Research Article GU J Sci, Part A, 13(2): 692-709 (2026) 10.54287/gujsa.1901011

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Comparative Analysis of Time-Frequency Transformations and Deep Learning Architectures in the Classification of FMCW Radar Micro-Doppler Signatures

Harun SEVİNÇ[1*] Levent SEYFİ[1]

[1] Konya Technical University, Konya, Türkiye

Keywords

FMCW Radar
Micro-Doppler Effect
Time-Frequency Analysis
Wigner-Ville Distribution
Deep Learning

Abstract

Frequency Modulated Continuous Wave (FMCW) radars present a robust and privacy-friendly alternative for human activity recognition (HAR) systems. In the classification of micro-Doppler signatures obtained from these radars via deep learning algorithms, the structural quality of the time-frequency (TF) transformation used as input directly affects the model's performance. Therefore, the explicit objective of this study is to design and optimize lightweight Convolutional Neural Network (CNN) architectures capable of efficiently extracting micro-Doppler features from various TF representations under the constraints of limited radar data. To achieve this, signals belonging to 6 different human activities collected by an FMCW radar were converted into spectrograms using Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), and Wigner-Ville Distribution (WVD). Three novel convolutional neural network architectures (CNN–Base, CNN–Wide, CNN–LSTM), capable of providing different responses to the spatial features of these transformations, were designed and analyzed using a 6-fold cross-validation strategy. Experimental results demonstrated that the logarithmic scaling of CWT misled CNN filters, while the standard STFT provided a strong baseline with 80.00% accuracy. The highest performance of the study was achieved at 80.42% with the CNN–Wide architecture, which successfully processed the high-resolution texture of WVD containing cross-terms utilizing its wide receptive field. The WVD-based model identified life-critical falling activities with 100% precision, offering a promising approach for elderly care and autonomous health monitoring systems.

Cite

Sevinç, H., & Seyfi, L. (2026). Comparative Analysis of Time-Frequency Transformations and Deep Learning Architectures in the Classification of FMCW Radar Micro-Doppler Signatures. *GU J Sci, Part A*, *13*(2), 692-709. doi:10.54287/gujsa.1901011

Author ID (ORCID Number)

0009-0006-2408-2540 Harun SEVİNÇ
0000-0002-8698-5140 Levent SEYFİ

Article Process

Submission Date 02.03.2026
Revision Date 06.04.2026
Accepted Date 29.04.2026
Published Date 30.06.2026

1. INTRODUCTION

The detection and classification of human movements, known as Human Activity Recognition (HAR), is of increasing importance in numerous fields such as smart home systems, elderly care, border security, and human-machine interaction (Miao et al., 2025; Wu et al., 2025). Traditionally, optical cameras and wearable sensors have been widely used for this task (Zhang et al., 2022). However, the privacy violations caused by cameras, their susceptibility to lighting changes, and the restricted user comfort associated with wearable devices have directed researchers towards alternative sensing technologies (Faisal et al., 2023). In this context, radar sensors have emerged as one of the most powerful alternatives for HAR applications, thanks to their

*Corresponding Author, e-mail: hsevinc@ktun.edu.tr

ability to operate independently of environmental conditions, their through-wall sensing potential, and their privacy-preserving nature (generating signals instead of images) (Gong et al., 2023).

Frequency Modulated Continuous Wave (FMCW) radar signals produce additional frequency shifts caused by the periodic oscillations of the moving target's appendages, such as arms and legs, in addition to the main torso. This phenomenon, referred to as the Micro-Doppler ($m-D$) effect in the literature, ensures that each human movement (walking, running, falling, etc.) leaves a unique signature on the radar. To extract these signatures from the raw radar data and make them observable, the signal must be transformed from the 1-dimensional time domain into the 2-dimensional time-frequency (TF) domain (Dang et al., 2026).

Today, Deep Learning architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, which excel in modeling the temporal dynamics of signals, demonstrate superior performance in the classification of $m-D$ signatures (Abdu et al., 2022; Ullmann et al., 2023). However, the performance of deep learning models largely depends on the resolution and information content of the 2-dimensional TF image provided as input. A review of the literature reveals that researchers mostly adopt the Short-Time Fourier Transform (STFT), which utilizes a fixed window size, as a standard preprocessing step (Biswas et al., 2024). The inherent trade-off between time and frequency resolution in STFT can lead to information loss and smearing, especially in movements involving sudden frequency changes such as falling or jumping (Wang et al., 2025). Although high-resolution algorithms like the Continuous Wavelet Transform (CWT) and Wigner-Ville Distribution (WVD) theoretically possess the potential to overcome these limitations (Ünal & Pakfiliz, 2022; Mishra & Pachori, 2025), the practical effects of the cross-terms and spatial warping produced by these transformations on deep learning filters have not been sufficiently investigated in a comparative manner (Dang et al., 2026).

Motivated by this, the presented study investigates the relationship between TF transformation algorithms and the spatial capacities of deep learning architectures in the classification of micro-Doppler signatures belonging to human activities obtained by an FMCW radar. The main contributions of the study can be summarized as follows:

- STFT, CWT, and WVD transformations were obtained using the same raw FMCW radar data, and the performance of deep learning models on these three different TF representations was directly compared.
- Three distinct network architectures—lightweight (CNN–Base), wide receptive field (CNN–Wide), and hybrid spatiotemporal (CNN–LSTM)—were designed and optimized to respond differently to structural variations in spectrograms (smearing, cross-terms, warping).

- The generalization capabilities and overfitting tendencies of the models were meticulously analyzed using a 6-Fold Cross-Validation strategy; it was experimentally proven that the CNN–Wide architecture paired with WVD data constitutes the optimal match.

- Furthermore, the proposed custom architectures were benchmarked against standard, widely adopted deep learning models (AlexNet, ResNet50, and MobileNetV2) to empirically demonstrate the critical need for architectural optimization in radar-based HAR tasks under limited data conditions.

