RIS-Assisted mmWave Radar for Robust Hand Gesture Recognition Using Lightweight CNNs

F. Morabet, A. Lazaro, *Senior, IEEE*, M. Lazaro, R. Villarino, and D. Girbau *Senior, IEEE*

***Abstract*—Accurate and robust hand gesture recognition is a key enabler of intuitive, contactless human–machine interaction in applications such as healthcare, smart environments, automotive systems, and wearable electronics. While radar-based methods offer significant advantages over vision- and contact-based systems, conventional approaches that rely on Doppler or phase information are often limited by multipath interference, environmental clutter, and user variability. This paper presents a novel amplitude-domain sensing framework that integrates a 24 GHz continuous-wave (CW) radar with modulated frequency-selective surfaces (FSS) to achieve reliable hand gesture recognition without the need for Doppler or phase processing. Each FSS panel introduces a distinct modulation pattern, allowing spatial encoding of gesture-induced occlusions as temporally varying changes in the radar return signal. Rather than directly tracking hand motion, the system infers gestures by identifying attenuation patterns in the modulated reflections, which are transformed into time–frequency spectrograms and classified using a lightweight convolutional neural network optimized for real-time embedded inference. The system was evaluated using eight dynamic hand gestures, including swiping, zigzag, and circular motions. The findings demonstrate a peak classification accuracy of 97% over an interaction range of 0.5 m to 1.5 m.**

***Index Terms*—Amplitude-based sensing, convolutional neural networks (CNNs), embedded radar systems, reconfigurable intelligent surface (RIS), frequency-selective surface (FSS), hand gesture recognition, human-machine interaction (HMI), millimeter-wave radar, modulated backscatter, real-time signal processing, 24 GHz radar.**

I. Introduction

HAND gesture recognition is a key enabler of intuitive, contactless interaction in smart environments, automotive systems, extended reality (XR), assistive technologies, and healthcare [1], [2]. These diverse and rapidly evolving applications require systems that are fast, unobtrusive, and resilient to both user variability and environmental interference.

Vision-based gesture recognition techniques, including RGB and depth cameras, provide accurate motion tracking but are sensitive to ambient lighting, susceptible to occlusion, and raise privacy concerns [3]. Infrared (IR)-based systems perform better in low-light conditions but often degrade under excessively bright ambient lighting [4]. Wearable sensing technologies, such as data gloves, inertial measurement units (IMUs), and surface electromyography (sEMG) sensors, offer high accuracy but require physical contact, user-specific calibration, and may reduce long-term user comfort [5]–[7].

Radar-based gesture recognition has emerged as a promising alternative due to its contactless operation, robustness to variations in lighting, and strong privacy protections [2], [8]. Among radar technologies, millimeter-wave (mmWave) and frequency-modulated continuous wave (FMCW) radars have demonstrated the ability to capture fine motion details through Doppler shifts, micro-Doppler features, or phase variations [9], [10]. However, such systems typically require complex hardware, including multiple antennas and beamforming units, and rely on computationally intensive processes such as range-Doppler maps (RDMs) and angle-of-arrival (AoA) estimation or phase analysis [10], [11]. Their performance can also degrade significantly in cluttered environments due to multipath reflections and inter-user gesture variability, posing serious challenges for robustness and generalization [11], [12].

In response to these limitations, recent studies have adopted machine learning-based methods that analyze radar-derived features such as spectrograms, RDMs, and time-frequency signatures to improve recognition performance [10], [13]–[15]. While data-driven methods improve classification accuracy, their reliance on multibranch architectures, temporal modeling, or attention mechanisms incurs significant computational overhead, limiting suitability for embedded platforms [14], [16]. Furthermore, some recent efforts employing frequency-shift keying (FSK) radar and spectrogram-based convolutional neural networks (CNNs) introduce distance-adaptive inference but still rely on micro-Doppler extraction and frequency switching, preserving a dependence on complex signal features [17]. In contrast, amplitude-domain sensing has the potential to simplify the hardware and reduce processing complexity. Despite this potential, it remains largely underexplored, with most of the existing work still relying on Doppler information or antenna arrays [11], [12], [16].

A Reconfigurable Intelligent Surface (RIS) consists of arrays of tunable elements (often metasurfaces) that can dynamically control the phase, amplitude, or polarization of incoming electromagnetic waves [18]. This makes RIS technology extremely versatile for applications such as active beamforming [19], enhancing wireless communication [20], [21], and enabling advanced intelligent sensing [22]–[24], particularly at mmWave frequencies [25].

In this study, a real-time hand gesture recognition prototype is introduced, combining a 24-GHz continuous-wave radar with an RIS. Unlike Doppler-dependent systems, the proposed design encodes gesture-induced occlusions through amplitude-domain modulation and processes them with a lightweight

Manuscript received month xx, 202x; accepted month xx, 202x. Date of publication month xx, 202x; date of current version month xx, 202x. This research was funded by the project PID2021-122399OB-I00 MICIU/AEI/10.13039/501100011033/FEDER, UE.

The authors are with the Electronics, Electrical and Automatics Engineering Department, Rovira and Virgili University, Tarragona, Spain (e-mail: farid.morabet@urv.cat, antonioramon.lazaro@urv.cat, marc.lazaro@urv.cat, ramon.villarino@urv.cat, david.girbau@urv.cat). The corresponding author is A. Lazaro (e-mail: antonioramon.lazaro@urv.cat).

CNN for efficient classification. This eliminates reliance on short-range micro-Doppler features and avoids the need for computationally demanding deep models, thereby offering a hardware-efficient and robust solution for next-generation human–machine interfaces.

The remainder of this paper is organized as follows: Section II details the system design and methodology, including signal processing, machine learning, and hardware implementation. Section III presents the experimental results and discussion, covering robustness, generalization, and comparative evaluation. Finally, Section IV summarizes the findings and outlines directions for future work.

II. System Design and Methodology

A. Gesture Recognition System

The gesture recognition system integrates a 24 GHz CW radar with an RIS composed of four modulated frequency-selective surface (FSS) panels, arranged symmetrically around a central interaction zone. Each of the four panels is controlled by a square-wave modulation signal of different frequency, denoted as f_{m1} through f_{m4}, are generated by a microcontroller (MC). These modulation frequencies spatially encode gestures by inducing amplitude variations in the backscattered signal at offsets corresponding to the FSS modulation frequencies [26]. As illustrated in Fig. 1, the radar is mounted vertically above the RIS platform, with the FSS panels evenly distributed to provide directional coverage without the need for active beam steering mechanisms. When a user performs a gesture within the interaction zone, their hand partially obstructs the radar's line of sight to one or more panels. This occlusion introduces variations in the time–frequency spectrogram of the reflected signal level. These amplitude modulations encode both the spatial trajectory and the temporal evolution of the gesture.

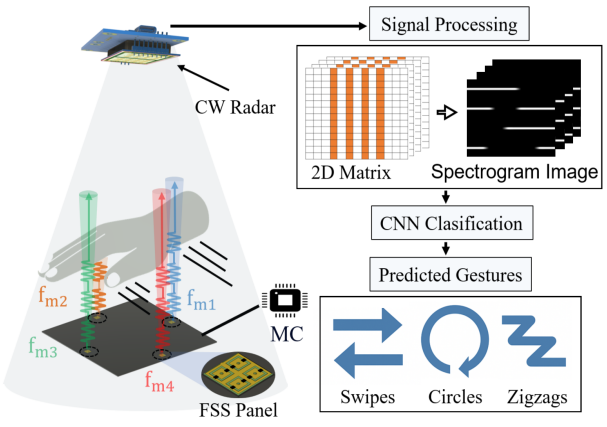


Fig. 1. Overview of the proposed radar–RIS gesture recognition system.

The radar signal is segmented to isolate gesture intervals. The resulting amplitude variations, converted into a time–frequency spectrogram, are mapped into a grayscale image, and input to a CNN for real-time classification. This architecture achieves accurate, contactless gesture recognition with low computational complexity, making it suitable for embedded human–machine interaction (HMI) applications.

