in the Mini Transformer is designed to exploit these distributional differences, but whether this cross-subcarrier mechanism provides sufficient discrimination under realistic object movement conditions remains experimentally unverified. Moving metallic objects pose a particular challenge: their high conductivity produces stronger and more spatially concentrated reflections whose instantaneous amplitude on affected subcarriers can exceed that produced by human body movement. Whether the attention mechanism can reliably separate these two effects under heavy metallic interference remains an open question and is identified as a high-priority validation target in future work.

Regarding antenna configuration, the current implementation uses omnidirectional antennas on both nodes to support spatial diversity and flexible device placement. Directional antennas could improve signal-to-noise ratio along specific propagation paths but would require precise orientation during installation and may reduce the multipath diversity that the statistical feature extraction relies on. A systematic comparison under controlled occupancy conditions is identified for future work. The current evaluation was conducted in a single office room with a fixed floor area and rectangular geometry. Room geometry directly affects multipath propagation structure: rooms with irregular shapes, open-plan layouts, or significantly different dimensions may produce CSI feature distributions that differ systematically from those observed in the training environment. While the four deployment cases in this study cover distinct Tx/Rx propagation paths and partially simulate geometric diversity at the link level, cross-room generalization has not been validated. Deploying the framework in a new room geometry would likely require recollecting a modest amount of labeled data to recalibrate the feature distribution, consistent with the periodic recalibration strategy already discussed for long-term environmental drift.

More broadly, CSI-based features are sensitive to long-term environmental drift, device variability, and room layout changes, any of which can shift the feature distribution relative to the training data without producing an obvious system failure. Periodic recalibration partially addresses this but places a maintenance burden on SMO operators. Online adaptation mechanisms, such as incremental learning on recently collected samples or environment-aware feature normalization, represent a more durable solution and are not yet supported in the current system. The multimodal data already collected by the BI-Tech node, including CO₂ concentration and illuminance measurements, provides a complementary basis for cross-validating occupancy estimates and detecting anomalous divergence from physically plausible values.

This study does not include long-term field validation of the proposed occupancy-aware control strategies. Prior BI-Tech work [22] demonstrated that behavioral reminders tied to real-time occupancy data can yield measurable energy reductions over multi-month

deployments; however, those results were obtained without CSI-based occupancy counting. Integrating the human counting framework into a longitudinal BI-Tech deployment would allow the energy-saving contribution of accurate occupancy estimation to be evaluated directly, and this remains the principal next step for the research program.

7. Conclusions

This study presents a deployable edge-AI occupancy-sensing framework for SMOs, addressing the system-level constraints that have limited the practical adoption of CSI-based methods. The primary contributions include a dual-core ESP32-S3 implementation that separates CSI preprocessing from network communication, a four-stage signal denoising pipeline that produces stable statistical features under multipath conditions, and a lightweight Mini Transformer that classifies six occupancy levels from a compact 56×8 feature matrix with 98.86% test accuracy. At the data level, the edge-side feature extraction pipeline reduces per-window transmission volume from 225 kB of raw CSI to 1.75 kB of statistical features, a $129\times$ reduction relative to raw CSI transmission that eliminates the MQTT bandwidth bottleneck. Compared with the DenseNet121 baseline, replacing high-dimensional spectrogram inputs with the compact 56×8 representation achieves a further reduction exceeding $5000\times$ in per-window input data volume, alongside an approximately $58\times$ reduction in model parameter count, while maintaining competitive classification performance.

The ablation analysis showed that data-level balancing via SMOTE and the adoption of ELU activation contribute more to generalization than increasing model capacity or strengthening regularization, reinforcing the principle that model architecture should be matched to the dimensionality of the feature space. The framework integrates with the BI-Tech behavioral intervention platform, providing the reliable occupancy estimates required to generate occupancy-aware energy-saving reminders in SMO environments.

Several directions remain for future work. Controlled validation involving non-human moving objects, metallic reflectors, and directional antenna configurations is needed to characterize deployment boundaries more precisely. Online adaptation mechanisms to handle long-term environmental drift should be incorporated into the system. Integrating the human counting module into a longitudinal BI-Tech field deployment would allow the energy-saving contribution of accurate occupancy estimation to be evaluated directly under real-world operating conditions.

