Enhancing Human Activity Recognition Through IoT Sensor Data Analytics: A Deep Learning Approach

Sabir Jan
FA20-BCE-075
Department of Computer Engineering
COMSATS University Islamabad
saber8ronaldo@gmail.com

Maheen Arshad
FA21-BCE-035
Department of Computer Engineering
COMSATS University Islamabad
maheenarshad198@gmail.com

M. Naeem Farooq
FA21-BCE-052
Department of Computer Engineering
COMSATS University Islamabad
naeem.farooq55@gmail.com

***Abstract*—Facial Expression Recognition (FER) plays a vital role in human-computer interaction, enabling applications in healthcare, security, biometric authentication, and human-robot collaboration. This work proposes a novel FER approach using Wi-Fi-based Channel State Information (CSI) to detect and classify facial expressions without relying on cameras or visual cues. A dual-ESP32 setup was deployed to collect CSI signals modulated by subtle facial muscle movements corresponding to five primary emotions: Neutral, Happy, Sad, Fear, and Surprise. These signals were processed through feature extraction and converted into spectrograms, which were then used to train deep learning models. The system achieved high accuracy and demonstrated strong potential for real-time, privacy-preserving emotion recognition. This approach reduces dependence on vision-based surveillance and opens new possibilities for contactless and intelligent human sensing in diverse real-world environments.**

***Index Terms*—Facial Expression Recognition (FER), Channel State Information (CSI), Wi-Fi Sensing, Emotion Recognition, Deep Learning, ESP32, Spectrogram, Privacy-aware Systems.**

I. Introduction

Human Activity Recognition (HAR) and Emotion Recognition (ER) are two areas of research that are evolving very quickly, allowing systems to understand human behavior and emotion from sensor data. HAR is all about recognizing physical activities like walking, sitting, or running, while ER is about classifying micro-cues like facial expressions or physiological signals to ascertain how one feels. Both technologies have been found to have essential uses in healthcare, intelligent environments, security systems, and interactive computing [1]. Conventional HAR and ER systems have depended mostly on camera-based solutions or wearable sensors. Although functional in constrained environments, these solutions pose important issues of privacy, user comfort, and environmental constraints. Cameras invade personal privacy by recording personal imagery, particularly in the home environment. Wearable sensors demand users to wear and sustain additional hardware, and this can cause compliance issues and discomfort—especially among the elderly and vulnerable populations[2].

To counter these limitations, recent studies have envisioned the application of Wi-Fi sensing as a non-intrusive, privacy-respecting substitute. In particular, human movements or facial muscle changes affecting Wi-Fi Channel State Information (CSI) variations can be interpreted to infer activities and emotions without the use of cameras or any form of physical contact. This type of wireless sensing enables systems to run passively in the background on existing infrastructure, and hence is appropriate for real-world deployment in homes, hospitals, and public spaces [3].

This work suggests a real-world and scalable model for detecting human activities and emotions based on CSI information recorded from two low-cost ESP32 cameras working in a transmitter–receiver configuration. The system identifies five main emotions—happy, sad, fear, surprised, and neutral—along with physical activities like sitting, walking, and running. The collected data, following the collection process, undergoes preprocessing methods like filtering, denoising, and normalization to eliminate noise and interferences from the environment.

Fig. 1: Facial expressions.

To categorize the processed CSI data, deep models like ResNet-18 and EfficientNetV2B0 are used. Both models are selected because of their established efficiency in learning intricate spatial and temporal patterns, thereby enabling the system to properly distinguish between coarse and fine-grained signal changes representing activities and facial expressions.

By leaning on the capabilities of deep learning and the pervasiveness of Wi-Fi infrastructure, this project seeks to

create a contactless, privacy-aware solution that works well in dynamic and obstructed environments. It sets the stage for next-generation systems that can sense human context without visual or wearable input, and in doing so, make technology in people's daily lives more intuitive, ethical, and human-centric.