B. Signal Processing and Classification Methodology

1) Radar–FSS Interaction and Signal Modulation: For a CW radar without RIS, the received signal can be expressed as the backscattered signal from the hand, the reflections from other objects or clutter, and the additive noise [27]:

$$r(t) = \alpha\, s(t-\tau) + s_c(t) + n(t), \tag{1}$$

where α is the attenuation that takes into account the reflection coefficient on the hand and the propagation losses, τ is the propagation delay, $s_c(t)$ is the clutter, and $n(t)$ is additive noise. This formulation yields only an attenuated and delayed version of the transmitted signal, and is therefore insufficient for gesture recognition, for which Doppler or phase characteristics are necessary to separate the signal from the clutter. Hand movements introduce variations in signal amplitude and propagation delay, resulting in micro-Doppler effects that enable the separation of the hand motion from static clutter. However, dynamic clutter caused by nearby moving objects (e.g., people in motion) can generate interference, making gesture detection more challenging.

The proposed system extends a frequency-modulated backscatter architecture [26]. For gesture recognition by leveraging dynamic hand occlusion each FSS panel operates as a reconfigurable reflector, toggling between reflective and non-reflective states under PIN-diode control.

The backscattered field at one of the FSS panels can be described using the FSS backscatter formulation [27]:

$$E_S(t) = E_{\text{structural}}(t) + \Gamma(t) E_{\text{incident}}(t), \tag{2}$$

where $E_{\text{incident}}(t)$ is the incident field, $E_{\text{structural}}(t)$ represents the static reflection and $\Gamma(t)$ denotes the PIN-diode-controlled reflection coefficient.

If each FSS is independently controlled and the switching rate of its PIN diodes is (f_m), then the reflection coefficient ($\Gamma(t)$) can be approximated by a square-wave function, assuming the switching time is much shorter than the modulation period—oscillating between two values, Γ_{OFF} and Γ_{ON} , corresponding to the two diode states. Since the time-varying ($\Gamma(t)$) is a periodic signal, it can be expressed as a Fourier series:

$$\Gamma(t) = \sum_{k=-\infty}^{k=+\infty} c_k e^{j2\pi k f_m t} \tag{3}$$

where c_k are the Fourier coefficients. The DC component C_0 represents the average reflection coefficient. The amplitude of these coefficients decreases rapidly with increasing harmonic order k. In practice, only the first-order components ($k = \pm 1$) are detectable, as higher-order terms fall below the noise floor as the distance to the radar increases.

This periodic switching imposes amplitude modulation on the incident continuous-wave (CW) signal, thereby producing frequency-shifted components at $f_c \pm f_{m_i}$, where f_{m_i} denotes the modulation frequency assigned to panel i ($i = 1, 2, 3, 4$). These sidebands act as frequency tags that uniquely identify the contribution of each panel. Due to this modulation, these contributions can be separated from the clutter, thereby facilitating their detection.

When RIS modulation is applied, the received signal generalizes to [26]:

$$r_{\text{RIS}}(t) = \sum_{i=1}^{N} \alpha_i \, s(t-\tau_i)\cos(2\pi f_{m_i} t + \Phi_i) + \alpha \, s(t-\tau) + s_c(t) + n(t), \tag{4}$$

where α_i and Φ_i represent the amplitude and phase of the first sideband associated with the (i^{th}) FSS panel. Under hand occlusion, the amplitude of the corresponding component is selectively attenuated, yielding gesture-dependent temporal signatures. Equation (1) is obtained as a special case of (4) when the modulation frequencies are canceled, i.e., $f_{m_i} = 0$.

As the user's hand obstructs different panels, the radar records variations through in-phase $I[n]$ and quadrature $Q[n]$ baseband signals, which are combined into a complex sequence

$$s[n] = I[n] + jQ[n]. \tag{5}$$

The spectral representation is then obtained via the fast Fourier transform (FFT):

$$Y[n,k] = \left|\text{FFT}_k\{s[n]\}\right|, \tag{6}$$

where $Y[n,k]$ represents the spectral magnitude at frequency bin k for frame n. Repetition of this process across frames produces a time–frequency representation that captures RIS-specific amplitude variations. This representation forms the input to the proposed learning framework, as illustrated in Fig. 2.

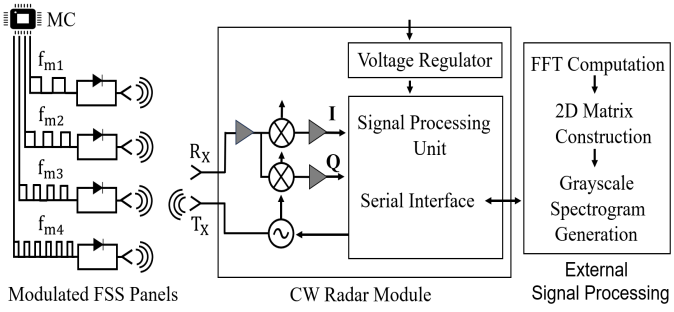


Fig. 2. Block diagram of the proposed radar–RIS gesture recognition system.

2) Machine Learning Model Architecture and Training: Conventional rule-based gesture recognition methods rely on hand-crafted features and fixed thresholds, which often fail to generalize across varying user behaviors and environmental conditions. To address these limitations, a lightweight CNN is employed to classify hand gestures directly from grayscale spectrograms. These spectrograms preserve key time–frequency features while minimizing memory usage, supporting efficient inference on embedded platforms.

The CNN architecture, illustrated in Fig. 3, consists of three convolutional blocks. Each block includes a 3×3 convolutional layer, batch normalization, ReLU activation, and 2×2 max pooling. The convolutional layers use 8, 16, and 32 filters, respectively, allowing the network to progressively expand its representational capacity while remaining compact. The output of the final convolutional block is flattened and passed through a fully connected layer with eight neurons, each representing a gesture class. A softmax layer then computes the probability distribution over all classes, and the predicted gesture corresponds to the class with the highest probability.

The classification head, including the flatten layer, fully connected layer, and softmax, is optimized for low-latency inference, enabling real-time deployment on resource-constrained hardware.

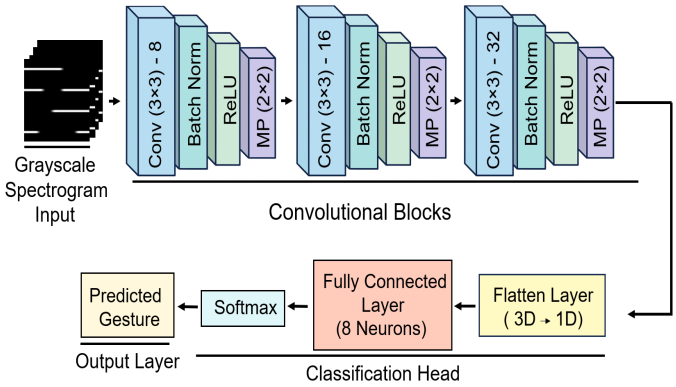


Fig. 3. Architecture of the CNN used for gesture classification.

The model is implemented in MATLAB and trained using stochastic gradient descent with momentum (SGDM) over 120 epochs. The initial learning rate is set to 0.001, with a mini-batch size of 8. The dataset is split into 80% training and 20% testing partitions. Cross-validation is used to assess model generalization across gesture classes and recording conditions.

C. Hardware Implementation

1) FSS Panel Design and Biasing: The FSS panel assembly is designed to enable reliable frequency-tagged modulation at 24 GHz, providing high reflection contrast and effective electromagnetic isolation. Each panel consists of a 2×2 array of periodic unit cells that operate at 24 GHz. The design follows the validated physical structure and biasing approach presented in [26]. In this work, the panels were fabricated on a Rogers RO4003C substrate [28] (thickness 0.51 mm, $\varepsilon_r = 3.45$, $\tan\delta = 0.0031$) with 17 µm copper metallization using a standard RF PCB process. As illustrated in Fig. 4(a), each unit cell consists of a square resonant loop with a PIN diode connected in the center. Aluminum Gallium Arsenide (AlGaAs) flip-chip PIN diodes from MACOM (model MADP-000907-14020x) [29] are used. The diodes were mounted onto the PCB using a standard reflow soldering process to ensure reliable electrical contact and mechanical stability. They exhibit low capacitance (C_{OFF} = 0.025 pF), low series resistance R_{ON} = 5.2 Ω, and fast switching speed (2 ns), making them ideal for this application. A square-wave control signal toggles the diode between high- and low-impedance states, enabling binary modulation of the radar cross-section. This configuration achieves a reflection coefficient difference greater than 15 dB over a 2.7 GHz bandwidth centered at 24 GHz (see Fig. 5 in [26]) by switching between the ON state (conductive, reflective surface) and the OFF state (non-conductive, resonant structure with suppressed reflection). The

complete 2×2 panel layout, including DC bias routing and two 1 kΩ resistors for improved RF isolation and switching stability, is shown in Fig. 4(b).

