Regular Article

Enhancing FMCW Radar-Based Human Activity Recognition using DI-ResNet

NgocBinh Nguyen[1], MinhNghia Pham[2], HuuCuong Thieu[1]

[1] Faculty of Telecommunications, Telecommunications University, Khanh Hoa, Vietnam
[2] Faculty of Radio Electronics Engineering, Le Quy Don Technical University, Hanoi, Vietnam

Correspondence: MinhNghia Pham, nghiapmmta@gmail.com
Communication: received 19 January 2026, revised 08 April 2026, accepted 13 May 2026
Online publication: 19 May 2026, Digital Object Identifier: 10.21553/rev-jec.433

Abstract– **This paper proposes a method to enhance the quality of Human Activity Recognition (HAR) based on Frequency Modulated Continuous Wave (FMCW) radar. The method utilizes a Dual Input-ResNet (DI-ResNet) model to address the limitations associated with relying solely on isolated range or micro-Doppler (m-D) features for activity classification. Specifically, two parallel ResNet-18 backbones are employed to extract deep semantic features from two distinct data domains: the Range-Time (RT) spectrogram and the Doppler-Time (DT) spectrogram. These features are subsequently fused via a concatenation layer to synthesize global context, thereby generating a more comprehensive representation of the performed activities. Experimental results demonstrate that the proposed method achieves superior recognition performance compared to single-stream baseline networks. Notably, the proposed model effectively mitigates confusion among activities exhibiting kinematic similarity, improving the Recall metric for the "SitDown" activity by 42.1% compared to traditional single-input methods. Furthermore, the accuracy for complex hand gestures such as "Drink" significantly increased by 19.8%, substantially minimizing the high misclassification rate associated with "Grab." Finally, the model achieves near-ideal reliability for safety-critical activities, attaining a recognition accuracy of 99.5% in fall detection, thereby confirming its potential for practical deployment.**

Keywords– **FMCW, range-time, Doppler-time, dual-input.**

1 Introduction

Nowadays, HAR has emerged as a research area of significant interest, offering practical applications across various industrial and societal sectors, including elderly healthcare, intelligent security surveillance, and smart home systems [1]. Conventional HAR approaches typically rely on wearable sensors [2] or camera-based systems [3]. However, both approaches exhibit significant limitations: wearable sensors can be intrusive and are constrained by battery life, while vision-based systems are susceptible to lighting conditions, occlusions, and, notably, raise serious privacy concerns [4]. To address these limitations, radar sensors have emerged as a superior alternative, leveraging distinct advantages such as non-contact sensing capabilities, robust operation in challenging environmental conditions (e.g., through-wall sensing or in darkness), and, crucially, the preservation of personal privacy [5]. Furthermore, advancements in FMCW radar operating in the millimeter-wave (mmWave) band have enabled the development of compact, low-power, and high-resolution radar systems, making them ideal tools for residential and civil HAR applications [6, 7].

To facilitate HAR, the extraction of 2-D spectrograms from the FMCW radar dataset offers a comprehensive and intuitive representation of information. The DT spectrogram visualizes m-D signatures, which are frequency variations induced by the micro-motion, thereby generating a unique signature for each specific activity [8, 9]. Furthermore, the RT spectrogram provides valuable information regarding the target's positional changes over time [10]. These 2-D spectrograms can be processed either as 2-D images [11] or as time-series data [12] to be fed into classification models.

Initially, researchers approached the classification problem by manually extracting statistical or geometric features from spectrograms, such as the Doppler centroid, bandwidth, or components derived from Singular Value Decomposition (SVD) [10]. These features were subsequently fed into traditional machine learning classifiers, such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) [13]. However, these methods require substantial domain expertise for key feature selection, and classification performance degrades significantly when faced with complex activities or noisy environments. The limitations associated with hand-crafted feature extraction have prompted a significant paradigm shift toward Deep Learning (DL) models, which can automatically learn features directly from the input dataset. Convolutional Neural Networks (CNNs) rapidly emerged as the standard by treating 2-D spectrograms as input images. Studies [14, 15] have demonstrated that CNNs can effectively learn m-D features, outperforming traditional approaches. Subsequently, customized deep learning architectures, such as DINN [16], SRCNN [17], and CRCNN [18] have been proposed to be better optimized for extracting m-D features from input spectrograms, thereby achieving high accuracy.

