2025 International Conference on Advanced Mechatronics and Intelligent Energy Systems

Lightweight Research on Identity Recognition based on HAR Data CNN-BiLSTM Transfer Learning

AIPCP25-CF-AMIES2025-00057 | Article

Lightweight Research on Identity Recognition based on HAR Data CNN-BiLSTM Transfer Learning

Zicheng Yin

Beijing-Dublin International College at BJUT, Beijing University of Technology, Beijing, 100124, China.

Corresponding author: ZichengYin@emails.bjut.edu.cn

Abstract. Human activity recognition (HAR) has a wide range of applications in the fields of intelligent security and health monitoring. Identification based on WiFi channel state information (CSI) can avoid the high deployment cost and privacy violation of traditional identity recognition methods (such as sensors, cameras, etc.), but its model still has insufficient generalization performance when recognizing different CSI data. Therefore, this paper proposes a bimodal transfer learning system such as CNN-BiLSTM to improve the generalization ability while ensuring recognition accuracy. Through experiments on the NTU-Fi HAR and NTU-Fi HumanID datasets, the performance of CNN-BiLSTM is verified, and a horizontal comparison is made with a variety of single-mode models and bimodal models. The experimental results show that the system is superior in feature extraction of HAR datasets, with an accuracy of 76.61%. Compared with single-mode models such as BiLSTM, the number of parameters is reduced while ensuring accuracy, and the effect is better than bimodal models such as CNN-LSTM, which can better adapt to transfer learning. However, its complex floating-point operations have high requirements on computer performance. For this reason, the experiment provides an improved method to reduce the amount of calculation and provides a new idea for the lightweight research of the model.

INTRODUCTION

Human activity recognition (HAR) technology is widely used in the field of identity recognition, including smart home and security monitoring scenarios. Traditional recognition methods mainly rely on cameras, wearable devices, or infrared sensors, but these methods are often susceptible to light changes, environmental interference, and inconvenience in wearing equipment. In recent years, the technology of using WiFi channel state information (CSI) for human activity recognition has gradually emerged. This method does not require additional equipment to be worn, has the advantages of low deployment cost and high privacy protection, and has opened up a new path for the commercial application of identity recognition.

However, due to the obvious multipath effect of WiFi signals in indoor environments, signal propagation is unstable, resulting in insufficient generalization ability of WiFi CSI-based identity recognition methods in different scenarios. In addition, this method currently requires high hardware computing power and is difficult to meet actual market needs. In response to the above challenges, a large number of research works have been devoted to developing new models and data preprocessing methods to enhance the generalization ability of different HAR datasets. LiWi-HAR [1] extracts key features in the process of compressing CSI data to achieve data lightweight and improves recognition performance through a double hidden layer BPNN classifier based on particle swarm optimization (PSO), but this model may easily fall into local minima or experience gradient vanishing. AutoFi[2] proposes a geometric self-supervised learning algorithm that effectively utilizes low-quality CSI samples by introducing Gaussian noise without destroying the internal information of CSI data, and uses a convolutional neural network-multilayer perceptron (CNN-MLP) model to achieve gait recognition. EfficientFi[3] proposed a quantitative feature algorithm to extract and compress CSI data at the WiFi router end, and then transmit it to the cloud for recovery and classification through CNN, thereby significantly reducing communication overhead, but it has high hardware requirements. SenseFi[4] built a benchmark library containing multiple models and data sets and

evaluated the performance of 12 models in supervised learning, unsupervised learning, and transfer learning tasks, providing a reference for model selection and scene adaptation. AFEE-MatNet[5] proposed an activity-related feature extraction, enhancement, and matching network, which reduces training complexity through data cleaning and enhancement and uses a confusion matrix to verify and correct coarse-grained action prediction. However, this algorithm can only be applied to "continuous action" CSI data, otherwise it will have a serious impact on the accuracy.