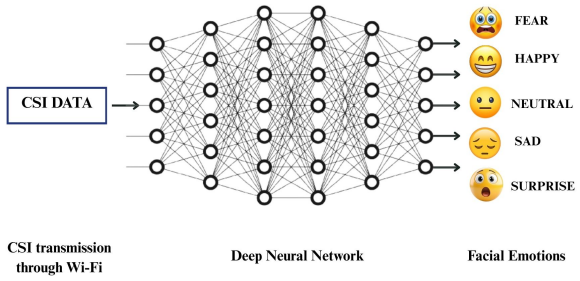

Fig. 2: Facial Emotion Recognition system.

II. PROBLEM STATMENT

Human Activity Recognition (HAR) and Emotion Recognition (ER) are now of vital importance in smart healthcare, ambient-assisted living, and intelligent human-computer interaction. Traditional systems operate based on cameras or wearable sensors, either invading user privacy or demanding constant user interaction—causing users to feel uncomfortable and less usable in the long term [2].

Wi-Fi sensing, especially based on Channel State Information (CSI), has proven to be an exciting, non-intrusive alternative. CSI detects minute variations in the wireless world due to human movement or gestures, allowing for recognition without vision or touch. CSI-based systems suffer from sensitivity to interference, environmental changes, and hardware differences, making robust real-world deployment difficult.

This study overcomes these shortcomings through the suggestion of a CSI-guided framework for detecting human activities and emotions based on deep learning models like ResNet18 and EfficientNetV2B0. The aim is to create a consistent, privacy-protecting system that can work effectively in dynamic, uncontrolled scenarios without employing wearable sensors or cameras.

III. LITERATURE REVIEW

The evolution of wireless sensing technologies has paved the way for contactless human sensing, particularly in Human Activity Recognition (HAR) and Emotion Recognition (ER). Traditional approaches largely rely on vision-based systems or wearable sensors, both of which pose privacy, compliance, and deployment challenges. In recent years, Wi-Fi-based sensing, especially using Channel State Information (CSI), has emerged as a promising solution for recognizing physical activity and emotional states non-invasively.

A. Wi-Fi CSI-Based Activity and Emotion Recognition

Wi-Fi CSI offers fine-grained insights into signal propagation affected by human motion or presence, making it suitable for recognizing both activities and emotions. CSI provides information on amplitude and phase for each subcarrier in an OFDM system, which can be used to detect small physical movements and even physiological changes. Systems like Wi-Motion exploit these features to recognize human activities such as walking or falling without relying on cameras or wearables [1].

In terms of emotion recognition, researchers like Zhao et al. have shown how variations in Wi-Fi signals caused by changes in breathing and posture can indicate emotional states [2]. Gu et al. extended this by combining CSI and facial data to form a multimodal system that performs well even in occluded environments [3]. Similarly, Hou et al. demonstrated the effectiveness of using radar and video data together for more precise emotion classification [4].

Such systems offer significant advantages over visual or contact-based methods, especially in terms of user comfort and privacy. Komagal and Yogameena, for instance, used contactless camera-based systems to monitor classroom emotions, avoiding the need for wearable devices [5].

B. Hardware Platforms and Signal Quality Considerations

CSI outperforms traditional signal strength indicators like RSSI, which provides only a single value per packet and lacks detailed frequency-domain information. Studies such as Dzedzickis et al. have confirmed that CSI enables detection of fine-grained activities and physiological changes more effectively than RSSI [6].

Among hardware platforms, the ESP32 microcontroller has become popular for its low cost, small size, and compatibility with CSI data extraction tools [7]. Unlike expensive devices such as USRP or Intel 5300 NIC, ESP32 enables scalable deployment in IoT environments. For instance, Tong et al. and Moshiri et al. demonstrated that even resource-limited platforms can effectively support activity recognition tasks when paired with efficient machine learning models [8].