Despite the successes achieved by DL, a critical research gap remains. The majority of state-of-the-art DL methods predominantly focus on DT spectrograms. While effective, this approach inadvertently neglects a valuable and complementary source of information: the RT spectrogram [10]. Information derived from the RT spectrogram encapsulates critical features regarding spatial displacement and body trajectory, which are essential for distinguishing activities exhibiting similar m-D signatures (e.g., "soft fall" versus "sit down"). The absence of this spatial information also renders the model susceptible to confusion caused by noise and multipath effects. Although several studies have attempted to integrate multi-domain information [19, 20], they typically employ late fusion strategies, thereby missing the opportunity for the network to learn complex inter-domain correlations at an early stage. To address this gap, this paper proposes a novel method aimed at enhancing classification accuracy by simultaneously exploiting information from both DT and RT spectrograms. The main contributions of this paper are summarized as follows:

- Proposal of the Dual-Input-ResNet (DI-ResNet) model for HAR based on FMCW radar dataset. The DI-ResNet model simultaneously utilizes features extracted from two independent inputs: (1) the RT spectrogram to capture spatial displacement features, and (2) the DT spectrogram to capture m-D features.
- Experimental validation of the proposed model's effectiveness using a publicly available real-world dataset.

The remainder of this paper is organized as follows: Section 2 details the FMCW radar signal processing workflow and describes the dataset. Section 3 describes the proposed DI-ResNet model. Section 4 presents the experimental results and discussion. Finally, Section 5 provides the conclusion and suggests directions for future research.

2 FMCW Radar Data Collection and Dataset Description

2.1 FMCW Radar Data Collection

The data collection and processing workflow of the FMCW radar is illustrated in Figure 1. The FMCW radar transmits chirp signals via the transmit antenna (TX). The backscattered signal from the target is captured by the receiving antenna (RX) and mixed with the transmitted waveform to generate the Intermediate Frequency (IF) signal. This analog signal is subsequently digitized by an Analog-to-Digital Converter (ADC) to form the raw radar data block.

To transform the raw data into a meaningful 2-D time-frequency representation, a signal processing

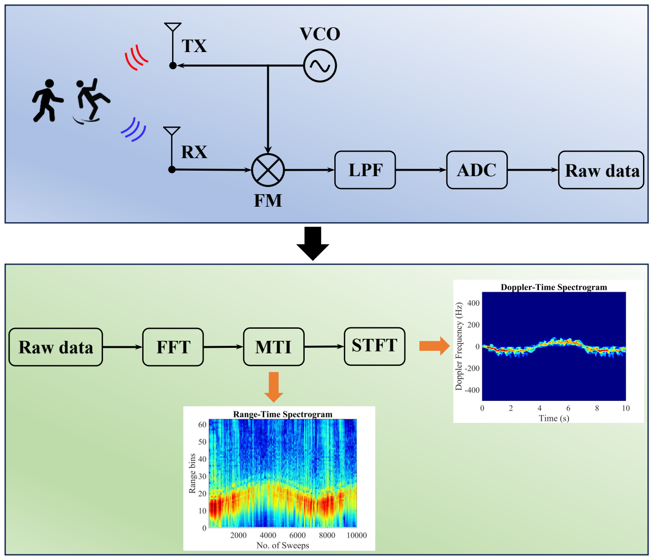


Figure 1. FMCW radar data collection and processing procedures.

pipeline is employed. Initially, the Fast Fourier Transform (FFT) algorithm is performed on the sampling points to generate the RT spectrogram, effectively resolving the target range. A Moving Target Indication (MTI) filter with a specific cut-off frequency is applied to the range-time matrix to eliminate reflections from stationary objects (static clutter). Subsequently, the Short-Time Fourier Transform (STFT) is utilized by employing a sequence of FFTs with short, overlapping sliding windows across the entire duration of the processed range-time matrix to obtain the time-frequency spectrum of the input IQ signal. Finally, the spectrogram is defined as follows [21]

$$\begin{aligned} S[n,k] &= |X[n,k]|^2 \\ &= \left| \sum_{m=0}^{M-1} x[m]h[m-n] \exp\left(-j\frac{2\pi km}{N}\right) \right|^2, \end{aligned} \quad (1)$$

where $S[n,k]$ represents spectrum energy at time index n and frequency index k, $h[m-n]$ denotes the window function.

2.2 Experimental Dataset Description

The experimental dataset employed in this study is the University of Glasgow dataset [22], which is publicly shared for scientific research purposes. The data was acquired using an FMCW radar operating at a carrier frequency of 5.8 GHz (C-band). The dataset comprises 1,632 samples collected from 72 participants performing six distinct activities of daily living: Drink, Fall, Grab, SitDown, StandUp, and Walk.