Based on the above research, this paper proposes a transfer learning method based on CNN and a bidirectional long short-term memory network (biLSTM). This method extracts spatial features from WiFi CSI data through CNN and combines biLSTM to effectively model temporal features, thereby improving the accuracy and generalization ability of feature extraction of the bimodal model and reducing the requirements for hardware. Subsequently, this paper evaluates from four dimensions and proposes corresponding improvement plans based on the evaluation results.

DATA AND METHODS

Dataset analysis

Table 1 shows that this study used two public CSI datasets (Data - Google Drive) for experiments. The NTU-Fi HAR dataset [3] has 6 labels, including 6 types of actions (walking, running, falling, boxing, circling arms, and cleaning the floor), completed by 20 subjects; the NTU-Fi HumanID [6] dataset has 14 categories, including the above 6 types of gait data of 14 subjects, for identity recognition. The CSI samples of these two datasets have a high degree of fit and are suitable for transfer learning [4].

TABLE 1. Statistics of two CSI datasets

	NTU-FI-HAR	NTU-Fi-HumanID
Collection Platform	Atheros CSI Tool	Atheros CSI Tool
Number of Categories	6	14
Category Name	20 subjects (13 males/7 females) box, circle, clean, fall, run, walk actions	Gait of 14 subjects
Data Size	(3,114,500) (Antenna, Subcarrier, Sampling frequency)	(3,114,500) (Antenna, Subcarrier, Sampling frequency)
Number of Training Samples	936	546
Number of Test Samples	264	294
Training Epochs	30	30

Methods

CNN-BiLSTM Transfer Learning Model

This study uses a convolutional neural network-bidirectional long short-term memory network (CNN-biLSTM) model for identity recognition. First, the spatial features are obtained through a feature extraction layer with two layers of convolution and ReLU activation function, and the pooling operation is used to reduce the amount of calculation and enhance the feature expression ability. Subsequently, the biLSTM network is used to model the bidirectional dependency of the time series data, and the dynamic characteristics of the CSI signal are captured by the time step; then, the spatiotemporal features are mapped to the identity classification label through the fully connected layer, and finally, the Softmax classifier is used to output the recognition result. To improve the generalization performance, this study adopts a transfer learning strategy: first pre-train the CNN-biLSTM on the NTU-Fi HAR dataset, input the obtained feature weights into the CNN, then freeze the CNN layer parameters, and fine-tune the biLSTM and fully connected layers on the NTU-Fi HumanID dataset to further improve the generalization ability of identity recognition. The overall architecture is shown in Figure 1:

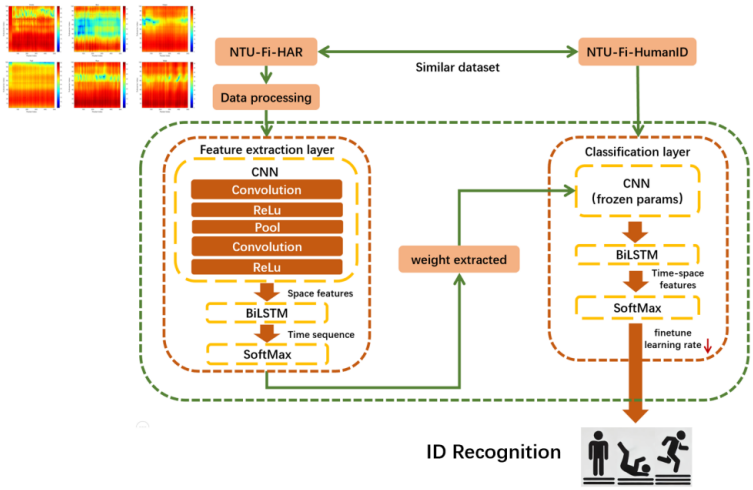

FIGURE 1. CNN-BiLSTM Transfer Learning Architecture (Photocredit : Original)

Experimental Procedures

Since WiFi CSI signals are greatly affected by noise, data preprocessing is required. The following four methods are summarized for learning-based CSI data cleaning:

First, denoise the data. The denoising process removes irrelevant noise and retains only CSI amplitude information to improve the stability and quality of the signal [2]. Second, perform a Doppler analysis. First, calculate the rate of change of the CSI amplitude to extract the human motion speed characteristics, and then use the short-time Fourier transform (STFT) to generate a time-frequency spectrogram to capture the dynamic changes of the motion pattern [7-9]. In addition, the one-dimensional linear interpolation method is used to fill in the missing data points to ensure the integrity and continuity of the CSI data stream [1, 10]. Finally, principal component analysis (PCA) is used to reduce the dimensionality of high-dimensional data, extract the main feature components, and reduce redundant information, thereby improving the computational efficiency and robustness of the model [9]. This study mainly uses the denoising method for the two data sets in Table 1.

Next, the overall experimental process is divided into two stages: pre-training and fine-tuning. The pre-training stage is conducted on the NTU-Fi_HAR dataset, using the CNN-BiLSTM model for preliminary learning. In the model, the CNN module is responsible for preliminary feature extraction, while the core BiLSTM part is used to capture time series information. Its bidirectional structure can process both forward and reverse time series features, thereby better modeling the time dependency within the data. The hyperparameter settings used in pre-training include a learning rate of 1×10^(-3), a training batch size of 16, and a training round of 100. The optimizer uses the Adam algorithm to ensure that the training loss converges quickly [4]. At the same time, the cross-entropy loss function is used for supervised learning, and the hyperparameters are continuously adjusted to obtain the best accuracy. After the pre-training is completed, the model parameters are saved in a path. Subsequently, the fine-tuning stage is entered, and task migration is performed on the NTU-Fi-HumanID dataset. At this time, the pre-training weights are loaded and the classifier part is removed, while the parameters of the CNN part are frozen, and only the BiLSTM and subsequent classification layers are updated to better adapt to the new data distribution. A lower learning rate (5×10^(-4)) was used in the fine-tuning phase, and the number of training rounds was reduced to 75, with the same optimizer and loss function. In the result evaluation, a test module was also provided to calculate the accuracy, loss, number of parameters, and floating-point operations of the model on the test set, ensuring that the performance of the model at each stage was fully monitored.

Evaluation Criteria

The experiment includes four reference standards, including accuracy, cross-entropy loss, number of parameters, and number of floating-point operations. Among them, accuracy measures the proportion of correctly classified samples to the total number of samples, which is used to evaluate the classification performance of the model; cross-entropy loss is used to measure the difference between the predicted distribution and the true label distribution. The smaller the value, the closer the prediction is to the true label; the number of parameters reflects the total number of weights and biases in the neural network. The fewer the parameters, the less storage the model requires; the number of floating-point operations per second is used to quantify the scale of operations performed by the model during forward propagation or training. The higher the FLOPs, the greater the computational complexity and execution time.

RESULT

Transfer learning effects of different models

Figure 2 shows the evaluation results of the CNN LSTM BiLSTM single-mode model and their bimodal combination. It can be seen that CNN performs best with an accuracy of 96.35%, but its parameter count is also high, at 0.478M; in contrast, the accuracy of CNN-BiLSTM is 76.61%, which is nearly 15% higher than that of CNN-LSTM, which also uses bimodal, and the parameter count is only 0.17M, but the required computational effort is significantly increased, with FLOPs of about 522M, which is almost 18 times that of CNN (30.22M FLOPs). This shows that the fusion of convolutional spatial feature extraction and bidirectional temporal modeling can significantly enhance the performance of dual-modal feature extraction, but it is also accompanied by higher computational overhead.