2. THEORETICAL BACKGROUND

In this section, the $m-D$ effect, which represents the physical reflections of human movements on the FMCW radar, is explained, and the mathematical foundations of the three primary TF analysis methods that convert 1-dimensional radar time-series signals into 2-dimensional feature maps for deep learning networks are presented.

2.1. Radar Micro-Doppler Effect

In a radar system, electromagnetic waves reflected from a moving target undergo a frequency shift (Doppler shift) depending on the target's radial velocity relative to the radar. However, in complex (non-rigid) targets such as humans, in addition to the translational movement of the target's main torso, the periodic oscillations of the arms and legs generate additional sidebands around the main Doppler frequency (Kim & Seo, 2022). Let $v(t)$ be the instantaneous velocity of the target, f_c the center frequency of the transmitted signal, and c the speed of light. The instantaneous Doppler frequency shift $f_D(t)$ is expressed in (1).

$$f_D(t) = \frac{2f_c}{c} v(t) \tag{1}$$

Since the acceleration and deceleration of the limbs change continuously during movements like human walking or falling, these $f_D(t)$ components create a highly complex and movement-specific $m-D$ signature over time (Yin et al., 2024).

2.2. Time-Frequency Transformations

To observe the time-dependent variation of micro-Doppler signatures, it is necessary to project the signal onto TF plane. In this study, the effects of three different TF transformations on deep learning performance were examined.

2.2.1. Short-Time Fourier Transform

STFT is the most widely used method for analyzing non-stationary signals (Sarna et al., 2024). The signal is divided into segments using a fixed-length sliding window function $w*(t)$, and the Fourier transform is calculated for each segment. The STFT of a time-domain signal $x(t)$ is defined in (2).

$$X_{STFT}(t,f) = \int_{-\infty}^{\infty} x(\tau) w * (\tau - t) e^{-j2\pi f \tau} d\tau \quad (2)$$

The major limitation of STFT is its use of a fixed window size, dictated by the Heisenberg-Gabor uncertainty principle. A narrow window increases time resolution but decreases frequency resolution, while a wide window has the opposite effect. This condition causes smearing in the spectrograms of movements involving sudden frequency changes, such as falling (Wang et al., 2025).

2.2.2. Continuous Wavelet Transform

CWT employs a multi-resolution approach to overcome the fixed resolution problem of STFT. Instead of sinusoidal waves, it uses a mother wavelet function $\psi(t)$ that can be scaled and shifted (Mishra & Pachori, 2025). The mathematical expression of CWT is in (3).

$$X_{CWT}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \quad (3)$$

Here, a is the scale parameter (inversely proportional to frequency), b is the shift (time) parameter, and the $*$ operator denotes the complex conjugate. CWT provides high frequency resolution at low frequencies and high time resolution at high frequencies, thereby capturing the edges of sudden changes more sharply (Avinash et al., 2025). However, this logarithmic scaling can distort the linear spatial structure that deep learning algorithms are accustomed to, leading to deformations in feature maps.

2.2.3. Wigner-Ville Distribution

WVD is a bilinear distribution that theoretically offers the highest time-frequency resolution by taking the Fourier transform of the signal's instantaneous autocorrelation without using any windowing or mother wavelet function (Ünal & Pakfiliz, 2022). It is mathematically defined in (4).

$$W_x(t,f) = \int_{-\infty}^{\infty} x\left(t + \frac{\tau}{2}\right) x^*\left(t - \frac{\tau}{2}\right) e^{-j2\pi f \tau} d\tau \quad (4)$$

WVD perfectly focuses the energy of $m - D$ signals in the time-frequency plane. However, in multi-component signals, due to its mathematical nature, it generates artificial frequency bands called cross-terms between the actual signal components, which have no physical counterpart (Faisal et al., 2023). In this study, the Smoothed Pseudo Wigner-Ville Distribution (SPWVD) variant (Kumawat & Raj, 2022), widely accepted in the literature, was utilized to partially suppress these cross-terms.

2.3. FMCW Radar Operating Principle and System Parameters

While traditional Continuous Wave (CW) radars can only measure the velocity of the target, FMCW radars can measure both the velocity (Doppler) and the range of the target with high precision by transmitting waves whose frequency increases or decreases linearly over time (chirp) (Patole et al., 2017).

When the signal transmitted (Tx) by the FMCW radar hits a target and reflects back (Rx), these two signals are passed through a hardware mixer to obtain a beat signal. The frequency of the obtained beat signal, f_b, is directly related to the target's distance to the radar, R, and is expressed in (5) (Hasch et al., 2012).

$$R = \frac{c \cdot T_c \cdot f_b}{2B} \tag{5}$$

Here, c represents the speed of light, T_c the chirp duration, and B the frequency sweep bandwidth of the radar. Phase changes caused by the target's movement are calculated over consecutive chirp signals (slow-time) to generate $m - D$ velocity profiles (Hakobyan & Yang, 2019).

In this study, a high-resolution Texas Instruments IWR6843ISK (60-64 GHz) mmWave radar module and a DCA1000EVM data capture card were used to losslessly acquire raw ADC (Analog-to-Digital Converter) data for collecting micro-Doppler data of human movements. The 60 GHz operating frequency band of the IWR6843 module ensures that even small-amplitude micro-movements of limbs, such as arms and legs, create distinct Doppler phase shifts, thereby enhancing classification performance. The fundamental radar configuration parameters used during the data collection phase are summarized in Table 1.

Table 1. *Radar Configuration Parameters*

Parameter	Value	Unit
Central Frequency	60	GHz
Bandwidth	4	GHz
Chirp Time	17.28	ms
Sampling Frequency	10	Msps
Number of Chirps per Frame	384	-

The acquired raw ADC data were converted into matrix form, target detection was performed by applying a Range-FFT process, and slow-time data were extracted from the range bin where the target was located. Subsequently, the data were subjected to the STFT, CWT, and WVD transformations detailed in Section 2.2.

3. PROPOSED METHODOLOGY

In this study, a comprehensive methodology integrating signal processing-based time-frequency transformations and the spatial capacities of deep learning architectures is proposed for the classification of different human activities obtained by an FMCW radar.