The raw data is processed following the workflow presented in Figure 1 to generate 2-D RT and DT spectrograms. Figure 2 and Figure 3 illustrate the RT spectrograms and DT spectrograms, respectively, corresponding to the different activities.

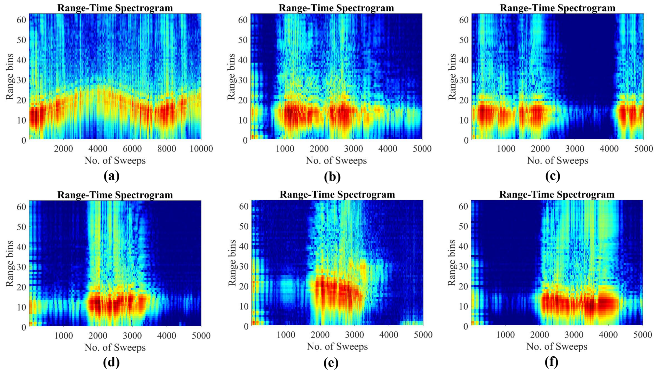


Figure 2. RT spectrograms of different activities: (a) Walk, (b) Grab, (c) Drink, (d) SitDown, (e) StandUp, (f) Fall.

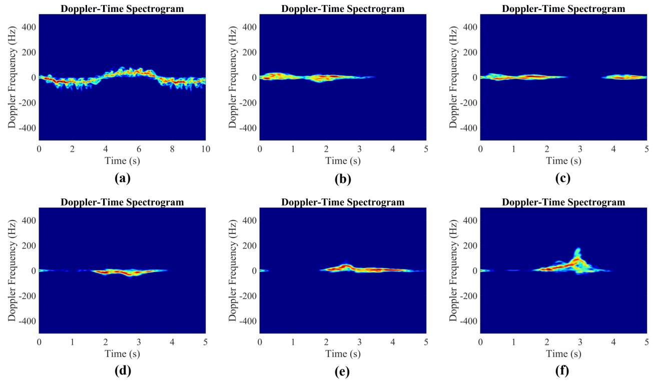


Figure 3. DT spectrograms of different activities: (a) Walk, (b) Grab, (c) Drink, (d) SitDown, (e) StandUp, (f) Fall.

3 Proposed DI-ResNet Model

3.1 Overview of the DI-ResNet

The overall architecture of the proposed DI-ResNet model is illustrated in Figure 4. The DI-ResNet is designed as a parallel CNN, utilizing ResNet-18 as the backbone network [23]. The primary objective of this design is to extract features from two heterogeneous data domains: (1) the RT spectrogram, which represents the target's positional and range variations over time; and (2) the DT spectrogram, which characterizes the m-D frequency shifts induced by the target's micro-motions. Instead of stacking these two data modalities into the channels of a single input, the model processes them via two distinct branches to extract high-level features before fusion.

3.2 Design rationale of DI-ResNet

3.2.1 Selection Backbone is ResNet: The proposed model employs the Residual architecture to mitigate the vanishing gradient problem, a common challenge in deep CNN structures. The output of a residual block (conv-unit 2) is mathematically formulated as

$$y = \sigma(\mathcal{F}(x, \{W_i\}) + W_s x), \tag{2}$$

where:

- x and y denote the input and output vectors of the block, respectively

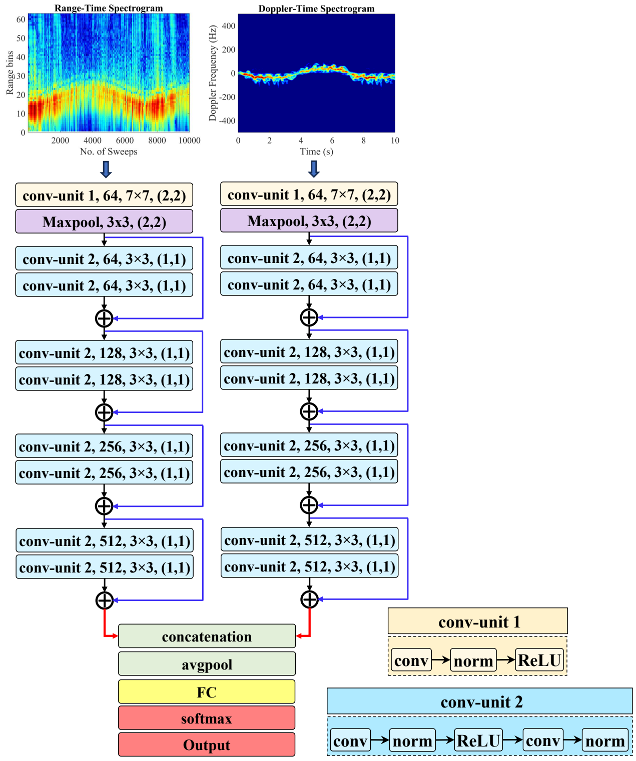


Figure 4. Proposed DI-ResNet model structure.