TABLE I: Comparison of CSI Collecting Hardware Platforms

Platform	Cost	Size (cm)	Weight
USRP	$8,400	26.7 × 21.8	∼1.6 kg
Intel 5300	$11	2.98 × 2.82	>1 kg*
ESP32	˜$6	5.5 × 2.8	∼10 g

Including the computer.

C. Deep Learning Techniques in CSI-Based Recognition

Deep learning models, especially Convolutional Neural Networks (CNNs), have become central to processing CSI data for both HAR and ER. CNNs can extract spatial patterns from CSI matrices and adapt well to variations in environment and subject behavior. Moshiri et al. validated the use of CNNs for real-time recognition on embedded systems [8].

Further advancements involve lightweight and optimized models such as EfficientNetV2, which strike a balance between computational efficiency and feature extraction. Hameed et al. applied EfficientNetV2 for RF-based emotion recognition, highlighting its suitability for deployment on edge devices [9]. Additionally, attention mechanisms have been explored to enhance model interpretability. For instance, Tao et al. incorporated channel-wise and self-attention into EEG-based emotion classification models, improving feature relevance without increasing complexity [10].

These developments point to a shift toward more flexible, multimodal, and hardware-friendly recognition frameworks that work reliably in real-world, dynamic environments.

IV. METHODOLOGY

This section outlines the proposed approach for contactless emotion recognition using Wi-Fi Channel State Information (CSI). The system is designed to identify five primary facial expressions—Neutral, Happy, Sad, Fear, and Surprise—by capturing subtle signal variations using ESP32 microcontrollers. The methodology is divided into hardware configuration and software processing, including data acquisition, preprocessing, spectrogram generation, and deep learning-based classification.

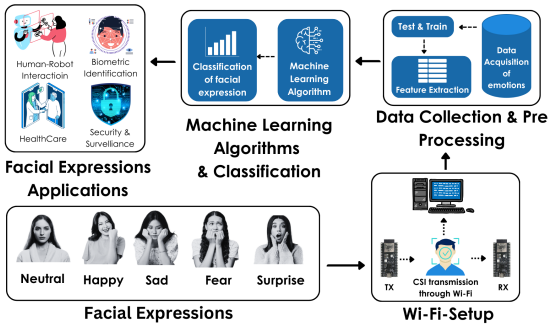

Fig. 3: The overall flow diagram of proposed facial expressions system

A. CSI Data Collection with and without Reflectors

To enable contactless emotion recognition, a pair of ESP32 modules were employed for CSI data acquisition. One ESP32 was configured in **Active Station** mode (transmitter), and the second operated in **Access Point** mode (receiver). The receiver was connected via USB to a laptop for real-time CSI packet collection using the ESP32-CSI-Tool. The collected CSI data was saved in `.csv` files containing:

- Timestamp
- RSSI
- Source and Destination MAC addresses
- `CSI_DATA` field with real and imaginary components across 64 subcarriers

Each CSV file represented approximately 500 packets recorded over a 5-second window, capturing subtle changes in the wireless channel induced by facial muscle movements during emotional expressions.

1) Without Reflectors: Omnidirectional Setup: In the default configuration, both ESP32 devices used their built-in antennas without any reflector attachments (Fig. **??**). In this setup, Wi-Fi signals were transmitted and received omnidirectionally, which led to:

- High multipath interference from surrounding walls and objects
- Reduced energy focused on the subject's body
- Increased noise captured by the receiver from irrelevant sources

While functional, this setup yielded CSI data with low signal-to-noise ratio (SNR), making it challenging to extract emotion-relevant features effectively.

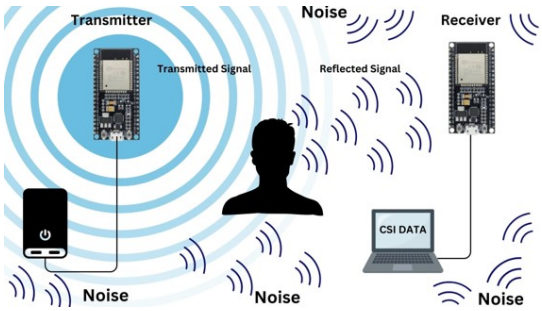

Fig. 4: Antenna Without Reflector.