3.1. Dataset and Preprocessing

The experiments were conducted in a typical indoor environment 5m x 2.64m room to capture realistic multipath effects and background clutter. To ensure the diversity and robustness of the dataset, data was collected from 10 different human subjects (6 males and 4 females, aged 20-53).

The dataset comprises a total of 240 samples distributed equally across 6 distinct human activity classes: Walking, Running, Standing Up, Sitting Down, Jumping and Falling. Each subject performed these movements 4 times at varying distances relative to the radar sensor.

The extraction of clean micro-Doppler signatures from the raw radar returns involved several critical preprocessing steps. First, the raw Analog-to-Digital Converter (ADC) data was organized into a fast-time/slow-time matrix. A 1D Fast Fourier Transform (FFT) was applied along the fast-time axis to obtain the Range-Time profile. To isolate the moving human subjects from stationary indoor objects (such as walls and furniture), a mean subtraction was implemented to remove static background clutter. Once the target bins were localized, the phase information across the slow-time axis was extracted. Finally, the time-frequency representations (STFT, CWT, and WVD) were generated from this phase data to visualize the micro-Doppler signatures. To feed these representations into the proposed deep learning architectures, all generated time-frequency spectrograms were converted to grayscale, resized to a standard dimension of 224x224 pixels, and normalized to a pixel value range of [0, 1]. The images obtained after the transformations are presented in Figure 1, 2, and 3, respectively.

3.2. Proposed Deep Learning Architectures

To fairly compare the impact of different time-frequency transformations (smearing, cross-terms, high resolution) on network perception, 3 distinct deep learning architectures with varying parameter counts and feature extraction capacities were designed. The structural details of the proposed architectures are summarized in the simplified tables. To ensure maximum feature extraction efficiency while preventing overfitting on our specific radar dataset, a careful layer-by-layer design philosophy was followed.

3.2.1. Lightweight Architecture (CNN–Base)

This architecture was designed with a minimum number of parameters to prevent overfitting on the limited dataset consisting of 240 samples. In the network, which consists of three consecutive convolution blocks, narrow kernels of size 3×3 were used, and the spatial dimension was gradually reduced using 2×2 max-pooling. The detailed layer architecture of the network is presented in Table 2. As detailed in the table, the baseline model utilizes a standard sequential progression. Initial convolutional layers employ smaller filter sizes to capture fine-grained time-frequency textures, followed by max-pooling layers to reduce spatial dimensions and computational load. The transition to the fully connected (dense) layers includes a 0.5 dropout rate, which acts as a strict regularization mechanism to prevent the network from memorizing the limited number of samples.

GU J Sci, Part A 13(2) 692-709 (2026) 10.54287/gujsa.1901011

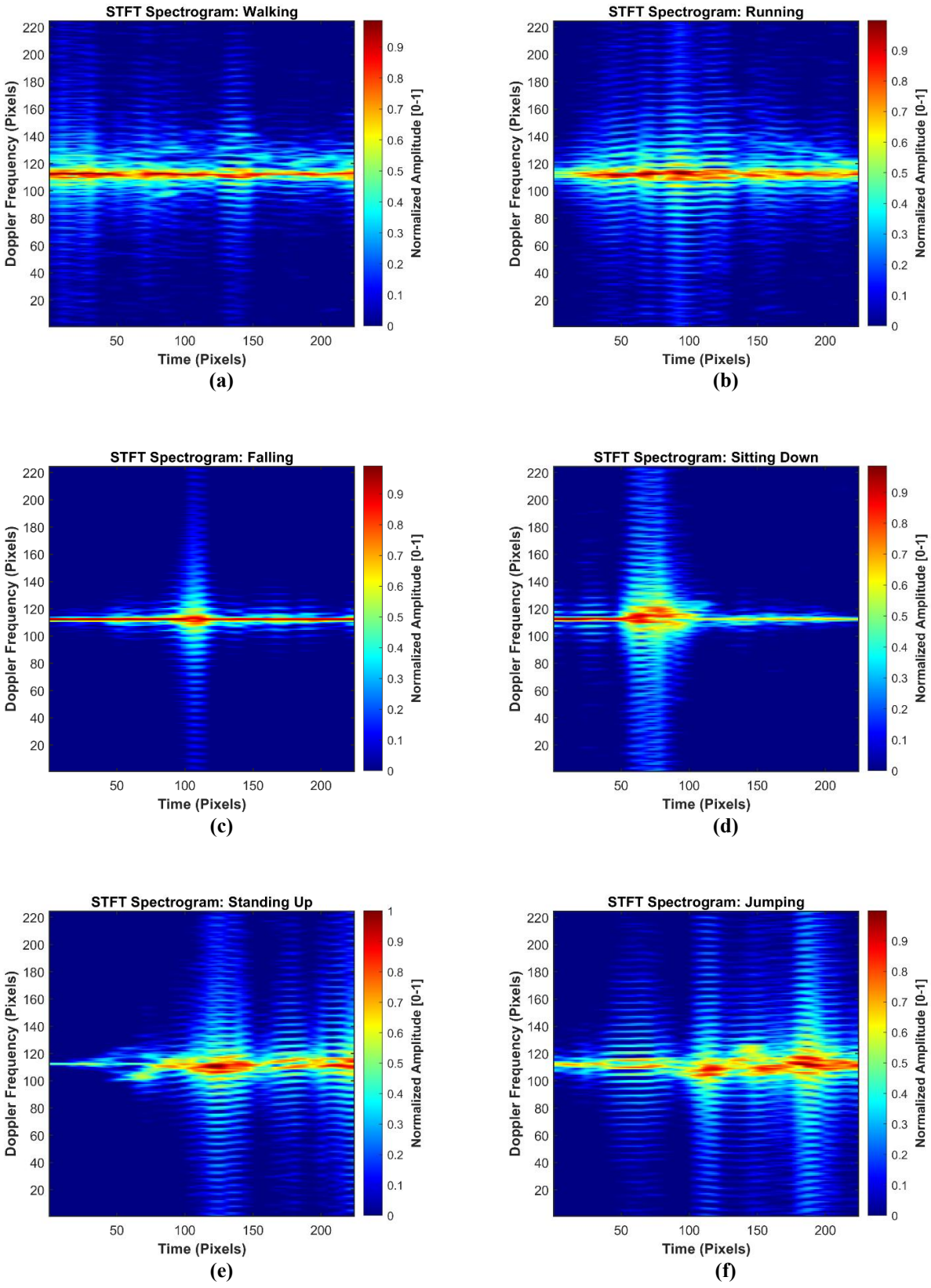

Figure 1. *Images obtained after STFT*
a) *walking,* ***b)*** *running,* ***c)*** *falling,* ***d)*** *sitting down,* ***e)*** *standing up,* ***f)*** *jumping*