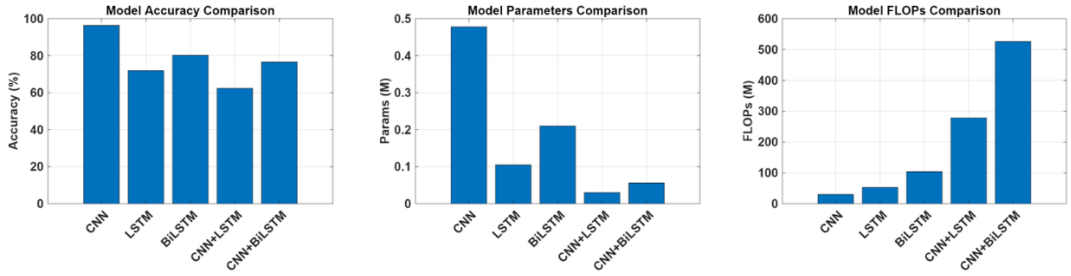

FIGURE 2. Evaluation of CNN LSTM BiLSTM single-mode models and their bimodal combinations (Photocredit : Original)

The impact of different training rounds on model performance

Figure 3 shows the changes in the accuracy and training loss of the model as the number of training epochs increases when the learning rate is 10^{-3}. It can be seen that the model has achieved overall convergence in Figure 3. As the training continues, the accuracy and loss tend to be stable, and the volatility continues to decrease. Figure 4 also uses accuracy and training loss to describe the performance changes of the model in different epochs. Although the overall changes are volatile, it can still be seen that the best performance period (the best performance is determined by high accuracy and low loss here) often occurs between 125 and 150 training epochs, rather than above 150.

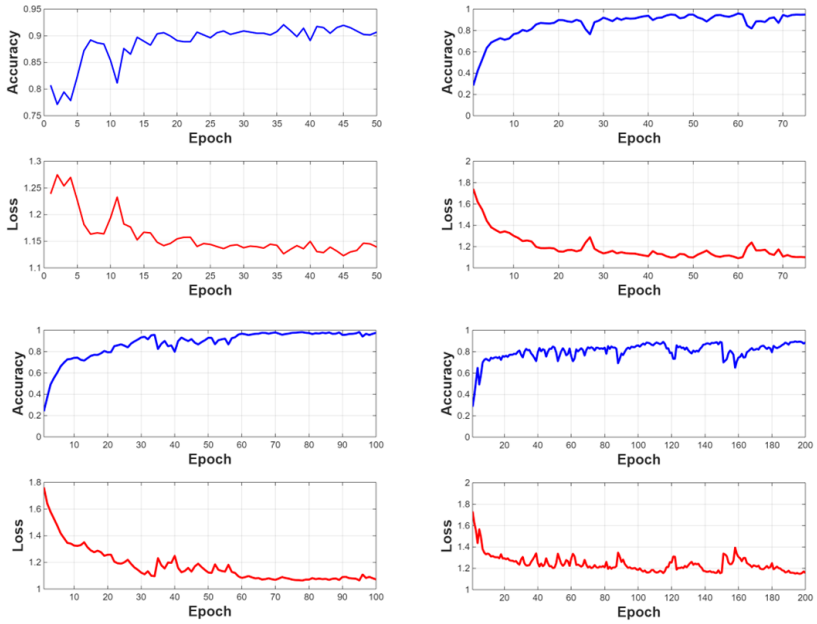

FIGURE 3. Accuracy and training loss at different training epochs (Photocredit : Original)

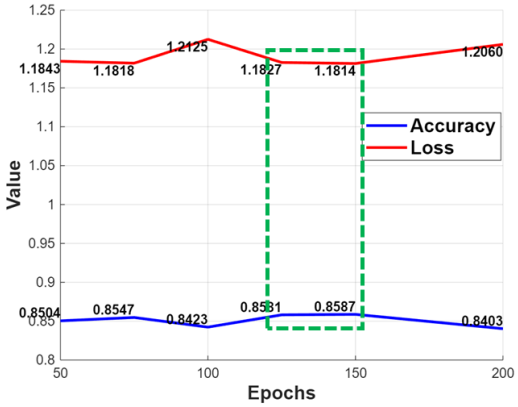

FIGURE 4. Mean of model accuracy and training loss at different training epochs (Photocredit : Original)