- $\mathcal{F}(x, \{W_i\})$ represents the residual mapping function (implemented via the conv layers shown in the Figure 4)
- W_s is the linear projection matrix, utilized when the dimensions of the input and output differ (upscaling from 64 filters to 128 filters) to ensure dimension matching)
- σ denotes the non-linear activation function (ReLU)
- The addition operation $(\mathcal{F} + W_s x)$ constitutes the shortcut connection (indicated by the blue arrow in the diagram)

3.2.2 Parallel Feature Extraction: The proposed model utilizes two separate ResNet-18 branches to extract distinct features from the two heterogeneous input data types:

Branch 1 (Left): Learns the mapping function $f_{range} : X_R \to V_R$, where the output V_R is a feature vector encapsulating trajectory information.

Branch 2 (Right): Learns the mapping function $f_{Doppler} : X_D \to V_D$, where the output V_D is a feature vector encapsulating detailed micro-motion information.

3.2.3 Feature Fusion Mechanism: A critical component of the proposed model is the Concatenation layer (highlighted in light green). Let T_R and T_D be the feature tensors obtained at the last layer before the average pooling (avgpool) layer, with dimensions $(C \times H' \times W')$. The fusion process is mathematically described as

$$V_{fused} = \text{Concat}(T_R, T_D). \tag{3}$$

Subsequently, V_{fused} passes through avgpool layer to generate the representative feature vector

$$z = \frac{1}{H' \times W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} V_{fused}(:, i, j). \quad (4)$$

Finally, the vector z is fed into a Fully Connected (FC) layer and a softmax function to compute the classification probabilities

$$P(y = k|x) = \frac{e^{w_k^T z + b_k}}{\sum_{j=1}^{K} e^{w_j^T z + b_j}}. \quad (5)$$

This combination creates a vector hyperplane capable of better data separation. Specifically, for the "Sit-Down" activity: V_R contributes information regarding height reduction, while V_D confirms negative velocity (downward motion). In addition, the combination $z = [V_R, V_D]$ establishes a double confirmation mechanism, effectively eliminating confusion with "Fall" or "Walk". With the "Grab" and "Drink" activities, which exhibit highly similar V_R (hand trajectories), the V_D (micro-motions of the wrist/fingers) remains distinct. The FC network effectively learns the weights w to "attend" more significantly to the V_D component within the z vector when distinguishing between these two classes.

4 Experimental Results and Discussion

4.1 Experimental Setup

The experimental environment established for the radar data pre-processing and model training phases comprises specific hardware and software components. In terms of hardware, the workstation is equipped with an NVIDIA GeForce RTX 3060 Ti graphics card, an Intel(R) Core™ i5-12400F CPU (2.5 GHz), and 32 GB of RAM to facilitate computational tasks. Regarding the software environment, all experiments were implemented using MATLAB R2023a.

The models were trained for 20 epochs with a batch size of 16. The initial learning rate was set to 0.01 and reduced by a factor of 10 every 4 epochs. The Stochastic Gradient Descent (SGD) optimizer was employed during the training process to update the model weights. Furthermore, a five-fold cross-validation scheme was implemented to validate and evaluate the performance of the proposed model.

4.2 Comparision Results and Discussion

To evaluate the efficacy of the proposed DI-ResNet architecture, a two-stage comparative analysis was conducted. Initially, the comparison between the DI-ResNet and a baseline DI-SimpleNet (a truncated architecture comprising two convolutional layers and two max-pooling layers) was performed to verify the overall effectiveness of the multi-input proposal for feature representation. Subsequently, the performance of the DI-ResNet model (utilizing both RT and DT spectrograms) was benchmarked against the conventional ResNet18 model trained on a single input (RT-only and DT-only). The quantitative assessment was carried out based on five primary metrics: accuracy, confusion matrix, precision, recall, and the F1-score. All experimental results were collected using the same experimental dataset described in Section 2.2.