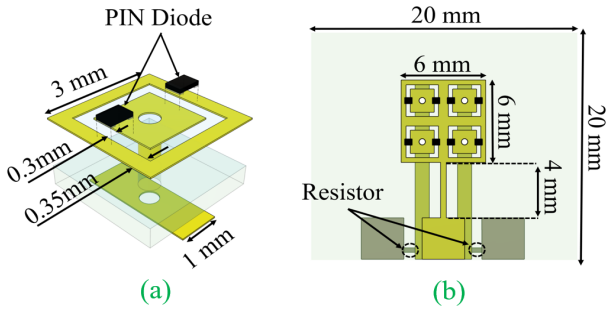


Fig. 4. (a) Exploded view of the unit cell showing PIN diode placement. (b) Layout of a 2×2 FSS panel with bias routing and resistor positions.

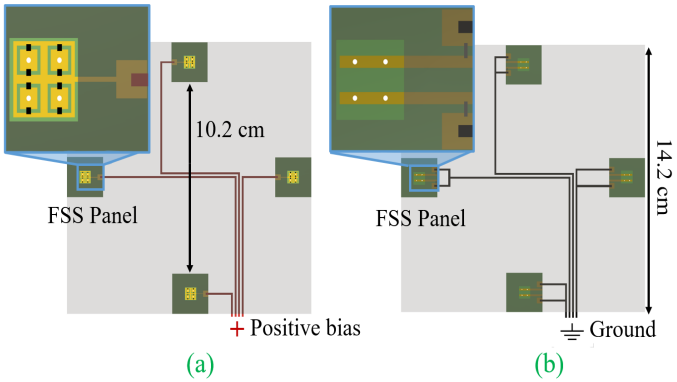


Fig. 5. (a) Top view of the FSS panel assembly showing spatial layout and positive bias routing. (b) Rear view illustrating ground return paths and diode placement.

The complete FSS panel assembly consists of four identical panels arranged in a 2×2 configuration, with diagonal spacing of approximately 10.2 cm and a total platform span of 14.2 cm. As illustrated in Fig. 5, the layout provides directional modulation coverage of the central interaction region. Positive supply lines and ground return paths are routed to each panel, with diode placement located on the rear side. A central ARM Cortex-M0 microcontroller controls each FSS by switching the bias lines with a square-wave signal at the desired modulation frequency.

2) Experimental Setup and Data Acquisition: The system is evaluated in a controlled indoor environment using a compact 24 GHz radar module (K-MD7-EVAL) [30], whose main debugging and configuration parameters are summarized in Table I. The radar is mounted vertically above the sensing region and aligned with the FSS panel assembly. Each panel in the FSS array is modulated using a square-wave signal of different frequencies generated from pulse-width modulation (PWM) outputs of an ARM Cortex-M0 ATSAM21G18A-MU microcontroller with a clock frequency of 48 MHz. The modulation frequencies, labeled f_{m1} to f_{m4}, are set to 2.00 kHz, 2.25 kHz, 2.50 kHz, and 2.75 kHz, respectively, with 250 Hz spacing to ensure spectral isolation. The modulation frequency can be selected within permitted limits, provided that the signal levels exceed typical environmental noise and hand-induced micro-motions while remaining below the radar's maximum Doppler bandwidth (∼8.9 kHz at 24 GHz). Harmonic interference is also mitigated by selecting non-overlapping frequencies, since square-wave modulation inherently generates sidebands at integer multiples of the fundamental frequency.

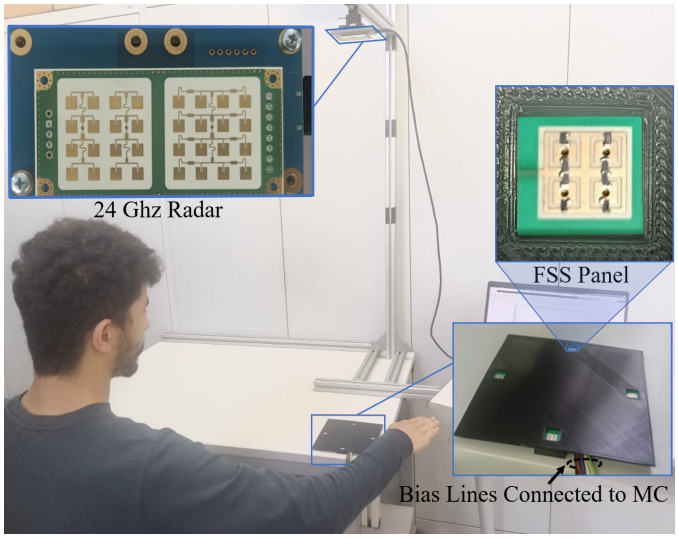


Fig. 6. Experimental configuration showing the 24 GHz radar module mounted above the sensing zone, FSS panel array, and MC-driven bias lines.

TABLE I
KEY SPECIFICATIONS OF THE K-MD7 RADAR MODULE.

Category	K-MD7 Value
Operating frequency	24.075–24.175 GHz
Field of view	TX: 30°× 30°, RX: 46°× 30°
Power consumption	55–105 mA (RMS) @ 3.3–5 V DC
Interface	UART up to 3 Mbps

System response is evaluated at five distances: 0.5 m, 0.75 m, 1.0 m, 1.25 m, and 1.5 m. These baseline measurements are performed without interrupting the line-of-sight (LoS) between the radar and the FSS panels. This ensures that the received signal is the one originated by the modulation of the FSSs. An FFT analysis is performed for each distance to extract the amplitude-modulated components associated with the modulation frequency of each panel. As illustrated in Fig. 7, the resulting spectra clearly show separated sidebands at the expected frequencies, confirming reliable frequency isolation over the entire tested range. Interference caused by nearby moving objects usually occurs at low velocity, so it mainly affects frequencies that are close to zero, which are well separated from the selected modulation frequencies. This is a key advantage of using modulated FSS compared to traditional Doppler radar systems, which can only detect moving objects and have difficulty in effectively filtering out non-stationary interference.

III. EXPERIMENTAL RESULTS AND DISCUSSION

1) Signal Preprocessing and Feature Extraction: Continuous radar acquisition generated baseband I/Q signals that

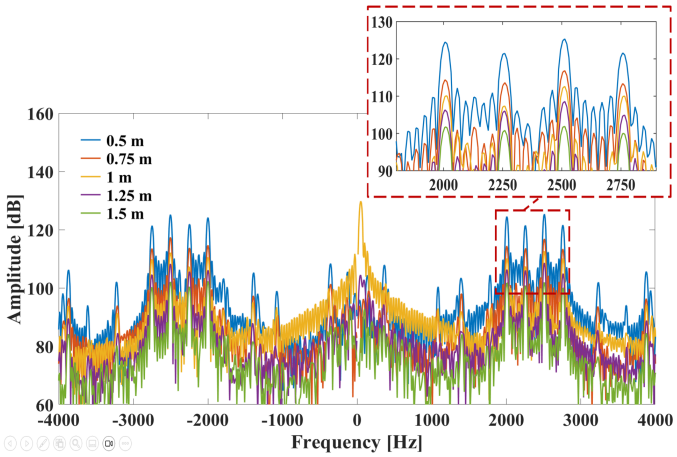


Fig. 7. Frequency spectra showing modulation sidebands at different radar-to-FSS distances from 0.5 m to 1.5 m.