Author Contributions: Conceptualization, Y.C.; methodology, Y.C.; software, Y.C. and X.W.; validation, Y.C.; formal analysis, Y.C.; investigation, Y.C.; resources, D.S.; data curation, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, Y.C., D.S., X.W., T.Y., T.U., and J.O.; visualization, Y.C.; supervision, D.S., T.Y., T.U., and J.O.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2132, and JST SPRING, Japan Grant Number JPMJSP2136.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

- Chen, S.; Zhang, G.; Xia, X.; Chen, Y.; Setunge, S.; Shi, L. The impacts of occupant behavior on building energy consumption: A review. *Sustain. Energy Technol. Assess.* **2021**, *45*, 101212. [\[CrossRef\]](#)
- Andersen, R.V.; Olesen, B.W.; Toftum, J. Simulation of the effects of occupant behaviour on indoor climate and energy consumption. In *Proceedings of CLIMA 2007: 9th REHVA World Congress: WellBeing Indoors*; REHVA: Helsinki, Finland, 2007.
- Udrea, I.; Alionte, C.G.; Ionaşcu, G.; Apostolescu, T.C. New research on People Counting and Human Detection. In *Proceedings of the 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Pitesti, Romania, 1–3 July 2021; pp. 1–6. [\[CrossRef\]](#)
- Uras, M.; Ferrara, E.; Cossu, R.; Liotta, A.; Atzori, L. MAC address de-randomization for WiFi device counting: Combining temporal- and content-based fingerprints. *Comput. Netw.* **2022**, *218*, 109393. [\[CrossRef\]](#)
- Bouazizi, M.; Ohtsuki, T. An Infrared Array Sensor-Based Method for Localizing and Counting People for Health Care and Monitoring. In *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 20–24 July 2020; pp. 4151–4155. [\[CrossRef\]](#)
- Jiang, C.; Masood, M.K.; Soh, Y.C.; Li, H. Indoor occupancy estimation from carbon dioxide concentration. *Energy Build.* **2016**, *131*, 132–141. [\[CrossRef\]](#)
- Ma, Y.; Zhou, G.; Wang, S. WiFi sensing with channel state information: A survey. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 46. [\[CrossRef\]](#)
- Yan, B.; Li, Y.; Dong, L.; Ren, Z.; Liu, H.; Gao, X.; Cheng, W. Crowd counting with WiFi sensing based on iterative attentional feature fusion. *Comput. Commun.* **2025**, *241*, 108245. [\[CrossRef\]](#)
- Zhang, L.; Zhang, Y.; Wang, B.; Zheng, X.; Yang, L. WiCrowd: Counting the Directional Crowd with a Single Wireless Link. *IEEE Internet Things J.* **2021**, *8*, 8644–8656. [\[CrossRef\]](#)
- Khan, D.; Ho, I.W.H. CrossCount: Efficient Device-Free Crowd Counting by Leveraging Transfer Learning. *IEEE Internet Things J.* **2023**, *10*, 4049–4058. [\[CrossRef\]](#)
- Yang, J.; Zou, H.; Jiang, H.; Xie, L. Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes. *IEEE Internet Things J.* **2018**, *5*, 3991–4002. [\[CrossRef\]](#)
- Cianca, E.; De Sanctis, M.; Di Domenico, S. Radios as Sensors. *IEEE Internet Things J.* **2017**, *4*, 363–373. [\[CrossRef\]](#)
- Zou, H.; Zhou, Y.; Yang, J.; Spanos, C.J. Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT. *Energy Build.* **2018**, *174*, 309–322. [\[CrossRef\]](#)
- Natarajan, A.; Krishnasamy, V.; Singh, M. Design of a Low-Cost and Device-Free Human Activity Recognition Model for Smart LED Lighting Control. *IEEE Internet Things J.* **2024**, *11*, 5558–5567. [\[CrossRef\]](#)
- He, Q.; Sumiyoshi, D. Investigating the Relationship Between Environmental Awareness and Energy-Saving Behavior in Office Buildings Using the Theory of Planned Behavior. *J. Environ. Eng. (Trans. AIJ)* **2025**, *90*, 1–12. [\[CrossRef\]](#)
- Shahbazian, R.; Trubitsyna, I. Human Sensing by Using Radio Frequency Signals: A Survey on Occupancy and Activity Detection. *IEEE Access* **2023**, *11*, 40878–40904. [\[CrossRef\]](#)
- Chen, Y.; Sumiyoshi, D.; Pan, Y.; Yamamoto, T.; Ueno, T.; Oh, J. A Low-Cost IoT Framework for Device-Free Human Counting and Lighting Energy Control in Commercial Spaces. In *Proceedings of the 2025 7th International Conference on Computer Communication and the Internet (ICCCI)*, Tokushima, Japan, 27–29 June 2025; pp. 50–56. [\[CrossRef\]](#)
- Alajlan, N.N.; Ibrahim, D.M. TinyML: Enabling of Inference Deep Learning Models on Ultra-Low-Power IoT Edge Devices for AI Applications. *Micromachines* **2022**, *13*, 851. [\[CrossRef\]](#)
- Armenta-Garcia, J.A.; Gonzalez-Navarro, F.F.; Caro-Gutierrez, J.; Garcia-Reyes, C.I. Tools and Methods for Achieving Wi-Fi Sensing in Embedded Devices. *Sensors* **2025**, *25*, 6220. [\[CrossRef\]](#)
- Oliveira, H.S.; Oliveira, H.P. Transformers for Energy Forecast. *Sensors* **2023**, *23*, 6840. [\[CrossRef\]](#)
- Wang, Z.C.; Li, D.; Liu, J.; Zhao, Y. A modified transformer and adapter-based transfer learning for fault detection and diagnosis in HVAC systems. *Energy Storage Sav.* **2024**, *3*, 110–121. [\[CrossRef\]](#)
- Chen, Y.; Sumiyoshi, D.; Morita, Y.; Ishibashi, S.; Wang, X.; Yamamoto, T.; Ueno, T.; Oh, J. BI-Tech: An IoT-Based Behavioral Intervention System for User-Driven Energy Optimization in Commercial Spaces. *IEEE Access* **2025**, *13*, 166853–166872. [\[CrossRef\]](#)
- Chaudhari, P.; Xiao, Y.; Cheng, M.M.C.; Li, T. Fundamentals, Algorithms, and Technologies of Occupancy Detection for Smart Buildings Using IoT Sensors. *Sensors* **2024**, *24*, 2123. [\[CrossRef\]](#)
- IEEE. *IEEE Standard for Information Technology—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 5: Enhancements for Higher Throughput (IEEE Std 802.11n-2009)*; IEEE: New York, NY, USA, 2009. Available online: <https://standards.ieee.org/ieee/802.11n/3952/> (accessed on 3 March 2026).
- Choi, H.; Fujimoto, M.; Matsui, T.; Misaki, S.; Yasumoto, K. Wi-CaL: WiFi Sensing and Machine Learning Based Device-Free Crowd Counting and Localization. *IEEE Access* **2022**, *10*, 24395–24410. [\[CrossRef\]](#)