Table I summarizes the classification performance across various input configurations. The results indicate that the recognition accuracy of the DT spectrograms is consistently higher than that of the RT spectrograms, achieving 90.3% versus 78.9% for ResNet and 75.2% versus 62.6% for SimpleNet. Furthermore, ResNet outperforms the SimpleNet baseline across all tests, highlighting the efficacy of deeper architectures in extracting complex semantic features. Notably, the dual-input fusion strategy significantly enhances performance; the proposed DI-ResNet achieves the highest accuracy of 93.8%, while DI-SimpleNet reaches 83.7% (representing gains of 21.1% and 8.5% over their respective single-input counterparts). These findings demonstrate that DT spectrograms exhibit higher reliability compared to RT spectrograms, and the fusion of RT and DT features effectively compensates for the limitations inherent in single-feature recognition.

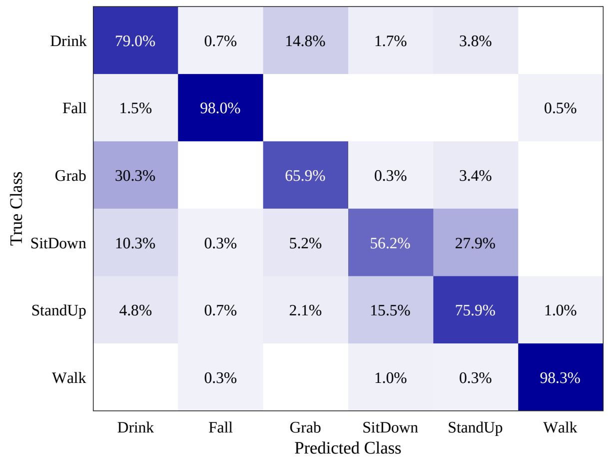
Table I
The comparison of average accuracy results

Model	Input Data	Accuracy
SimpleNet	RT	62.6
SimpleNet	DT	75.2
SimpleNet	RT, DT	83.7
ResNet [23]	RT	78.9
ResNet [23]	DT	90.3
DI-ResNet	RT, DT	93.8

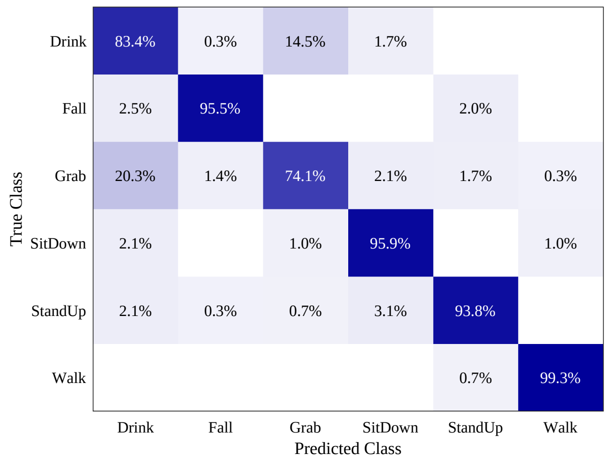
To provide a deeper analysis of the classification accuracy and misclassification rates for individual activities across the experimental scenarios, confusion matrices were computed for the ResNet model using RT spectrograms, the ResNet model using DT spectrograms, and the DI-ResNet model utilizing the combination of both RT and DT spectrograms.

The results from Figure 5a and Figure 5b indicate that the recognition accuracy for in-place activities such as Grab, SitDown, StandUp, and Drink is subject to significant variation depending on the input data modality. Specifically, the RT spectrograms yielded the lowest accuracy for SitDown (achieving only 56.2%) and Grab (65.9%). Due to the symmetry in the trajectory of the body center of mass, 27.9% of SitDown samples were misclassified as StandUp. Conversely, the DT spectrograms demonstrated superior performance for the SitDown activity (95.9%) by effectively exploiting m-D features; however, it encountered difficulties in distinguishing between Grab and Drink (resulting in a 20.3% misclassification rate). Nevertheless, the fact that RT recognized "Fall" more accurately than DT (2.5% higher) emphasizes the mutually complementary features of the two data domains in covering the full spectrograms of the six target activities.