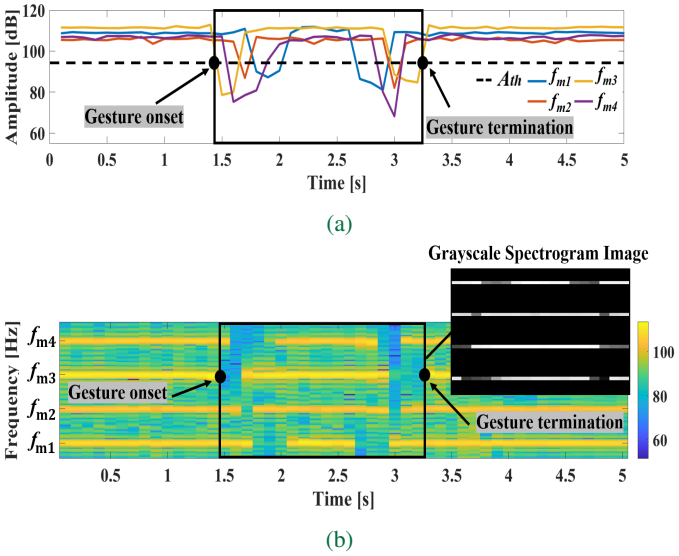


Fig. 8. Segmentation and preprocessing of radar signals. (a) Amplitude traces from four modulated FSS sidebands showing gesture onset and threshold-based expiration A_{th}. (b) Corresponding time–frequency spectrogram and final grayscale image used as CNN input.

included amplitude-modulated reflections from both dynamic gestures and static background elements. A multistage preprocessing strategy was used to isolate gesture-relevant activity. Segmentation was performed by monitoring the amplitude variations in the frequency-modulated sidebands associated with the four FSS panels (f_{m1} to f_{m4}). Gesture onset was detected when any sideband amplitude dropped below a threshold A_{th}, indicating partial hand occlusion. The completion of the gesture was established when all sidebands remained above A_{th} for at least two seconds. The threshold was set at approximately 5 dB below the minimum gesture-induced amplitude observed during calibration. All gestures were performed within 10 cm of the FSS array to ensure reliable encoding and consistent occlusion. This short-range configuration maximizes the probability of LoS interruption between the radar and one or more modulated panels, thereby inducing well-defined, gesture-dependent amplitude modulations in the backscattered signal. This process is illustrated in Fig. 8(a), where threshold crossings indicate gesture boundaries.

After segmentation, each gesture interval was converted into a time-frequency spectrogram using a sliding-window FFT, as shown in Fig. 8(b). The spectrogram captures the temporal evolution of amplitude components linked to the modulation frequency of each panel. Noise was reduced using dynamic range compression and exponential averaging along the time axis. Spectrograms were mapped to 8-bit grayscale (0–255) and resized to 256×414 pixels. These standardized grayscale images were then used as input to the CNN-based gesture classification model.

2) Evaluation of Robustness Under Diverse Conditions: System robustness was evaluated under three conditions: mechanical vibration, static FSS occlusion, and variation in gesture speed. These experiments were designed to assess the stability of frequency-tagged amplitude encoding in the presence of motion-induced interference, LoS disruption, and different users.

A mechanical disturbance was simulated by applying a low-frequency vibration to the radar mount and introducing an unrelated hand. As shown in Fig. 9, the resulting time–frequency spectrogram displays disturbance-induced energy centered near 0 Hz. These disturbance-induced components appeared outside the FSS modulation bands, which remained stable and isolated, demonstrating resilience against ambient motion.

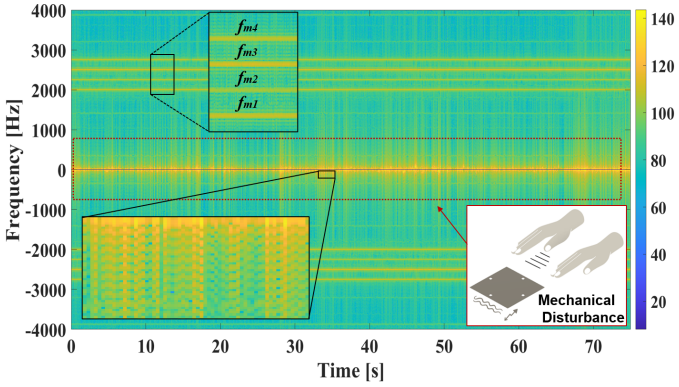


Fig. 9. Time–frequency spectrogram of radar backscatter under mechanical disturbance.

A second evaluation investigated the system's behavior under partial occlusion using common low-loss non-metallic materials such as a rigid board, a folded cloth, and a stack of books. Each time one of these elements obstructed one of the FSS panels, there was an attenuation of the corresponding modulated frequency, while the remaining panels preserved spatial information. The results are shown in Fig. 10, which presents the spectrograms alongside their corresponding grayscale images, illustrating the modulation patterns under each occlusion condition. Despite partial band degradation, the overall spectral structure remained intact, confirming resilience to localized obstructions.

Robustness to variation in gesture speed was evaluated using swipe gestures. In conventional CW radars, such gestures gen-

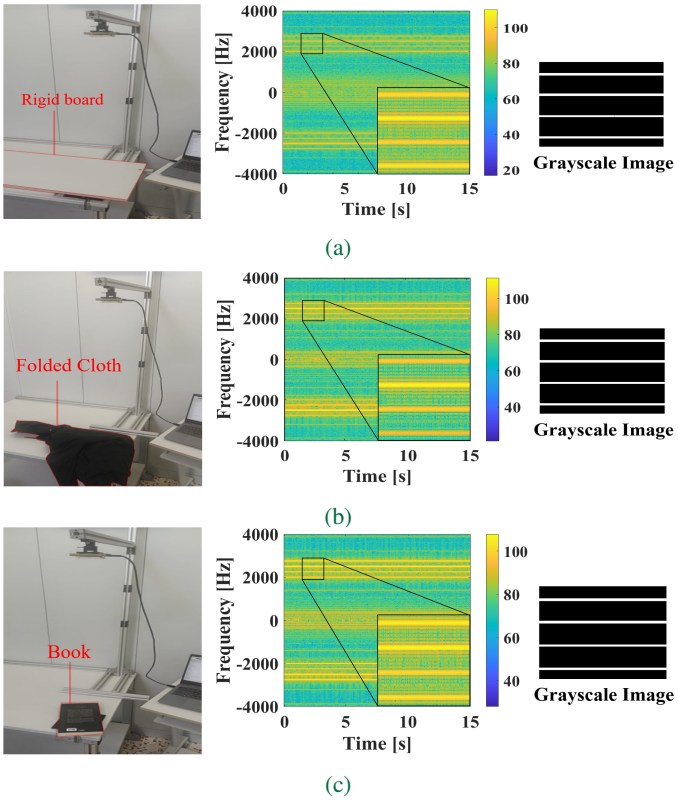


Fig. 10. Spectrogram and grayscale amplitude images under three static occlusion scenarios using low-loss, non-metallic materials: (a) Rigid board, (b) folded cloth, and (c) a single book placed in front of individual FSS panels in the radar's field of view.

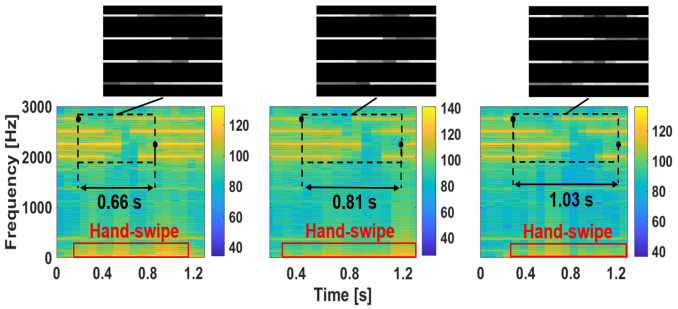


Fig. 11. Spectrograms of swipe gestures with different durations

erate low-frequency Doppler shifts near 0 Hz, which directly couples recognition accuracy to hand speed and increases susceptibility to clutter. The proposed RIS prototype avoids this problem. In this case, the hand is modeled as a passive occluder, which produces a temporally varying attenuation of the RIS-tagged sidebands. As shown in Fig. 11, the spectrograms of swipe gestures with durations of 0.66 s, 0.81 s, and 1.03 s confirm that, although the baseband content varies with hand speed, the RIS sideband signatures are preserved by the preprocessing stage described in Section III-1

3) Dataset Overview and Evaluation Protocol: Representative examples of processed data are shown in Fig. 13. Each grayscale image encodes the spatiotemporal amplitude dynamics across the four modulated FSS panels, capturing gesture-specific patterns that reflect differences in hand trajectory and occlusion timing. These visual distinctions support robust class separation across both gesture types and interaction distances. All samples were labeled with both gesture class and radar-to-panel distance. No subject-specific calibration or adaptation was applied during either data collection or model evaluation.