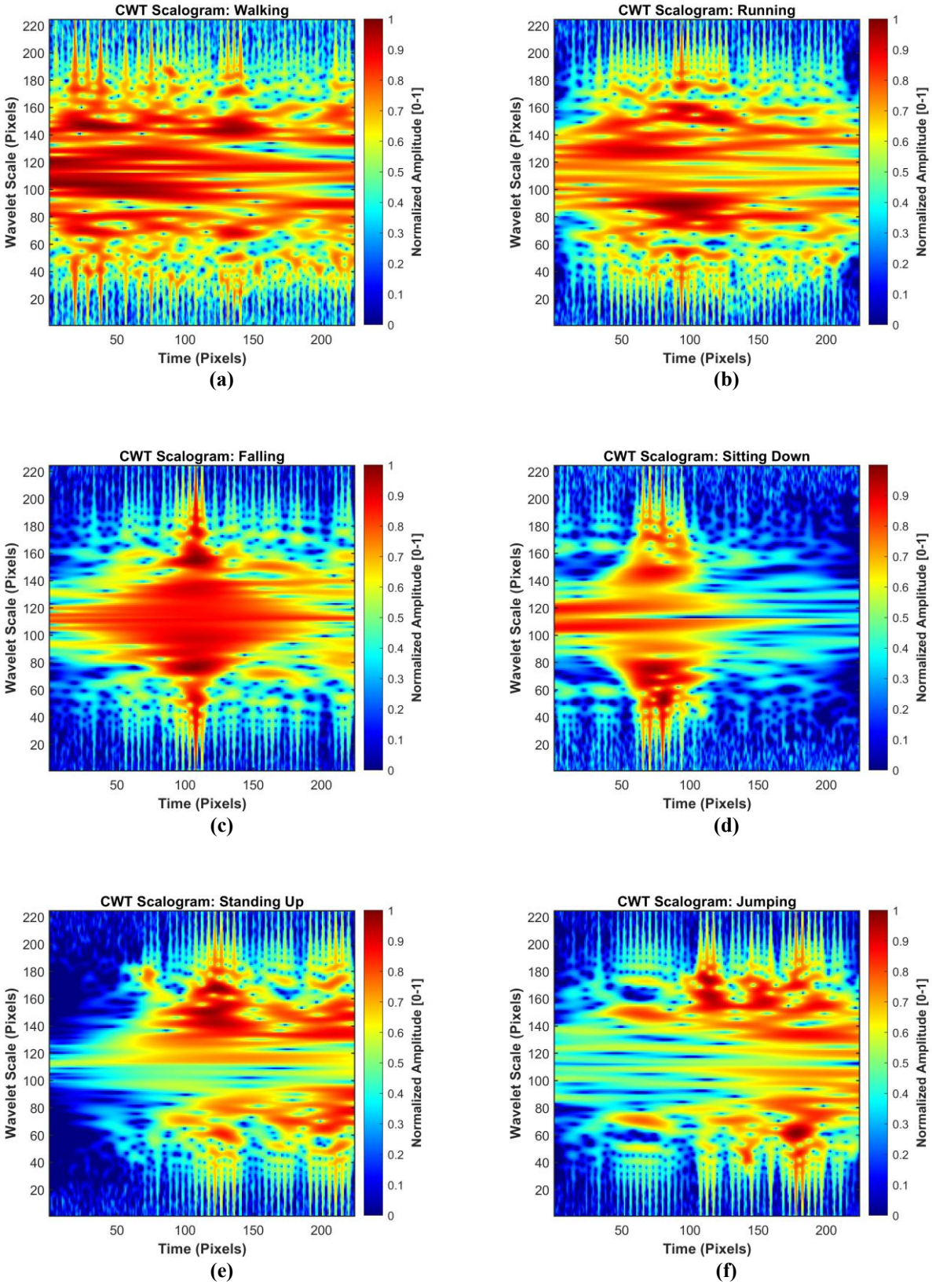

Figure 2. *Images obtained after CWT*
a) *walking,* ***b)*** *running,* ***c)*** *falling,* ***d)*** *sitting down,* ***e)*** *standing up,* ***f)*** *jumping*