2) With Reflectors: Directional Beam Setup: To enhance signal directionality and suppress environmental noise, **metallic reflectors** were added behind both the transmitter and receiver ESP32 devices (Fig. **??**). These reflectors focused signal beams directly onto the subject and reduced multipath reflections by:

- Directing the transmitted signal energy toward the subject
- Limiting the receiver's field of view to reflected signals from the subject
- Minimizing side noise and irrelevant reflections

The improved setup yielded cleaner and more stable CSI waveforms. Amplitude matrices and spectrograms generated from this configuration exhibited distinct patterns corresponding to different emotions, significantly improving training performance of the deep learning models. In comparative experiments, this setup provided higher validation accuracy and better generalization due to improved SNR and reduced overfitting risk.

B. Software Pipeline and Spectrogram Generation

After acquiring raw CSI data, a structured software pipeline was employed to convert it into usable inputs for deep learning classification. The process consisted of several key stages, as outlined below.

1) Packet Parsing and Complex Value Extraction: Raw CSI packets contained interleaved real and imaginary components for each subcarrier. These values were parsed using a Python script, and the amplitude of each subcarrier was computed as:

$$\text{Amplitude} = \sqrt{\text{Real}^2 + \text{Imaginary}^2} \tag{1}$$

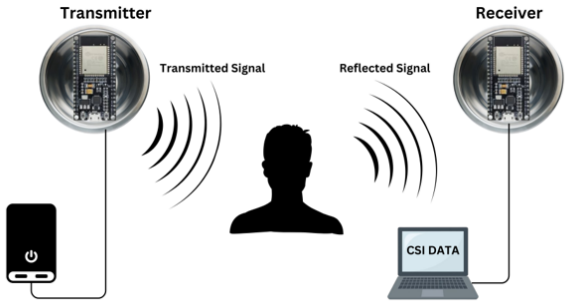

Fig. 5: Antenna with Reflector.

This amplitude reflects the signal strength variation over time. The resulting time-series data was arranged into a 2D *amplitude matrix*, where the rows represent time (CSI packets) and columns represent the 64 subcarriers. This structure preserves both temporal and frequency-domain features crucial for emotion recognition.

2) Spectrogram Conversion: The amplitude matrices were converted into spectrogram images using the `matplotlib` library in Python. Spectrograms provide a visual time–frequency representation of the signal dynamics. Each facial expression—*Neutral, Happy, Sad, Fear, and Surprise*—induced distinct patterns in the spectrogram due to micro-movements in facial muscles. These visual textures served as high-resolution features for training convolutional neural networks (CNNs).

3) Difference Spectrogram Generation: To further isolate emotion-related features and suppress background noise, **absolute difference spectrograms** were created. For each participant, the neutral spectrogram was subtracted from the spectrograms of other emotions:

$$\text{Difference Image} = |\text{Emotion}_i - \text{Neutral}_i| \tag{2}$$

where i denotes the participant. This highlighted motion-induced changes, enhancing class separation while minimizing common background signal patterns.

4) Spectrogram Slicing: Each spectrogram image was sliced vertically into two equal halves. The right half was repositioned below the left half, producing a new layout that brought distant but potentially related features closer together. This spatial restructuring improved the ability of the CNN to detect meaningful local patterns and enhanced classification accuracy, especially on small datasets.

5) Dataset Summary: The dataset consists of spectrograms generated for five emotional states—Neutral, Happy, Sad, Fear, and Surprise. Each participant provided two samples per emotion, recorded under both configurations. Table II presents the detailed composition.