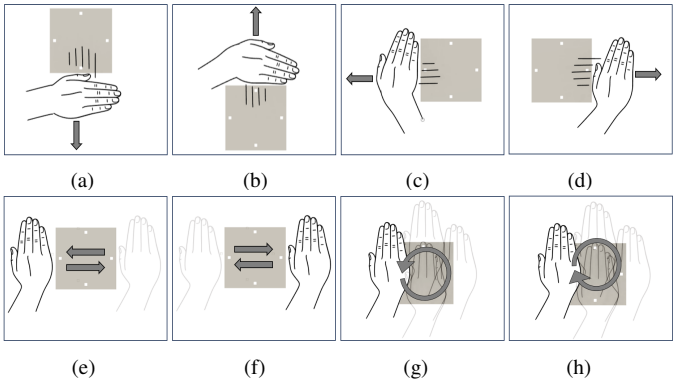


Fig. 12. Visual representation of the eight predefined right-hand gesture classes used in the dataset: (a) DW, (b) UP, (c) LF, (d) RT, (e) LRL, (f) RLR, (g) CW, and (h) CCW.

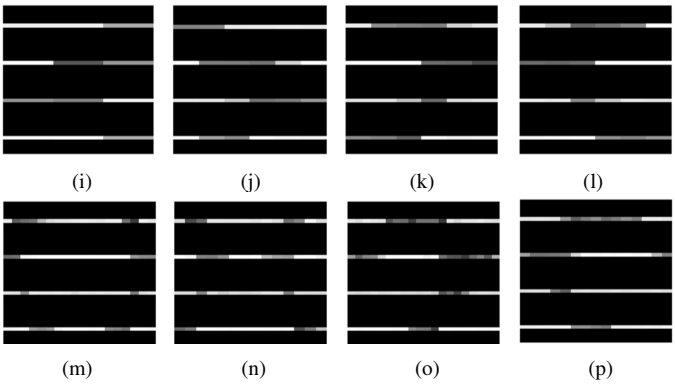


Fig. 13. Representative grayscale radar amplitude images corresponding to the eight gesture classes: (i) DW, (j) UP, (k) LF, (l) RT, (m) LRL, (n) RLR, (o) CW, and (p) CCW. Each image encodes spatiotemporal amplitude variations across the four modulated FSS panels during gesture execution.

4) CNN Performance at Fixed Distances: The spatial generalization capabilities of the radar–FSS gesture recognition system were evaluated using three CNN models, each independently trained and tested using collected data obtained from gestures performed at fixed radar-to-FSS distances of 0.5 m, 1.0 m, and 1.5 m. The learning behavior of these models is shown in Fig. 14(a), which presents the first 300 training iterations to emphasize early stage convergence. Each iteration represents a single weight update computed over a small batch of training samples. Among the three configurations, the model trained at 1.0 m exhibited the most stable and efficient learning dynamics, characterized by a rapid increase in accuracy and consistently lower training error. In comparison, the 0.5 m and 1.5 m models followed similar convergence trends but demonstrated slightly slower improvement and marginally higher error rates. These patterns align closely with the corresponding classification outcomes and reflect the 1.0 m model's superior ability to generalize across gesture instances.

This trend is further illustrated in the confusion matrices shown in Fig. 14(b)–(d), corresponding to the 0.5 m, 1.0 m, and 1.5 m configurations, respectively. The models achieved overall classification accuracies of 95.25%, 97.5%, and 94.75%. At 1 m, the model maintained high class-wise consistency, with all gesture classes exceeding 96% accuracy. Notably, DW, RT, LF, RLR, and UP reached 98%, reflecting the benefit of balanced radar coverage and distinct amplitude modulation patterns. The 0.5 m model also performed well, with several gestures (DW, RT, LF, CCW, and UP) achieving 98% accuracy. However, mirrored and rotational gestures showed reduced performance. RLR dropped to 86%, largely due to confusion with LF (8%), and to a lesser extent, CW and RT (2% each). CW also dropped to 92%, with 6% misclassified as RT. These errors are due to limited angular resolution and near-field effects at close range. At 1.5 m, the overall accuracy decreased slightly to 94.75%, with the greatest impact seen in rotational gestures. Both CW and CCW achieved 92%, but showed increased mutual confusion and minor misclassification with neighboring gestures. For example, CCW was confused with RLR and UP in 4% of cases, while CW was misclassified as CCW and RT in 2%. In contrast, linear gestures such as DW, LRL, RLR, and UP retained high classification rates between 96% and 98%, confirming the robustness of amplitude-modulated encoding for well-defined motion trajectories.

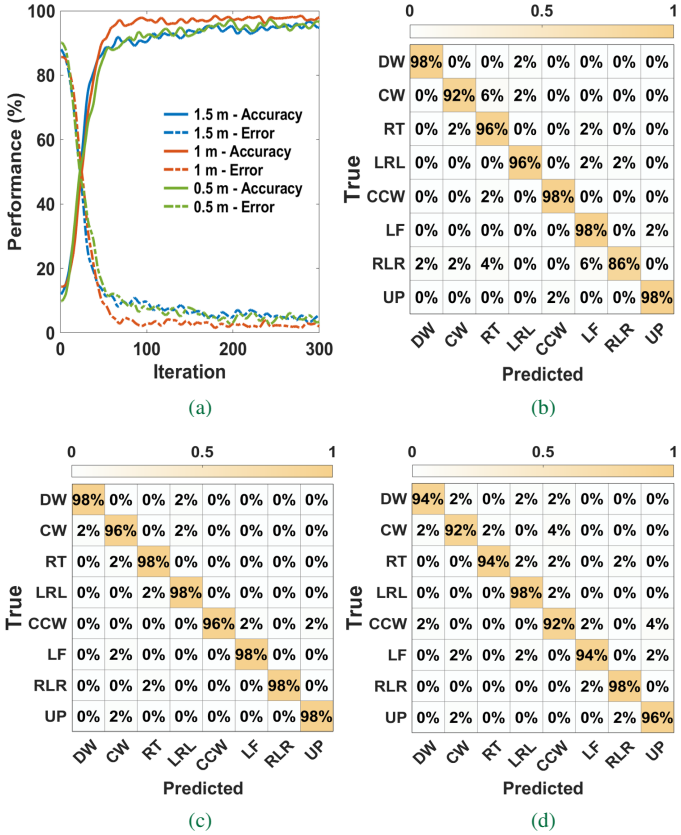


Fig. 14. CNN training and classification at fixed distances. (a) Training performance at 0.5 m, 1.0 m, and 1.5 m. (b–d) Confusion matrices for models trained at 0.5 m, 1.0 m, and 1.5 m, respectively.

5) Cross-Distance Generalization: Addressing spatial variability in user positioning, a generalized CNN model was trained using gesture data collected at three discrete radar-to-FSS distances spanning 0.5 m to 1.5 m in 0.5 m increments. This configuration exposed the model to coarsely spaced spatial viewpoints, enabling it to learn distance-invariant representations more effectively. The classification performance of this model is detailed in the confusion matrix shown in Fig. 15(a), where it achieved an overall accuracy of 95.92%. Gesture classes previously impacted by distance sensitivity, such as RLR and CW, showed measurable improvement, reaching 93% and 95%, respectively. In contrast, gestures that were already stable, such as DW and UP, retained high accuracy with only minimal variation. This confirms generalization without loss of baseline performance.