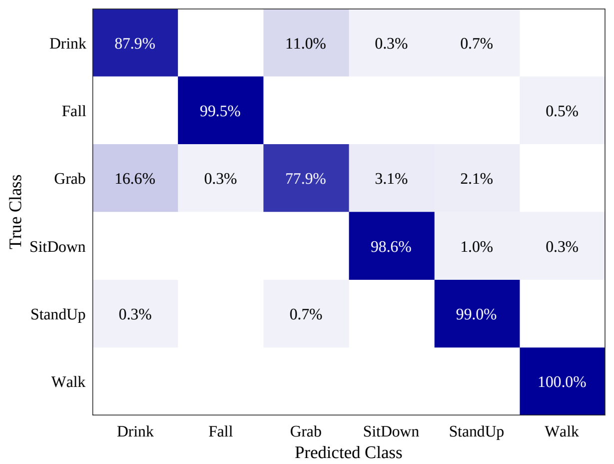
As shown in Figure 5c, the multi-domain fusion



(a) The ResNet model using RT spectrograms



(b) The ResNet model using DT spectrograms



(c) The DI-ResNet model using RT and DT spectrograms

Figure 5. Confusion matrixes of different models.

strategy demonstrates comprehensive superiority, particularly for the SitDown, Fall, and Walk classes. The incorporation of positional information from the RT spectrograms helps mitigate the blind spots of single-stream models, boosting the accuracy of SitDown from 56.2% to 98.6% and Grab from 65.9% to 77.9%. For the pair of activities exhibiting high kinematic similarity, such as Grab and Drink, the multi-input model reduced the cross-misclassification rate from 20.3% (using DT spectrogram) to 16.6%, while simultaneously increasing the accuracy of Drink to a peak of 87.9%.

Notably, for safety-critical activities such as "Fall," the fusion method achieved near-ideal accuracy (99.5%), effectively eliminating the confusion observed in the RT spectrograms. Although the overlap of kinematic features between Grab and Drink still induces a marginal error rate, the experimental results confirm that the multi-domain model has substantially improved the system's reliability and overall performance compared to single-domain baselines.

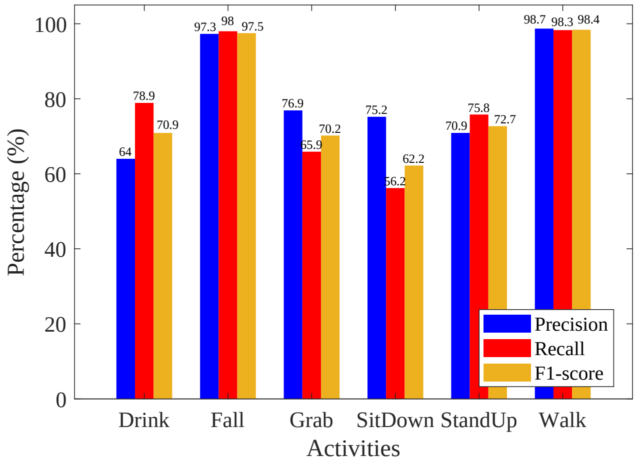
Figure 6 illustrates the Precision, Recall, and F1-score metrics for each activity classified using three distinct neural network architectures. Experimental results indicate that the DI-ResNet model achieves the highest Recall metric in 5 out of 6 activity groups: Drink, Fall, SitDown, StandUp, and Walk. The superiority is most evident in the SitDown activity: while using RT spectrogram attained only 56.2% due to the inherent limitations of range-only data, the incorporation of Doppler features in the DI-ResNet elevated this metric to 98.3% (an improvement of 42.1%). Notably, the proposed model achieves near-perfect performance for Walk (100%) and Fall (99.5%), significantly outperforming the two single-domain models.

Regarding the Precision metric, using RT spectrograms reveals a significant weakness in the "Drink" activity (achieving only 64%), indicating a high rate of confusion with other activities such as "Grab." Although using DT spectrograms improves this figure to 77.2%, it is the fused DI-ResNet model that truly optimizes the precision for "Drink" to 83.8% (an increase of 19.8% compared to RT). For high-risk activities such as "Fall," the Precision reaches 99.5%, ensuring maximum reliability for safety alerts.