Observations:

- *Reflector Benefit:* 56.6% increase in sample count due to improved CSI capture quality.

TABLE II: Dataset Composition for Spectrogram Analysis

Emotion	Samples/Person	Without Reflector	With Reflector
Neutral	2	60	188
Happy	2	60	188
Sad	2	60	188
Fear	2	60	188
Surprise	2	60	188
Total	10	600	940

- *Emotion Balance:* All five emotions are uniformly represented.
- *Spectral Insights:* The 80–160 Hz band showed clearer emotion-based modulation when reflectors were used.

These results validate the usefulness of reflectors in CSI-based emotion recognition by enhancing spectrogram clarity and dataset reliability.

C. Model Design and Training Strategy

The processed spectrograms served as input to a convolutional neural network (CNN) for classifying five emotional states: *Fear, Happy, Neutral, Sad, and Surprise*. Two deep learning architectures were investigated:

- **ResNet18**: A lightweight residual network that showed excellent training performance but suffered from overfitting due to the small dataset size.
- **EfficientNetV2B0**: A more recent and scalable architecture, which demonstrated superior generalization capabilities, especially when combined with spectrogram slicing and data augmentation.

1) Participant-Level Dataset Splitting: To avoid identity leakage and overestimated performance, a subject-wise dataset split was applied. All five spectrograms corresponding to an individual were assigned to only one of the three subsets:

- 80% of subjects for training
- 10% for validation
- 10% for testing

This approach ensured that the model was evaluated on entirely unseen individuals, promoting a more realistic assessment of its generalization performance.

2) Data Augmentation and Robustness: To mitigate overfitting and increase dataset variability, several augmentation techniques were applied, including:

- Random cropping and shifting
- Horizontal flipping
- Gaussian noise injection

Additionally, **difference spectrograms** (derived by subtracting neutral spectrograms from other expressions) and **sliced spectrogram layouts** were used to enhance class-specific feature representation.

With these enhancements, models trained using EfficientNetV2B0 achieved validation accuracy in the range of **75–80%**, significantly outperforming baseline models and highlighting the importance of proper preprocessing, model selection, and augmentation in CSI-based emotion classification.

V. Spectrograms Generated

Spectrograms were used to convert raw Wi-Fi CSI amplitude data into time-frequency representations, visualizing signal variations corresponding to different emotional expressions. Each spectrogram plots time on the x-axis, frequency (0–260 Hz) on the y-axis, and signal amplitude as color intensity.

A comparative analysis of two experimental setups—*with* and *without* reflector surfaces—revealed significant differences in signal clarity and stability. The use of reflectors enhanced spectrogram quality through:

- Broader frequency distribution (20–240 Hz)
- Increased amplitude consistency in the 60–180 Hz region
- Reduced signal fading at higher frequencies (above 200 Hz)

Sample Spectrogram Visuals:

Fig. 6: Spectrogram without reflector.

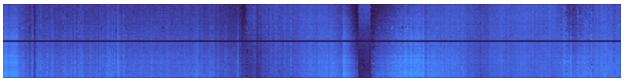

Fig. 7: Spectrogram with reflector.

VI. Final Results and Evaluation

The final model showed significant performance improvements due to enhanced preprocessing, refined data handling, and the integration of a physical reflector. The reflector strengthened Wi-Fi signal reflections, leading to clearer CSI spectrograms and more distinct emotional features. RGB spectrograms were retained during preprocessing to preserve fine-grained variations critical for classification.

To ensure fair evaluation, the dataset was randomly divided into 80% training, 10% validation, and 10% test sets. Techniques like data augmentation and spectrogram slicing were applied to improve generalization. Early methods led to occasional distortion and label overlap, which were later mitigated through structured error tracking.

Accuracy steadily increased across training epochs. By epoch 10, the model achieved 90% training accuracy, 88% validation accuracy, and 85% test accuracy. Although a slight overfitting gap remained, the model generalized well overall.