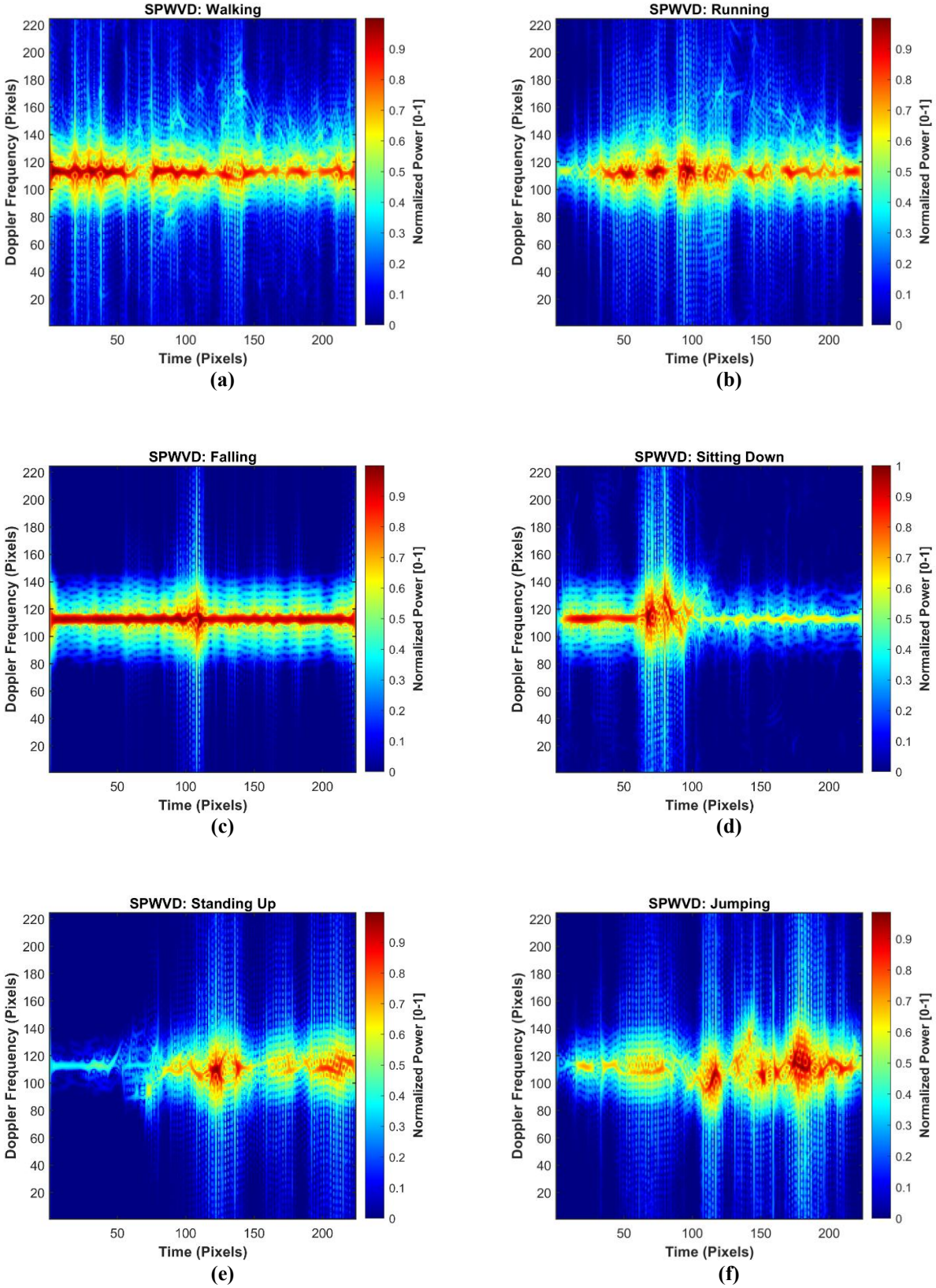

Figure 3. *Images obtained after WVD*
a) *walking,* ***b)*** *running,* ***c)*** *falling,* ***d)*** *sitting down,* ***e)*** *standing up,* ***f)*** *jumping*

Table 2. *Proposed Lightweight Convolutional Neural Network (CNN–Base) Architecture*

Layer	Layer Type	Parameters (Filters / Units)	Kernel Size
Input	Input Layer	-	-
Block 1	Conv2D + MaxPooling2D	16 Filters	[3x3]
Block 2	Conv2D + MaxPooling2D	32 Filters	[3x3]
Block 3	Conv2D + MaxPooling2D	64 Filters	[3x3]
Classifier	Flatten + Dense	64 Units	-
Regularization	Dropout	Rate: 0.5	-
Output	Dense	6 Units	-

Note: Unless otherwise specified, all Conv2D layers utilize a stride of 1 and the 'ReLU' activation function. All MaxPooling2D layers apply a 2x2 pool size with a stride of 2. The final Output Dense layer utilizes the 'Softmax' activation for classification.

3.2.2. Extended Feature Extractor Architecture (CNN–Wide)

This architecture was specifically designed to capture the spatial features of inputs containing cross-terms or high-resolution complex textures, such as WVD, from a broader perspective. To expand the receptive field of the network, 5×5 kernels were used in the first layer, the number of filters was increased, and zero-padding (padding='same') was applied to prevent information loss. This wide receptive field allows the network to observe a larger portion of the movement cycle in a single pass, significantly enhancing the extraction of macro-movement patterns before passing the feature maps to deeper layers. To prevent the increased number of parameters from leading to memorization, the Dropout rate was raised to 0.6. The architecture of this model is presented in Table 3.

3.2.3. Hybrid Spatiotemporal Architecture (CNN–LSTM)

Micro-Doppler signatures in radar spectrograms are sequential data progressing along the time axis (horizontal axis). This architecture, shown in Table 4, extracts the spatial texture of the spectrogram with the first two CNN blocks, reshapes the obtained feature maps into a time-series format, and feeds them into an LSTM layer with 64 units. Thus, this design choice, explicitly shown in the sequential steps of the table, enables the network to track the chronological progression of the micro-Doppler signatures, making it highly sensitive to sequential activities like falling or sitting.

Table 3. *Extended Feature Extractor Architecture (CNN–Wide)*

Layer	Layer Type	Parameters (Filters / Units)	Kernel Size
Input	Input Layer	-	-
Block 1	Conv2D + MaxPooling2D	32 Filters	[5x5]
Block 2	Conv2D + MaxPooling2D	64 Filters	[3x3]
Block 3	Conv2D + MaxPooling2D	128 Filters	[3x3]
Classifier	Flatten + Dense	128 Units	-
Regularization	Dropout	Rate: 0.6	-
Output	Dense	6 Units	-

Note: To maintain feature map spatial dimensions, all Conv2D layers in this architecture utilize 'Same' padding. Standard settings apply: ReLU activation for intermediate layers, convolutions, and a stride of 2 for pooling. The Output layer uses Softmax.