The role of fine-tuning

Figure 5 shows the accuracy and loss curves of the CNN-BiLSTM transfer learning process. When the learning rate was 5×10^{-4} and the training was performed for 150 rounds, the model achieved the best pre-training effect. Subsequently, by fixing the CNN parameters and fine-tuning, an accuracy of 76.61% was finally achieved on the test set. The accuracy and loss of the CNN-BiLSTM model show a smoother curve in Figure 6 with smaller fluctuations. This shows that the feature extractor of the model can be transferred between similar tasks (such as NTU-Fi HAR and NTU-Fi Human-ID), and can reduce the volatility of the training loss and the convergence speed. The volatility of the training loss is caused by the feature extractor, so the transfer learning that only trains the classifier performs more smoothly. This is the same result obtained in [4].

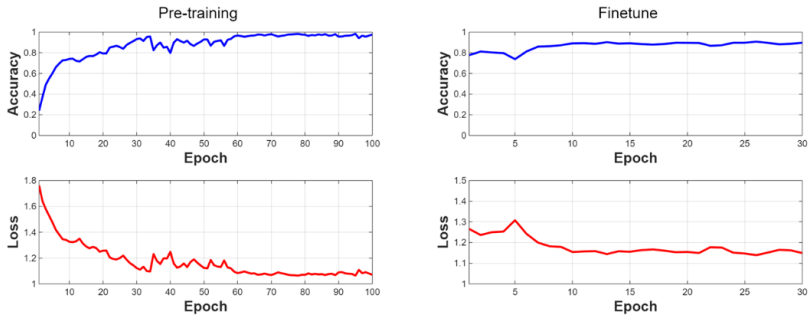

FIGURE 5. CNN-BiLSTM transfer learning process accuracy and loss curve (Photocredit : Original)

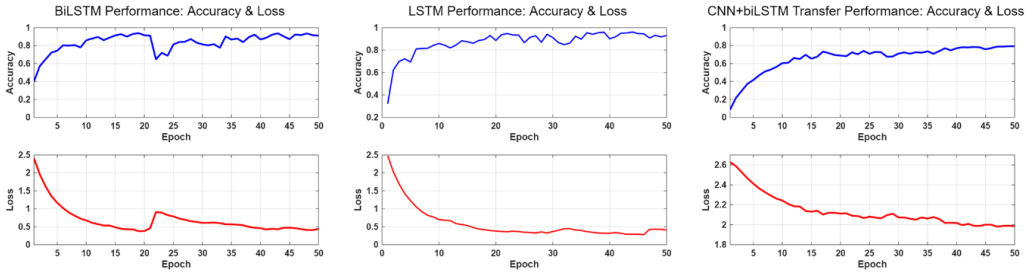

FIGURE 6. Performance curves of the CNN-BiLSTM model and two other single-mode models (Photocredit : Original)

Exploring the reasons for excessive floating point operations

The computational complexity of CNN-BiLSTM shown in Figure 2 is very large. The researchers summarized the following reasons, as shown in Table 2:

TABLE 2. The main factors affecting FLOPs growth

Influencing factors	Impact on FLOPs
Number of CNN layers (L)	Linear growth O(L)
CNN convolution kernel size (K)	Quadratic growth O(K^2)
Number of CNN channels (C)	Quadratic growth O(C^2)
Number of LSTM hidden units (H)	Quadratic growth O(H^2)
Number of LSTM layers (L)	Linear growth O(L)
Bidirectionality	FLOPs × 2
Batch Size (B)	Linear growth O(B)