Further enhancement of spatial representation and accommodation of intermediate user positions were achieved by training a second generalized model on gesture data sampled at five radar-to-FSS distances ranging from 0.5 m to 1.5 m in 0.25 m increments. Denser spatial coverage increased training diversity and spatial continuity. The resulting dataset contained 2,000 labeled instances. The corresponding confusion matrix is presented in Fig. 15(b), where the model achieved an improved average class-wise accuracy of 97.00%. Gains were particularly notable for mirrored and rotational gestures, with RLR and CW both reaching 96%. Linear gestures such as DW, LF, and UP maintained strong performance near 98% accuracy, with reduced inter-class confusion compared to the coarse model.

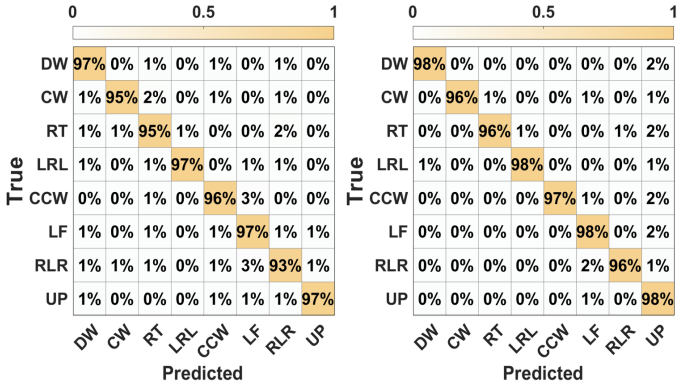


Fig. 15. Confusion matrices for cross-distance generalization. (a) Model trained at three distances from 0.5 m to 1.5 m in 0.5 m increments. (b) Model trained at five distances from 0.5 m to 1.5 m in 0.25 m increments.

A summary comparison of classification accuracies across all configurations is presented in Fig. 16. The 1.0 m model achieved the best single-distance accuracy (97.5%), but the 0.25 m model showed superior overall generalization, outperforming both the coarse generalized and fixed-distance models. These results highlight the importance of fine-grained spatial sampling and confirm the system's robustness under realistic user positions.

The proposed lightweight CNN was compared with two established architectures: ResNet18 [31], a deeper residual network, and EfficientNet-B0 [32], which employs compound scaling with squeeze–excitation modules. As shown in Table II , both models delivered only marginal accuracy gains

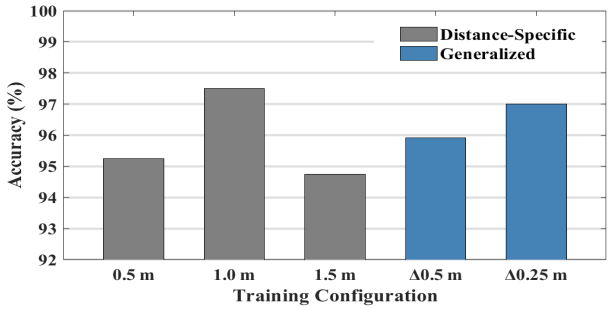


Fig. 16. Summary of classification accuracy for CNN models trained at fixed and generalized spatial configurations. Models were trained at 0.5 m, 1.0 m, and 1.5 m individually, and with spatial generalization using data from 0.5 m to 1.5 m in 0.5 m ($\Delta 0.5$ m) and 0.25 m ($\Delta 0.25$ m) increments.

($\leq$0.7%) over the lightweight CNN. However, these small improvements came at the cost of an increase of more than one order-of-magnitude in parameters and memory consumption. To further broaden the evaluation, transformer models based on attention and implemented in PyTorch were also assessed: DeiT-Tiny [33], a data-efficient vision transformer, and ViT-B/16 [34], a widely adopted baseline model. Both models were trained and evaluated on the same dataset and using the same cross-validation protocol as the CNN-based baselines. DeiT-Tiny achieved performance comparable to the proposed model, with only a minor reduction of 0.25% in validation accuracy. In contrast, ViT-B/16 showed a substantially larger degradation of 5.5%, despite its significantly higher model capacity. This behaviour can be explained by the characteristics of the radar spectrograms, which exhibit vertically aligned, horizontally consistent, and highly repetitive time–frequency patterns with only subtle intra-class variations, as seen in Fig. 13. Such structured representations strongly favour convolutional architectures, which naturally exploit local spatial correlations and repeated geometric patterns. Transformer models, on the other hand, generally require larger-scale training and stronger regularisation to fully leverage their capacity in this type of domain. As a result, although they can still perform competitively, the proposed lightweight CNN remains the most suitable choice for real-time radar-based gesture recognition, offering high accuracy, efficient computation, and strong alignment with the underlying characteristics of the input data.

TABLE II
ACCURACY–COMPLEXITY TRADE-OFF: LIGHTWEIGHT CNN VS. DEEP CNNS AND TRANSFORMERS

Model	Parameters	Size (MB)	Accuracy
Lightweight CNN	423,800	1.62	97.00%
ResNet18	11,689,512	46.7	97.45%
EfficientNet-B0	5,300,000	20.2	97.70%
DeiT-Tiny	5,700,000	22.0	96.75%
ViT-B/16	86,600,000	330.0	91.50%

6) Gesture Vocabulary and Accuracy under Different Panel Counts: The number of active RIS panels plays a decisive role in determining both the gesture vocabulary coverage and the overall recognition robustness, as summarized in Table III. With a single panel, the received modulation collapses to a one-dimensional trace, permitting discrimination of only two coarse gestures (downward swipe and zigzag). Employing two panels expands the vocabulary to six gestures, though coverage remains orientation-dependent: a left–right configuration enables horizontal gestures (LF, RT, CW, CCW) while collapsing vertical ones (DW, UP), whereas a top–bottom configuration enables vertical gestures (DW, UP, CW, CCW) while collapsing horizontal ones (LF, RT). With three panels, the full eight-gesture vocabulary is nominally recovered; however, the reduced spatial diversity introduces systematic ambiguities, particularly among mirrored swipes and rotational classes. In contrast, the complete four-panel configuration restores full angular and spectral diversity, ensuring robust separability across the entire vocabulary.

TABLE III
EFFECT OF RIS PANEL COUNT ON GESTURE COVERAGE

Panels (N)	1	2	3	4
Gestures Supported	2	6	8	8

This trend is quantified in Fig. 17(b), which presents the difference confusion matrix between the three-panel system and the fine 0.25 m-increment generalized model described in Section III-5. While both systems nominally support the complete vocabulary, the three-panel configuration achieves an average accuracy of 89.6%, representing a 7.4% degradation relative to the full four-panel design. The most pronounced losses occurred in swipes: DW (–12%), RT (–9%), LF (–17%), and UP (–12%), with strong mutual confusions (+10%, +8%, +16%, +12%). These degradations result from reduced angular diversity with fewer panels. In contrast, zigzag gestures remained stable within ±3% due to their distinctive temporal pattern, while CW and CCW showed only moderate decline, retaining separability through their circular trajectories.