Table 4. *Hybrid Spatiotemporal Architecture (CNN–LSTM)*

Layer	Layer Type	Parameters (Filters / Units)	Kernel Size
Input	Input Layer	-	-
Spatial Block 1	Conv2D + MaxPooling2D	32 Filters	[3x3]
Spatial Block 2	Conv2D + MaxPooling2D	64 Filters	[3x3]
Reshaping	Reshape	(Timesteps, Features)	-
Temporal Block	LSTM	64 Units	-
Regularization	Dropout	Rate: 0.5	-
Output	Dense (Classifier)	6 Units	-

Note: Spatial blocks extract features using ReLU activation (stride=1) and max-pooling (stride=2). The extracted spatial maps are reshaped into sequences and fed into the LSTM temporal block, which utilizes 'Tanh' activation. The final Output layer uses Softmax.

3.2.4. Standard Deep Learning Architectures for Benchmarking

To evaluate the efficiency and necessity of the proposed custom architectures under limited data conditions, three widely adopted standard deep learning models were employed for benchmarking purposes: AlexNet (Krizhevsky et al., 2012), known for its heavy parameterization; ResNet50 (He et al., 2016), a deep residual network; and MobileNetV2 (Sandler et al., 2018), a modern lightweight architecture designed for efficiency. To ensure a fair comparative analysis, these standard models were stripped of their pre-trained ImageNet weights and trained entirely from scratch using the exact same hyperparameters and 6-Fold Cross-Validation strategy applied to the proposed models.

3.3. Training Strategy and Evaluation Metrics

To eliminate bias resulting from random data splitting or memorization due to the limited size of the dataset, all models were trained and tested using a 6-Fold Cross-Validation strategy. The Adam optimization algorithm and the Sparse Categorical Crossentropy loss function were employed for network optimization. The hyperparameters of the models were fixed with a learning rate of 0.001, a batch size of 16, and 25 training epochs. The performance comparison was based on Mean Accuracy, Standard Deviation, and Computation Time metrics obtained across the folds.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the classification performances of the three proposed deep learning architectures (CNN–Base, CNN–Wide, CNN–LSTM) on the three different TF transformations (STFT, CWT, WVD) are presented, and the results are analyzed in-depth within the context of signal processing theory.

4.1. Comparative Performance of TF Transformations and Architectures

All models were trained using a 6-Fold Cross-Validation strategy to objectively measure their generalization capabilities on the limited dataset. Table 5 summarizes the mean accuracy, standard deviation (Std) across folds, and total training time metrics for each combination of TF input and network architecture.

An examination of Table 5 reveals that the highest mean classification performance was achieved by the CNN–Wide architecture trained on WVD data, with an accuracy of 80.42%. While the traditional STFT approach established a strong baseline with 80.00% accuracy using the lightweight CNN–Base architecture, it is noteworthy that CWT, despite theoretically possessing the best time resolution, exhibited the lowest performance across all architectures.

Table 5. *Mean Accuracy and Training Times*

Classifier Model	STFT	CWT	WVD
CNN–Base	%80.00 ± %5.59 (228.6 s)	%75.83 ± %4.49 (237 s)	%76.25 ± %4.73 (284.5 s)
CNN–Wide	%79.58 ± %7.28 (747.2 s)	%73.75 ± %6.08 (893.8 s)	**%80.42 ± %4.66** (1055.3 s)
CNN–LSTM	%68.75 ± %6.57 (555.5 s)	%65.83 ± %11.06 (745.2 s)	%68.33 ± %5.34 (887.3 s)
AlexNet	%60.83 ± %26.37 (1730.9 s)	%46.25 ± %29.32 (1641.8 s)	%8.75 ± %3.46 (2573 s)
ResNet50	%26.67 ± %7.73 (9066.9 s)	%20.42 ± %3.36 (9303.4 s)	%15.42 ± %5.67 (9484.7 s)
MobileNetV2	%17.08 ± %7.13 (3849.3 s)	%18.33 ± %3.73 (3917.7 s)	%12.08 ± %5.85 (4592 s)

4.2. In-depth Analysis of Experimental Findings

The obtained results reveal the critical relationship between deep learning receptive fields and the mathematical nature of time-frequency images:

Comparison of STFT and CWT: Although the Continuous Wavelet Transform (CWT) theoretically offers excellent time resolution at high frequencies, it fell behind the traditional STFT in this study. The primary reason for this is the translation invariance expectation, which is the working principle of CNN. STFT possesses a linear and uniform time-frequency grid; that is, the visual thickness of a micro-Doppler movement remains constant regardless of its position on the frequency axis. However, because CWT utilizes a logarithmic scale, the same micro-movement is visually stretched and compressed (spatial warping) in different frequency bands. This structural deformation made it difficult for standard 3x3 or 5x5 CNN filters to generalize the spatial features of the movement, dragging the model's success down to the 75.83% band.

The Superiority of WVD and CNN–Wide: Although the Wigner-Ville Distribution (WVD) provides superior time-frequency focusing compared to STFT and CWT, it inherently contains cross-terms. This phenomenon reflects the signal energy onto the plane in a much more complex and scattered structure. The narrow receptive field (3x3) of the lightweight CNN–Base architecture was insufficient to filter out this complex structure and the noise created by cross-terms (76.25%). In contrast, the wider kernel structure (5x5) in the first layer and the increased depth (128 filters) of the CNN–Wide architecture successfully deciphered the fine texture of WVD. The wide receptive field learned the cross-terms as 'distinctive spatial features' rather than random noise, translating WVD's theoretical resolution advantage into practice and securing the top position in the study with an 80.42% accuracy. Furthermore, maintaining the Dropout rate of CNN–Wide at the 0.6 level successfully prevented overfitting on the complex data of WVD (±4.66 Std).

The Collapse of the Hybrid Architecture: Designed with the hope of capturing the temporal progression of the spectrograms, the CNN–LSTM architecture failed by remaining below the 70% threshold for all three data types. The reason for this situation is that while the 224x224 dimensional images are spatially compressed in the convolution layers and fed into the LSTM layer as a 1-dimensional vector, the inter-axis relationships (the time-frequency bond) are severely destroyed. In small datasets, there is not enough diversity to learn the long-term dependencies required by recurrent networks.