Since this experiment uses a pre-trained model and a fixed dataset, the number of layers, convolution kernel size, number of channels of the CNN part, and the bidirectionality of the BiLSTM have been determined; while the number of hidden units of the LSTM, the number of LSTM layers, and the batch size can be adjusted according to demand. Although the FLOPs of CNN (e.g. 28.23M) and the FLOPs of BiLSTM (e.g. 105.81M) are relatively fixed in terms of their respective computational load, in the combined model, the total FLOPs are much greater than the result of simply adding the two together. This is mainly due to the influence of the nonlinear superposition effect: after data rearrangement and preprocessing, the output of the CNN will be converted into a longer or higher-dimensional sequence, so that the BiLSTM needs to perform calculations on more time steps, and the bidirectional mechanism further multiplies the amount of calculation, thereby significantly increasing the FLOPs of the entire model.

The transfer effect of the bimodal model on NTU-Fi-HumanID

As can be seen from Table 3, CNN-BiLSTM is significantly better than other models in terms of accuracy (76.61%), but its parameter count is as high as 525.81M, much higher than CNN-LSTM (144.04M) and Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) (48.39M), and its FLOPs is 0.056M; in contrast, although CNN-LSTM has an accuracy of only 62.32%, its parameter count and FLOPs are lower; CNN-GRU's accuracy (53.88%) is relatively low in this result, and its parameter count and FLOPs are 48.39M and 0.059M respectively; ResNet-GRU's accuracy reaches 71.79%, its parameter count is 206.41M, and its FLOPs is relatively high (3.265M). In the transfer learning of bimodal tasks, CNN-BiLSTM can effectively extract spatiotemporal features: CNN is responsible for capturing local spatial patterns, and BiLSTM can learn bidirectional temporal dependencies, thereby achieving higher accuracy in scenarios similar to the source domain or with sufficient data. The low number of parameters allows it to reduce the requirements for hardware resources and data size. However, compared with other bimodal models, the extremely high floating-point operations lead to a significant increase in its computational workload, resulting in longer output time.

TABLE 3. Effect of transfer learning of bimodal model on NTU-Fi-HumanID

Model	Accuracy (%)	Params (M)	FLOPs (M)	Computation time (s/test sample)
CNN-BiLSTM	76.61	525.81	0.056	0.68
CNN-LSTM	62.32	144.04	0.030	0.17
CNN-GRU	53.88	48.39	0.059	0.07
ResNet18-GRU	71.79	206.41	3.265	0.69

Bimodal model pre-training loss

Although the four models in Figure 7 eventually converge, there are slight differences in their convergence behaviors. Overall, the loss of the CNN-BiLSTM model decreases most stably, while the GRU-based models exhibit less fluctuation in loss compared to the LSTM-based models. This can be attributed to the fact that GRU generally has fewer parameters and a simpler gating structure than LSTM, which results in smoother gradient updates during backpropagation. Moreover, within the GRU series, the CNN-GRU converges faster than the ResNet18-GRU, possibly because the CNN architecture is typically simpler, shallower, and contains fewer parameters, allowing the training loss to drop more quickly. In contrast, ResNet18, being a deeper network with a more complex structure and a larger number of parameters, may experience a relatively slower convergence process. Therefore, in training, a deeper network does not necessarily yield better performance.

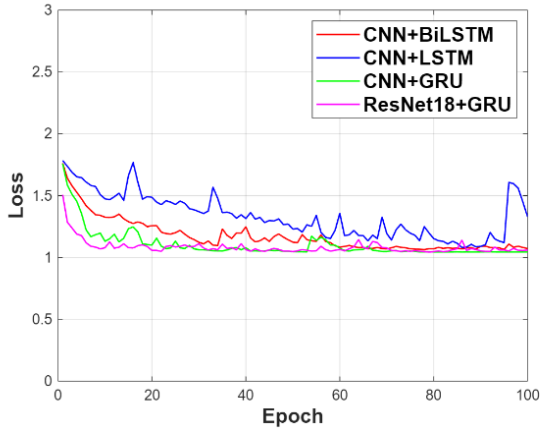

FIGURE 7. Pre-training loss for bimodal models (Photocredit : Original)