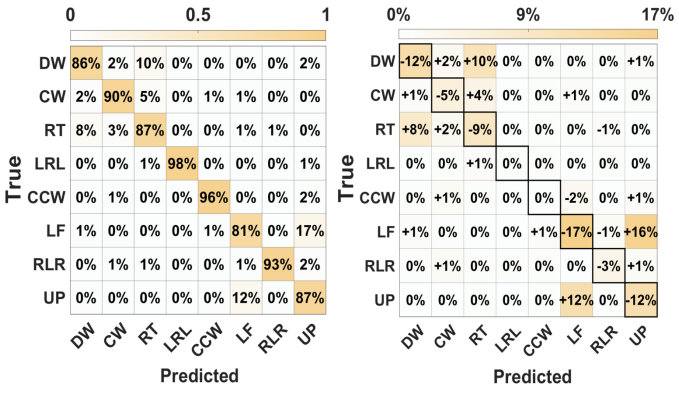
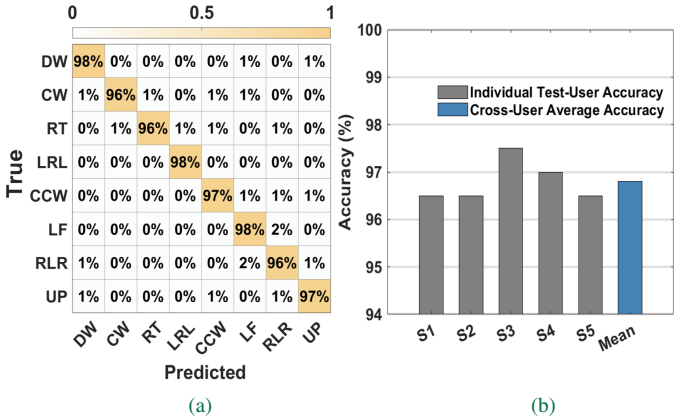


Fig. 17. Confusion matrices: (a) three-panel RIS configuration; (b) difference relative to the four-panel configuration.

7) Cross-User Evaluation: The subject-independent generalization capability of the proposed system was evaluated using a five-fold leave-one-subject-out (LOSO) cross-validation protocol across the five participants. In each fold, gesture data from four participants were used for training, and data

from the remaining participant (unseen during training) were reserved for testing. This was repeated until every participant had served once as the test subject, ensuring accuracy was assessed on unseen users. Evaluation used the fine 0.25 m-increment generalized model. The system achieved an average cross-user recognition accuracy of 96.8%, with per-participant performance ranging from 96.5% to 97.5%, as illustrated in Fig. 18(b). Accuracy therefore remained consistently above 96% for all participants.



(a) (b)

Fig. 18. Cross-user evaluation: (a) overall average confusion matrix across participants under LOSO testing; (b) individual participant accuracies (S1–S5) together with the overall cross-user average. Data collected across five distances (0.5–1.5 m in 0.25 m increments).

8) Comparative Analysis with Related Works: The proposed RIS-assisted continuous-wave (CW) radar is positioned within the broader landscape of radar-based gesture sensing by comparing it with representative systems reported in the literature, as summarized in Table IV. Conventional radars generally treat the hand as an active scatterer and derive gesture features from velocity-dependent observables such as Doppler, micro-Doppler, or range–angle signatures. This approach makes them sensitive to gesture speed, vulnerable to clutter, and performance-limited at longer ranges. Achieving higher accuracy typically requires additional radar hardware and computational complexity.

Certain FMCW systems report slightly higher recognition accuracy, but this improvement is achieved at the cost of limited range and high complexity, including wideband chirps, multi-channel receivers, angle estimation, and computationally intensive networks. FSK-based solutions extend the usable range and can operate with lightweight CNNs, yet they remain tied to velocity-dependent features and require active modulation/demodulation circuitry, which reduces robustness in cluttered or variable environments. By contrast, the proposed RIS-assisted CW radar achieves competitive accuracy with only a single Tx/Rx chain and a simple CNN, while maintaining robust recognition performance across extended ranges up to 1.5 m.

IV. Conclusion

The novelty of the proposed solution lies in the integration of a 24 GHz continuous-wave radar with frequency-modulated FSS panels and a lightweight CNN for real-time hand gesture recognition. By leveraging frequency-specific amplitude modulation, the system eliminates the need for Doppler processing and beamforming, while offering robust spatial discrimination and resilience to occlusion, ambient motion, and user variability. The system achieves a classification accuracy of up to 97.0% in recognizing eight dynamic hand gestures at interaction distances ranging from 0.5 m to 1.5 m, confirming its reliability under various operating conditions. The compact neural network architecture enables fast and memory-efficient inference on embedded platforms, making the system suitable for deployment in low-power, real-time human–machine interfaces where user privacy and minimal computational load are critical.

Future work will aim to increase the system's robustness and versatility by expanding the gesture vocabulary, incorporating more complex dynamic hand motions, and enhancing spatial coverage. Potential improvements include increasing the number or size of modulated FSS panels. Further exploration of advanced learning architectures will also be considered to accommodate richer gesture sets and more complex temporal dynamics. These enhancements aim to advance the system toward a robust and privacy-aware gesture interface suitable for diverse real-world embedded applications.

References

[1] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, and A. A. Shiam, "A methodological and structural review of hand gesture recognition across diverse data modalities," *arXiv preprint arXiv:2408.05436*, 2024.

[2] H. Xu, Z. Wang, H. Huang, and F. Chen, "Radar-based hand gesture recognition: A review," *Sensors*, vol. 22, no. 1, p. 295, 2022.

[3] W. Mucha and M. Kampel, "Addressing privacy concerns in depth sensors," in *Computers Helping People with Special Needs*. Springer, 2022, pp. 492–500.

[4] J. R. Medina-Quero, L. M. Jimenez, and R. Medina, "Gesture recognition using infrared sensors and a machine learning approach," *Sensors*, vol. 21, no. 19, p. 6563, 2021.

[5] Y. Dong, J. Liu, and W. Yan, "Dynamic hand gesture recognition based on signals from specialized data glove and deep learning algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[6] Y. Jiang, L. Song, J. Zhang, Y. Song, and M. Yan, "Multi-category gesture recognition modeling based on semg and imu signals," *Sensors*, vol. 22, no. 15, p. 5855, 2022.

[7] H. Zhang, Y. Huang, and W. Zheng, "A survey of wearable-based hand gesture recognition systems with deep learning," *IEEE Access*, vol. 9, pp. 137 240–137 256, 2021.

[8] S. M. R. Islam, O. Boric-Lubecke, and V. Lubecke, "Contactless radar-based sensors: Recent advances in vital-signs monitoring of multiple subjects," *IEEE Microwave Magazine*, vol. 23, no. 7, pp. 72–91, 2022.

[9] Z. Zhang, Z. Tian, Y. Zhang, and M. Zhou, "U-deephand: FMCW radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9926–9934, 2019.

[10] J. Wang, Y. Wu, C. Li, and L. Wang, "Hand gesture recognition based on dual FMCW radar and transformer networks," *IEEE Access*, vol. 10, pp. 12 709–12 718, 2022.

[11] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-Gesture Recognition Using Two-Antenna Doppler Radar With Deep Convolutional Neural Networks," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, 2019.

[12] J. Y. Eom, W. S. Jeon, and D. G. Jeong, "UWB impulse radar-based open-set gesture recognition using transformer and one-versus-rest classifier," *IEEE Sensors Letters*, vol. 8, no. 1, pp. 1–4, 2024.

[13] T. Jhuang, H. Hsu, H. Chen, and C. Lin, "A real-time radar-based gesture recognition system using CNN-BiLSTM architecture," *Sensors*, vol. 22, no. 4, p. 1582, 2022.

[14] L. Zhu, Z. Chen, and S. Wu, "Robust hand gesture recognition using deformable dual-stream cnn-tcn fusion network," *Sensors*, vol. 23, no. 3, p. 1045, 2023.

TABLE IV
HAND GESTURE RECOGNITION STUDIES USING RADAR

Hardware	Accuracy	Radar Complexity	Range	Radar Antennas	ML Model	Gesture Features
24 GHz CW + RIS	97%	Low	0.50–1.50 m	1 Tx + 1 Rx	Lightweight CNN	RIS amplitude
24 GHz CW [35]	84.1%	Low	∼0.30–0.40 m	1 Tx + 1 Rx	Quadratic SVM	Micro-Doppler
24 GHz Doppler [11]	95%	Medium	0.10–0.30 m	1 Tx + dual-Rx	Deep CNN	Micro-Doppler
24 GHz FSK [17]	93.67%	High	0.30–1.80 m	1 Tx + 1 Rx	Lightweight CNN	Micro-Doppler
24 GHz FMCW [36]	96.73%	High	0.30 m	1 Tx + 1 Rx	Multi-scale GNN	Micro-Doppler
60 GHz FMCW [37]	93.87%	High	∼0.50 m	1 Tx + 3 Rx	CNN + LSTM	RDM + AoA
60 GHz FMCW [38]	98%	High	0.40–0.75 m	1 Tx + Multiple Rx	NN + BiLSTM	RDM + AoA

[15] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-doppler signatures with convolutional neural network," in *Proc. IEEE Radar Conference*, 2016, pp. 1–6.