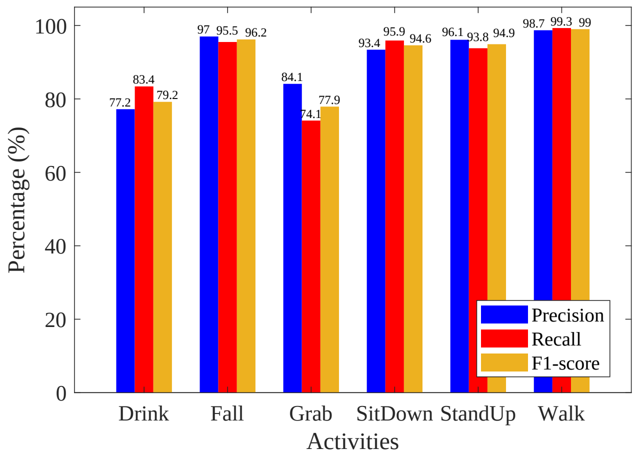
In terms of the F1-Score, DI-ResNet is the sole model to achieve an optimal balance, surpassing a threshold of 82% across all classes. For complex hand manipulation activities (Grab, Drink) and posture transition activities (SitDown), the RT model yielded significantly low results (70.2%, 70.9%, and 62.2%, respectively). The DI-ResNet model not only improved the F1-Score for Grab to 82.8% (increasing by 5% when using DT spectrogram) but also achieved a substantial leap for SitDown, reaching 97.4% (an increase of 35.2% compared to using RT spectrogram). This demonstrates that the multi-domain fusion strategy has effectively resolved local weaknesses, yielding a robust and consistent classifier across the entire dataset.

5 Conclusion

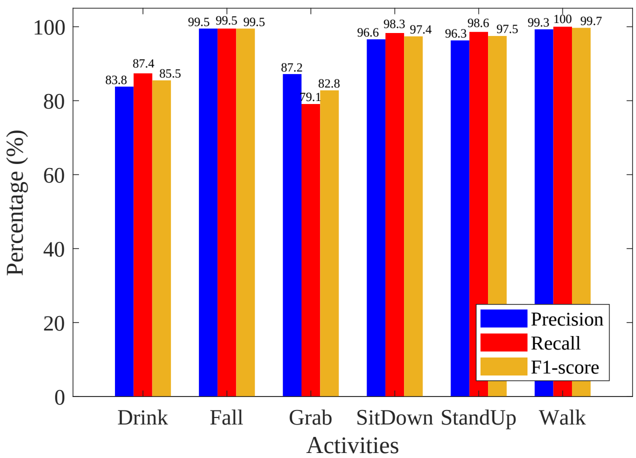
This paper presents an improved multi-input model based on FMCW radar sensors designed to enhance HAR accuracy by fusing the Range-Time (RT) and Doppler-Time (DT) domains of human activities. The proposed DI-ResNet model was evaluated on the University of Glasgow dataset, and experimental results demonstrate its effectiveness, achieving an average accuracy improvement of 3.5% compared to single-input



(a) The ResNet model using RT spectrograms



(b) The ResNet model using DT spectrograms



(c) The DI-ResNet model using RT and DT spectrograms

Figure 6. The Precision, Recall and F1-score of different models.

models. Specifically, the classification accuracy for all activities improved by a margin of 3% to 6%, with Fall detection notably achieving an accuracy of 99.5%. In conclusion, by integrating multi-domain features from RT and DT maps, the proposed model achieves high and robust accuracy for radar-based HAR, confirming its strong potential for deployment in real-world surveillance applications. Future work will address the classification of fine-grained gestures characterized by significant kinematic overlap and evaluate model generalizability in non-idealized environments subject to multipath effects and noise.

References

[1] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Activities recognition and fall detection in continuous data streams using radar sensor," in *Proceedings of the 2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, vol. 1. IEEE, 2019, pp. 1–4.

[2] T. R. Bennett, J. Wu, N. Kehtarnavaz, and R. Jafari, "Inertial measurement unit-based wearable computers for assisted living applications: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 28–35, 2016.

[3] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[4] R. Igual, C. Medrano, and I. Plaza, "Challenges, issues and trends in fall detection systems," *Biomedical engineering online*, vol. 12, no. 1, p. 66, 2013.

[5] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 2, pp. 655–665, 2017.

[6] M. Chakraborty, H. C. Kumawat, S. V. Dhavale *et al.*, "DIAT-RadHARNet: A lightweight DCNN for radar based classification of human suspicious activities," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[7] W. Taylor, K. Dashtipour, S. A. Shah, A. Hussain, Q. H. Abbasi, and M. A. Imran, "Radar sensing for activity classification in elderly people exploiting micro-doppler signatures using machine learning," *Sensors*, vol. 21, no. 11, p. 3881, 2021.

[8] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE transactions on geoscience and remote sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.

[9] V. C. L. Ho and H. Wechsler, "Micro-Doppler effect in radar: Phenomenon model and simulation study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 2–21, 2006.