CONCLUSION

This paper investigates methods for human activity and identity recognition based on WiFi CSI data, proposing and implementing a transfer learning model that integrates CNN and bidirectional LSTM. The following conclusions were drawn from the study:

a) The performance of the model does not necessarily improve with deeper architectures. The CNN-BiLSTM model achieves the best generalization performance among the bimodal models and is highly hardware-friendly. The only drawback is its larger computational cost, which may lead to longer training times.

b) In comparison, unimodal models such as CNN and BiLSTM are more recommended in the transfer learning setting. This is because, in terms of computational cost, hardware requirements, accuracy, and generalization on other models, they outperform the bimodal models.

Due to the inherent complexity of bimodal models like CNN-BiLSTM, achieving model lightweight by only adjusting hyperparameters yields minimal improvements. Therefore, a more in-depth research is needed for lightweight bimodal models. Possible directions include introducing algorithms to replace the hidden layers of the model or reducing the feature extraction burden by cleaning the CSI data.

REFERENCES

1. W. Liang et al., “LiWi-HAR: Lightweight WiFi-Based Human Activity Recognition Using Distributed AIoT,” in IEEE Internet of Things Journal, vol. 11, no. 1, pp. 597–611, 1 Jan. 2024, doi: 10.1109/JIOT.2023.3286455.
2. J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, “AutoFi: Toward Automatic Wi-Fi Human Sensing via Geometric Self-Supervised Learning,” in IEEE Internet of Things Journal, vol. 10, no. 8, pp. 7416–7425, 15 April 2023, doi: 10.1109/JIOT.2022.3228820.
3. J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, “EfficientFi: Toward Large-Scale Lightweight WiFi Sensing via CSI Compression,” in IEEE Internet of Things Journal, vol. 9, no. 15, pp. 13086–13095, 1 Aug. 2022, doi: 10.1109/JIOT.2021.3139958.
4. J. Yang, X. Chen, H. Zou, C. X. Lu, D. Wang, S. Sun, and L. Xie, “SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing,” Patterns, vol. 4, no. 3, 100703, 2023, doi: 10.1016/j.patter.2023.100703.
5. Z. Shi, Q. Cheng, J. A. Zhang, and R. Y. D. Xu, “Environment-Robust WiFi-Based Human Activity Recognition Using Enhanced CSI and Deep Learning,” in IEEE Internet of Things Journal, vol. 9, no. 24, pp. 24643–24654, 15 Dec. 2022, doi: 10.1109/JIOT.2022.3192973.
6. D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, “CAUTION: A Robust WiFi-Based Human Authentication System via Few-Shot Open-Set Recognition,” in IEEE Internet of Things Journal, vol. 9, no. 18, pp. 17323–17333, 15 Sept. 2022, doi: 10.1109/JIOT.2022.3156099.
7. Y. Zhang et al., “Widar3.0: Zero-Effort Cross-Domain Gesture Recognition With Wi-Fi,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 8671–8688, 1 Nov. 2022, doi: 10.1109/TPAMI.2021.3105387.
8. J. Hu, F. Ge, X. Cao, and Z. Yang, “RGANet: A Human Activity Recognition Model for Extracting Temporal and Spatial Features from WiFi Channel State Information,” Sensors, vol. 25, no. 3, 918, 2025, doi: 10.3390/s25030918.
9. I. A. Showmik, T. F. Sanam, and H. Imtiaz, “Human Activity Recognition from Wi-Fi CSI data using Principal Component based Wavelet CNN,” Digital Signal Processing, vol. 138, 104056, 2023, doi: 10.1016/j.dsp.2023.104056.
10. M. G. Moghaddam, A. A. N. Shirehjini, and S. Shirmohammadi, “A WiFi-Based Method for Recognizing Fine-Grained Multiple-Subject Human Activities,” in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–13, 2023, Art no. 2520313, doi: 10.1109/TIM.2023.3289547.