[16] T. Ali, F. Iqbal, M. A. Jan, and M. A. Javed, "End-to-end dynamic hand gesture recognition using mmwave radar and deep learning," *IEEE Access*, vol. 10, pp. 102 357–102 368, 2022.

[17] T. Yang, K. Ma, H. Wang, Z. He, and J. Wang, "Hand gesture recognition using fsk radar sensors," *Sensors*, vol. 24, no. 1, p. 349, 2024.

[18] S. Guo, S. Lv, H. Zhang, J. Ye, and P. Zhang, "Reflecting modulation," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2548–2561, 2020.

[19] P. Gao, L. Lian, and J. Yu, "Wireless area positioning in ris-assisted mmwave systems: Joint passive and active beamforming design," *IEEE Signal Processing Letters*, vol. 29, pp. 1372–1376, 2022.

[20] M. Jung, W. Saad, M. Debbah, and C. S. Hong, "On the optimality of reconfigurable intelligent surfaces (riss): Passive beamforming, modulation, and resource allocation," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4347–4363, 2021.

[21] M. H. Khoshafa, O. Maraqa, J. M. Moualeu, S. Aboagye, T. M. Ngatched, M. H. Ahmed, Y. Gadallah, and M. Di Renzo, "Ris-assisted physical layer security in emerging rf and optical wireless communication systems: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2024.

[22] S. Milici, J. Lorenzo, A. Lazaro, R. Villarino, and D. Girbau, "Wireless breathing sensor based on wearable modulated frequency selective surface," *IEEE Sensors Journal*, vol. 17, no. 5, pp. 1285–1292, 2016.

[23] M. Asif Haider and Y. D. Zhang, "Ris-aided integrated sensing and communication: A mini-review," *Frontiers in Signal Processing*, vol. 3, p. 1197240, 2023.

[24] X. Wang, Z. Fei, and Q. Wu, "Integrated sensing and communication for ris-assisted backscatter systems," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13 716–13 726, 2023.

[25] X. Zhu, W. Chen, Z. Li, Q. Wu, Z. Zhang, K. Wang, and J. Li, "Ris-aided spatial scattering modulation for mmwave mimo transmissions," *IEEE Transactions on Communications*, vol. 71, no. 12, pp. 7378–7392, 2023.

[26] F. Morabet, A. Lazaro, M. Lazaro, R. Villarino, and D. Girbau, "Driver activity monitoring based on modulated frequency selective surface and millimeter-wave radar," *IEEE Sensors Journal*, 2025.

[27] M. A. Richards, J. A. Scheer, and W. A. Holm, *Principles of Modern Radar: Basic Principles*. SciTech Publishing / IET, 2010.

[28] R. Corporation, "Ro4000 series high frequency circuit materials - ro4003c and ro4350b laminates data sheet," Rogers Corporation, Technical Report 1, 2022, [Online]. [Online]. Available: https://www.rogerscorp.com/advanced-electronics-solutions/ro4000-series-laminates/ro4003c-laminates

[29] MACOM Technology Solutions, "Madp-000907-14020x: Silicon pin diode datasheet," MACOM Technology Solutions, Technical Report, 2023, [Online]. [Online]. Available: https://cdn.macom.com/datasheets/MADP-000907-14020x.pdf

[30] RFbeam Microwave GmbH, "K-md7 evaluation kit datasheet," RFbeam Microwave GmbH, Technical Report Rev. A, 2023, [Online]. [Online]. Available: https://rfbeam.ch/wp-content/uploads/dlm_uploads/2025/03/K-MD7-Datasheet.pdf

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," pp. 6105–6114, 2019.

[33] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Training data-efficient image transformers and distillation through attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[35] A. Bannon, R. Capraru, and M. Ritchie, "Exploring gesture recognition with low-cost cw radar modules in comparison to fmcw architectures," in *Proc. Radar 2020 Conf.*, 2020.

[36] Z. Xiong, K. Ma, and N. Yan, "Hand gesture recognition based on micro-doppler radar using graph neural network," *Electronics Letters*, vol. 60, no. 3, 2024.

[37] Y. Wang *et al.*, "Real-time hand gesture recognition in clinical settings: A low-power fmcw radar integrated sensor system with multiple feature fusion," *Sensors*, vol. 25, no. 13, p. 4169, 2025.

[38] Y.-C. Jhaung, Y.-M. Lin, C. Zha, J.-S. Leu, and M. Köppen, "Implementing a hand gesture recognition system based on range–doppler map," *Sensors*, vol. 22, no. 11, p. 4260, 2022.

Farid Morabet received the Bachelor's degree in Physics and Electronic Science and the Master's degree in Telecommunication Systems Engineering from Abdelmalek Essaâdi University, Tetouan, Morocco, in 2018 and 2021, respectively. He is currently a Pre-doctoral Research Staff in Training with the Department of Electronic, Electrical, and Automatic Engineering at Rovira i Virgili University (URV), Tarragona, Spain, where he started in 2023. His current research interests include the development of advanced electromagnetic systems, millimeter-wave technologies, and innovative sensor applications.

From October 2022 to March 2023, he was awarded an Erasmus+ Mobility Grant for research collaboration at AntennaLab, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, focusing on wireless systems for biomedical applications.

Antonio Lazaro (M'07–SM'16) was born in Lleida, Spain, in 1971. He received the M.S. and Ph.D. degrees in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1994 and 1998, respectively. He then joined the faculty of UPC, where he currently teaches a course on microwave circuits and antennas. Since July 2004, he is a Full-Time Professor at the Department of Electronic Engineering, Universitat Rovira i Virgili (URV), Tarragona, Spain. His research interests are microwave device modeling, on-wafer noise measurements, monolithic microwave integrated circuits (MMICs), low phase noise oscillators, MEMS, RFID, UWB and microwave systems.

Marc Lazaro was born in Tarragona, Spain, in 1995. He received the BS in Industrial Electronics and Automation Engineering and the MS in Electronic Systems Engineering and Technology (METSE) from Rovira i Virgili University, Tarragona, Spain, in 2017 and 2018, respectively. Up until now, he has accumulated professional experience as a data acquisition engineer and as embedded systems developer. Since 2019 he has been working toward the Ph.D. degree in the Department of Electronics at the Rovira i Virgili University. His research activities are focused on semipassive RFID technologies based on backscattering communication and novel applications based on mmWave identification (MMID).

Ramon Villarino was awarded a degree in Telecommunications Technical Engineering by Ramon Llull University (URL) in Barcelona, Spain, in 1994, a degree in Senior Telecommunications Engineering by the Universitat Politècnica de Catalunya (UPC) in Barcelona, Spain, in 2000 and a doctorate by the UPC in 2004. In 2005-2006, he was a Research Associate at the Technological Telecommunications Center of Catalonia (CTTC) in Barcelona, Spain. He worked as a Researcher and Assistant Professor at the Universitat Autònoma de Barcelona (UAB) from 2006 to 2008. Since January 2009 he has been a full-time professor at Universitat Rovira i Virgili (URV) in Tarragona, Spain. His research activities focus on radiometry, microwave devices, and systems based on UWB, RFIDs, and frequency-selective structures using MetaMaterials (MM).

David Girbau (M'04–SM'13) was awarded a BSc in Telecommunication Engineering, a Master's in Electronics Engineering, and a doctorate in Telecommunication by Universitat Politècnica de Catalunya (UPC) in Barcelona, Spain, in 1998, 2002 and 2006, respectively. From February 2001 to September 2007 he was a Research Assistant at UPC. From September 2005 to September 2007 he was a part-time Assistant Professor at Universitat Autònoma de Barcelona (UAB). Since October 2007 he has been a full-time professor at Universitat Rovira i Virgili (URV) in Tarragona, Spain. His research interests include microwave devices and systems, with an emphasis on UWB, RFIDs, RF-MEMS and wireless sensors.