[10] Z. Li, F. Fioranelli, S. Yang, L. Zhang, O. Romain, Q. He, G. Cui, and J. Le Kernec, "Multi-domains based human activity classification in radar," in *Proceedings of the IET International Radar Conference (IET IRC 2020)*, vol. 2020. IET, 2020, pp. 1744–1749.

[11] U. Saeed, S. Y. Shah, S. A. Shah, J. Ahmad, A. A. Alotaibi, T. Althobaiti, N. Ramzan, A. Alomainy, and Q. H. Abbasi, "Discrete human activity recognition and fall detection by combining FMCW RADAR data of heterogeneous environments for independent assistive living," *Electronics*, vol. 10, no. 18, p. 2237, 2021.

[12] J. Zhu, H. Chen, and W. Ye, "A hybrid CNN–LSTM network for the classification of human activities based on micro-Doppler radar," *IEEE Access*, vol. 8, pp. 24 713–24 720, 2020.

[13] S. A. Shah and F. Fioranelli, "Human activity recognition: Preliminary results for dataset portability using FMCW radar," in *Proceedings of the 2019 International radar Conference (RADAR)*. IEEE, 2019, pp. 1–4.

[14] M. S. Seyfioğlu, S. Z. Gürbüz, A. M. Özbayoğlu, and M. Yüksel, "Deep learning of micro-Doppler features for aided and unaided gait recognition," in *Proceedings of the 2017 IEEE Radar Conference (RadarConf)*. IEEE, 2017, pp. 1125–1130.

[15] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall mo-

tion detection using deep learning," in *Proceedings of the 2016 IEEE Radar Conference*. IEEE, 2016, pp. 1–6.

[16] N. Nguyen, M. Pham, V. Le, D. DuongQuoc, and V.-S. Doan, "Micro-doppler signatures based human activity classification using dense-inception neural network," in *Proceedings of the 2022 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2022, pp. 268–273.

[17] N. Nguyen, V.-S. Doan, M. Pham, and V. Le, "SR-CNN: Stacked-Residual Convolutional Neural Network for Improving Human Activity Classification Based on Micro-Doppler Signatures of FMCW Radar," *Journal of Electromagnetic Engineering and Science*, vol. 24, no. 4, pp. 358–369, 2024.

[18] N. Nguyen, M. Pham, V.-S. Doan, and V. Le, "Improving human activity classification based on micro-doppler signatures of FMCW radar with the effect of noise," *PloS One*, vol. 19, no. 8, p. e0308045, 2024.

[19] W. Ding, X. Guo, and G. Wang, "Radar-based human activity recognition using hybrid neural network model with multidomain fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2889–2898, 2021.

[20] X. Wang, S. Guo, J. Chen, P. Chen, and G. Cui, "GCN-enhanced multidomain fusion network for through-wall human activity recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[21] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," in *Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4, 1997, pp. 21–24.

[22] D. F. Fioranelli, D. S. A. Shah, H. Li1, A. Shrestha, D. S. Yang, and D. J. L. Kernec, "Radar sensing for healthcare: Associate editor francesco fioranelli on the applications of radar in monitoring vital signs and recognising human activity patterns," *Electronics Letters*, vol. 55, no. 19, pp. 1022–1024, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

NgocBinh Nguyen was born in 1989, he received his Bachelor's degree in 2011 from the Telecommunications University, Khanh Hoa, Viet Nam. He obtained his Master's degree in 2017 and Ph.D. degree in 2025 from the Le Quy Don Technical University, Hanoi, Vietnam. He is currently a lecturer at Telecomunication University, Khanhhoa, Vietnam. His current research interests include radar signal processing, image processing, and deep learning.

MinhNghia Pham was born in 1980. He received the B.S., M.S. degrees in electronics engineering from the Le Qui Don Technical University, Vietnam, in 2005 and 2008, respectively, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology, China in 2014. He is currently a lecturer at Le Qui Don Technical University, Vietnam. He currently focuses on polarimetric synthetic aperture radar image processing, radar signal processing, deep learning and signal processing.

HuuCuong Thieu was born in 1986, he received his Bachelor's degree in 2011 from the Telecommunications University, Khanh Hoa, Vietnam. He obtained his Master's degree in 2017 and Ph.D. degree in 2022 from the Le Quy Don Technical University, Hanoi, Vietnam. Currently, he is the Head of the Department in the Faculty of Telecommunications at the Telecommunications University. His current research interests include radar signal processing, image processing, and deep learning.