The Impact of Model Complexity and Comparison with Standard Architectures: To explicitly demonstrate the efficiency of the proposed optimizations, the custom architectures were benchmarked against standard, heavy-weight deep learning models (AlexNet, ResNet50, and MobileNetV2). As shown in Table 5, applying highly deep and heavily parameterized networks to a limited-size radar dataset (240 samples) resulted in severe overfitting. Although these standard models rapidly memorized the dataset and achieved training accuracies exceeding 90%, they suffered a catastrophic drop in validation performance. In a 6-class classification problem where the random guess probability is approximately 16.6%, ResNet50 and MobileNetV2 completely failed to extract generalized features, yielding validation accuracies that hovered around random chance (e.g., 15.42% and 12.08% on WVD, respectively) while requiring dramatically higher computation times (up to 9484 seconds). Furthermore, AlexNet suffered from extreme instability, evidenced by massive standard deviations reaching up to ±29.32% across folds. These comparative results empirically prove that while heavy-weight models tend to memorize the limited data and fail to generalize, the proposed CNN-Base and CNN-Wide architectures, with their highly optimized parameter spaces and structural design, successfully suppress overfitting and efficiently decipher the complex time-frequency textures.

4.3. Error Analysis and Class-Based Performances of the Best Models

To visualize the class-based discriminative power of different time-frequency transformations, the Confusion Matrices of the best models that achieved the highest success for each TF transformation (CNN–Base for STFT, CNN–Base for CWT, and CNN–Wide for WVD) on the test data are presented in Figure 4.

A detailed examination of the confusion matrices reveals distinct character changes in the class-based discriminative capabilities of different time-frequency transformations. First, examining the life-critical Falling class, it is observed that the STFT (Figure 4a) and WVD (Figure 4c) based models detected this critical movement flawlessly, correctly classifying all 8 test samples. However, the CWT-based model (Figure 4b) experienced significant difficulty in this class, correctly identifying only 4 out of 8 samples, confusing half of the samples with other movements such as running and jumping. This finding directly proves the reliability of the proposed WVD and STFT-based mmWave radar systems in critical health applications such as elderly monitoring and fall detection.

The fundamental performance differences among the transformations stem from the kinematic nature of the movements and the ways the algorithms process the signal. Although the CWT-based model exhibited perfect

(100%) success in the Running, Sitting Down, and Standing Up classes, it confused a significant portion (3 out of 8 samples) of the samples in the Walking class with Running and experienced a collapse in the falling class. This situation confirms that the spatial warping caused by logarithmic scaling, discussed in Section 4.2, while making the features of certain movements very distinct, disrupts the visual integrity of movements like walking and falling, thereby misleading the CNN filters.

In contrast, although STFT (Figure 4a) captured the falling movement perfectly, it could not prevent the signal energy from scattering to other classes due to the time-frequency smearing caused by the fixed window size, particularly in the Jumping and Walking (4/8 correct) classes, where it correctly identified only 4 out of 8 samples for each class.

The overall best-performing model of the study, the WVD + CNN–Wide combination (Figure 4c), exhibited the most balanced and stable performance. WVD increased the number of correct predictions in the Walking and Jumping classes (accurately predicting 6 and 5 out of 8 samples, respectively), where STFT struggled, while maintaining its flawless (100%) success in the Falling and Standing Up classes. Although the traditional confusion between Running and Walking movements was partially observed in WVD as well (3 errors in the Running class), the successful learning of the high-resolution texture containing cross-terms by the wide receptive field architecture elevated the overall classification accuracy above the other methods.

In addition to the classification performance, to verify the training stability and generalization capacity of the model, the Training and Validation Accuracy and Loss learning curves recorded over 25 epochs by the optimal WVD-based CNN–Wide architecture is presented in Figure 5. When the learning graphs are examined, it is observed that the training and validation curves exhibit consistent convergence from the initial epochs. Despite the extended filter structure of the network (depth reaching up to 128 filters) and the complex cross-terms of WVD, no sharp fluctuations or early divergence were observed in the validation loss curve. This balanced learning characteristic proves that the high Dropout strategy of 0.6 applied in the architecture and the wide receptive field of 5×5 dimensions successfully suppressed overfitting on the limited radar dataset, extracting permanent spatial features instead of memorizing the data.

5. CONCLUSION AND FUTURE WORKS

In this study, the performances and spatial compatibilities of different TF transformations (STFT, CWT, WVD) combined with novel deep learning architectures (CNN–Base, CNN–Wide, CNN–LSTM) were comparatively investigated for the classification of human micro-Doppler signatures obtained using a FMCW radar. Experimental analyses revealed that traditional STFT provides an acceptable baseline accuracy (80.00%), whereas CWT, despite theoretically possessing high time resolution, misled the deep learning filters due to logarithmic scaling deformation (warping), thereby exhibiting the lowest performance (73.75% - 75.83%). The WVD, despite the complex texture created by the cross-terms it contains, achieved excellent harmony with the CNN–Wide architecture—which features a wide receptive field (5×5 kernel size) and a high dropout rate—

securing the top position in the study with an 80.42% accuracy. Furthermore, the best-performing model's ability to detect life-critical sudden-acceleration movements like Falling with 100% accuracy proved the reliability of the system even under constrained hardware and data conditions.