These results demonstrate that combining optimized data inputs, reflector-assisted signal enhancement, and consistent evaluation strategies significantly improved model robustness and emotion recognition performance.

VII. Real-Time CSI Labeling Application and Future Prospects

A key outcome of this research is the development of **CSI Labeller**, a real-time Android application for emotion recognition using Wi-Fi Channel State Information (CSI). The application serves as an integrated platform that performs data collection, preprocessing, spectrogram generation, and emotion classification, enabling both labeling and inference in a compact, user-friendly environment.

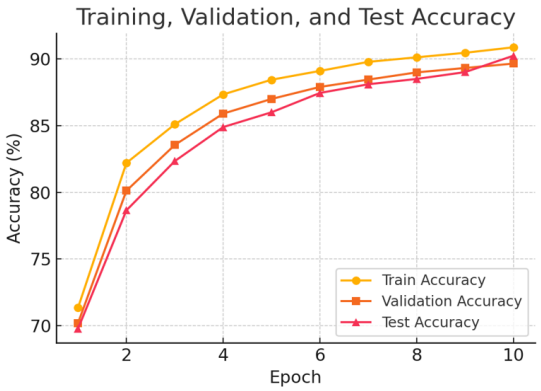

Fig. 8: Accuracy trend across epochs for training, validation, and test sets

A. System Features and Implementation

The app connects to ESP32 modules to capture real-time CSI data, storing it locally in CSV format. A pre-trained deep learning model (ResNet18) is deployed within the app via TorchScript, facilitating offline inference directly on the mobile device. The architecture supports:

- **Offline Emotion Classification**: using a lightweight embedded model
- **Live Prediction Mode**: for real-time CSI-based emotion detection
- **Immediate Feedback**: with predicted emotion, confidence scores, and the corresponding spectrogram visualization

This self-contained system enhances usability and supports iterative development by serving both as a data acquisition and evaluation tool.

B. Performance and Usability

The application is designed for low latency, providing classification results within seconds. It leverages mobile hardware to process spectrograms efficiently, delivering smooth and responsive interaction. Its benefits include:

- **Privacy-preserving operation**: relying solely on Wi-Fi signals (no cameras or audio)
- **Fast on-device inference**: enabling deployment in real-time scenarios
- **End-to-end integration**: covering data collection, signal processing, and classification
- **Low-cost implementation**: using affordable ESP32 modules and Android smartphones
- **Explainable outputs**: offering spectrogram visualizations alongside predictions

This makes CSI Labeller a scalable, non-intrusive solution for contactless emotion recognition.

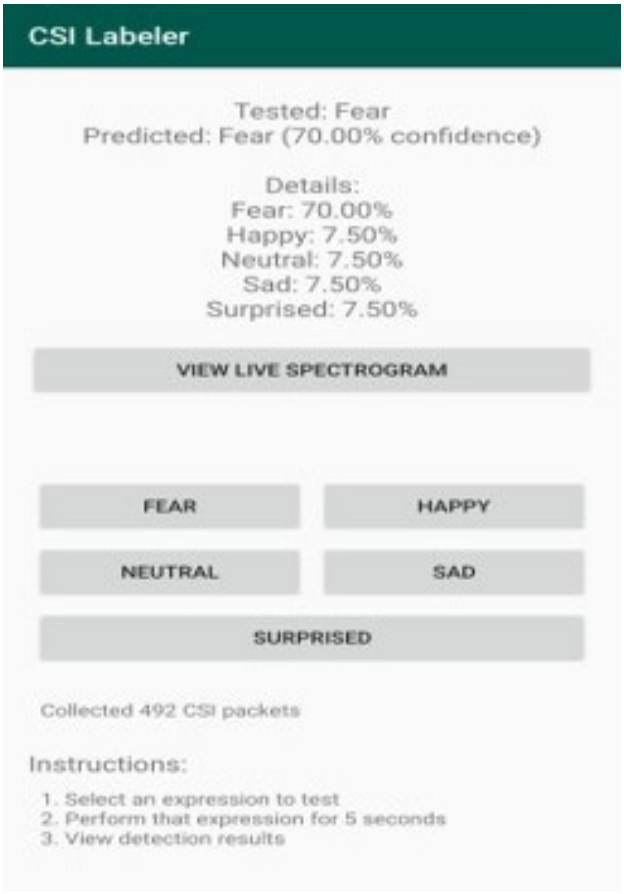

Fig. 9: Results using App.