26. Espressif Systems. *ESP32-S3 Wi-Fi Driver: Channel State Information (CSI); ESP-IDF Programming Guide*, v5.5.3; Espressif Systems: Shanghai, China, 2024.
27. Shen, L.H.; Hsiao, A.H.; Chu, F.Y.; Feng, K.T. Time-Selective RNN for Device-Free Multiroom Human Presence Detection Using WiFi CSI. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2505817. [[CrossRef](#)]
28. Chu, F.Y.; Chiu, C.J.; Hsiao, A.H.; Feng, K.T.; Tseng, P.H. WiFi CSI-Based Device-free Multi-room Presence Detection using Conditional Recurrent Network. In Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Helsinki, Finland, 25–28 April 2021; pp. 1–5. [[CrossRef](#)]
29. Lu, K.I.; Chiu, C.J.; Feng, K.T.; Tseng, P.H. Device-Free CSI-Based Wireless Localization for High Precision Drone Landing Applications. In Proceedings of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019; pp. 1–5. [[CrossRef](#)]
30. Mach, T.K.T.; Pham, C.T.; Le, M. UWB Impulse Radar for People Counting With Convolutional Neural Network on Microcontrollers. *IEEE Sens. J.* **2024**, *24*, 15643–15650. [[CrossRef](#)]
31. Pandkar, M.; Nambiar, S.; Sinha, A.; Sharma, P. Optimizing room occupancy estimation on the edge: A TinyML and sensor network approach. *Results Eng.* **2026**, *29*, 109107. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.