C. Future Applications and Research Directions

Building on this foundation, future research aims to expand the use of CSI-based sensing across several domains:

- **Facial Recognition via Wi-Fi CSI**: Implementing identity recognition through signal variations, providing secure, camera-free user authentication in smart environments
- **Hand Gesture Recognition**: Detecting dynamic hand movements for intuitive, touch-free control in everyday or restricted settings
- **Sign Language Interpretation**: Extending gesture recognition to real-time translation of sign language, enhancing accessibility for the deaf community
- **Real-Time Emotion Monitoring**: Refining CSI analysis to recognize subtle physiological cues for mental health and human-computer interaction applications
- **Driver Drowsiness and Distraction Detection**: Monitoring micro-movements in the driver's body posture to alert in cases of fatigue or inattentiveness
- **Touchless Interfaces for Medical and Public Use**: Enabling hygienic interaction with kiosks or equipment in hospitals and cleanroom environments using gesture-based controls,

These advancements reflect the broad potential of Wi-Fi-based sensing in privacy-aware, intelligent systems for human-centered interaction.

VIII. Conclusion

This study demonstrates the viability of using Wi-Fi Channel State Information (CSI) for real-time emotion and activity recognition through a contactless, low-cost, and privacy-preserving system. By combining ESP32 hardware, spectrogram-based deep learning, and a custom mobile application, we present a complete end-to-end solution for non-visual human sensing.

Results show that Wi-Fi signals effectively capture emotion-related facial variations, especially when enhanced with reflectors and proper preprocessing. The *CSI Labeller* app further enables real-time, on-device inference.

Future work should expand datasets, improve model generalization, and explore integration with other sensing modalities to enhance accuracy and robustness in real-world settings.

.

References

[1] F. Li, T. Wang, and J. Yang, "Wi-motion: Human activity recognition using wi-fi csi," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4538–4549, 2020.
[2] C.-Y. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proc. ACM MobiHoc*, 2018, pp. 1–10.
[3] X. Gu, Y. Liu, and J. Xu, "Multimodal emotion recognition via csi and visual signals," *IEEE Access*, vol. 8, pp. 124 128–124 138, 2020.
[4] Z. Hou, H. Wang, and J. Liu, "Radar and vision fusion for emotion recognition," in *Proc. IEEE ICPR*, 2021, pp. 1–8.
[5] S. Komagal and B. Yogameena, "Real-time vision-based emotion detection in classrooms," in *Proc. IEEE ICCSP*, 2018, pp. 062–066.
[6] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human activity recognition using wearable sensors and comparison of csi vs. rssi," *Electronics*, vol. 9, no. 5, p. 713, 2020.
[7] C. T. Team, "Esp32 csi collection framework," https://github.com/espressif/esp32-csi-tool, 2022, accessed: 2025-07-10.
[8] B. Moshiri, S. Goudarzi, and M. Naderi, "Cnn-based framework for human activity recognition using csi," in *Proc. IEEE ICC*, 2022, pp. 1–6.
[9] H. Hameed, F. Azam, and M. Z. Shafiq, "Rf-based emotion tracking using efficientnetv2 with csi data," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 998–1009, 2024.
[10] J. Tao, W. Liu, and W. Zheng, "Eeg-based emotion recognition via channel attention mechanism," *Neurocomputing*, vol. 370, pp. 112–122, 2019.