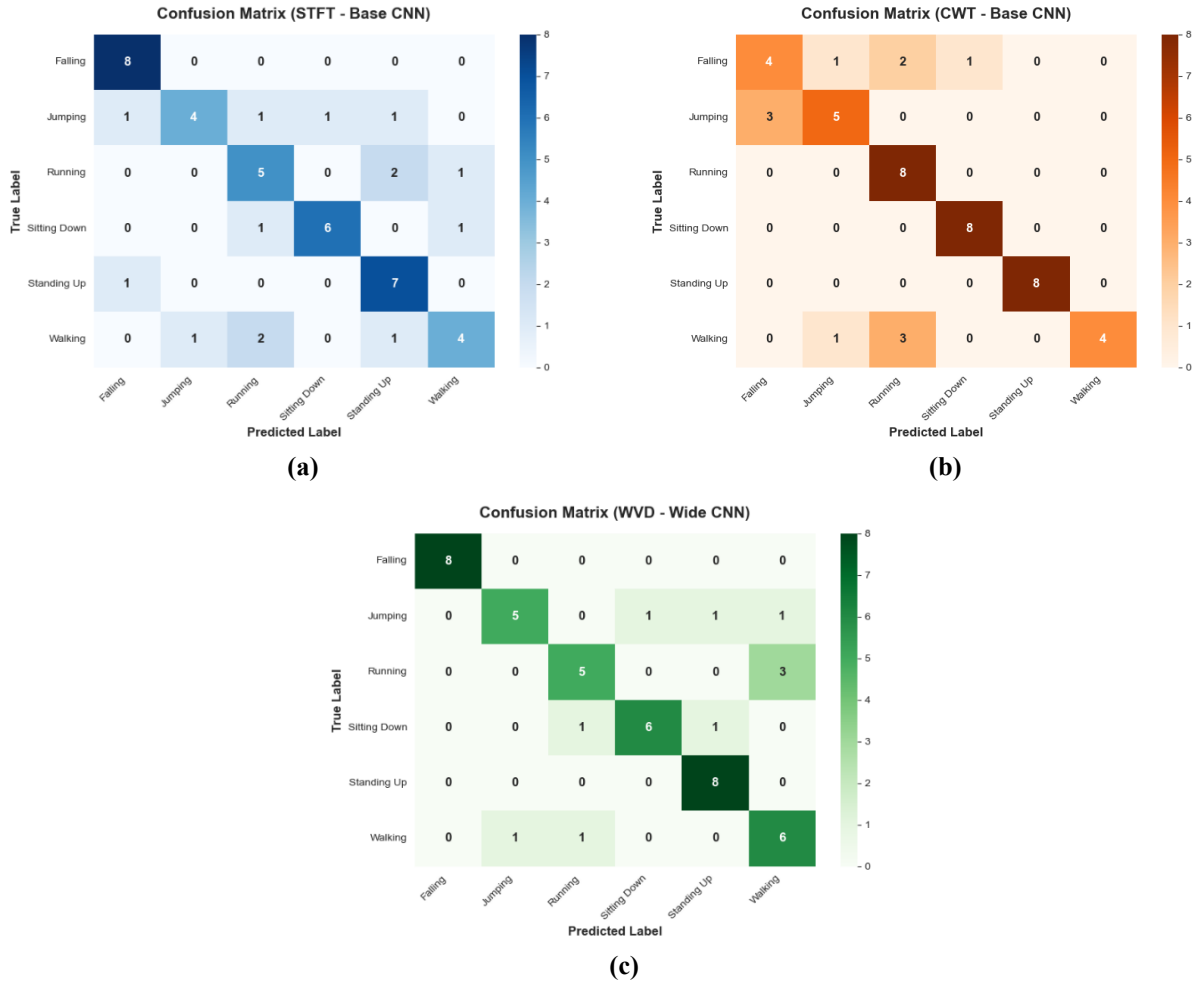

Figure 4. *Confusion matrices of best models*
a) *STFT-based model,* ***b)*** *CWT-based model,* ***c)*** *WVD-based model*

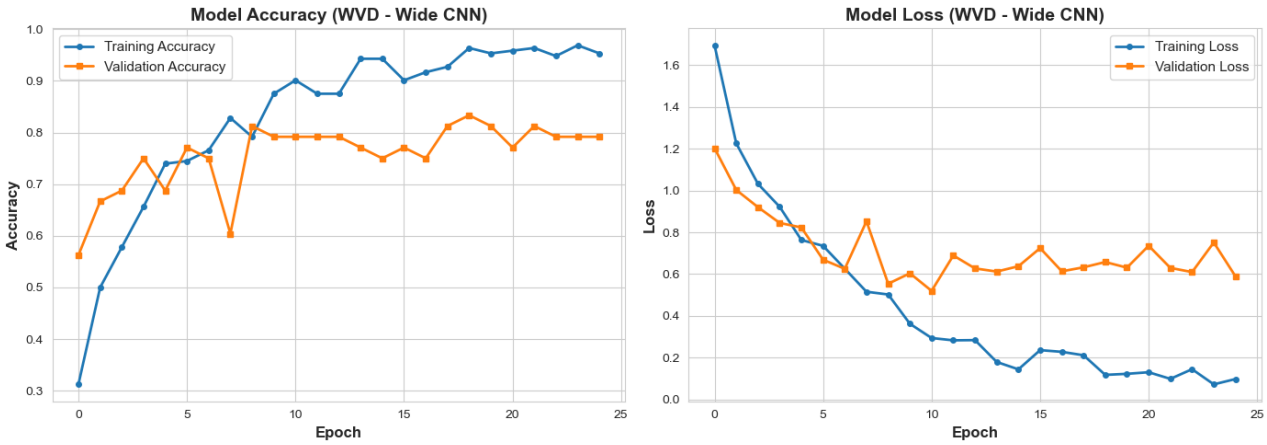

Figure 5. *Training and validation accuracy and loss curves for the optimal WVD + CNN–Wide architecture*

In future studies, to increase the generalizability of the proposed WVD-based wide convolutional architecture, it is planned to expand the dataset with participants from different age groups and body types. Additionally, it is aimed to optimize the developed algorithms to be lightweight for integration into edge computing devices and to test their real-time performances in simultaneous multi-target tracking scenarios.

AUTHOR CONTRIBUTIONS

Conceptualization, L.S. and H.S.; methodology, H.S.; software, H.S.; validation, H.S. and L.S.; formal analysis, H.S.; investigation, H.S.; resources, L.S.; data curation, H.S.; writing—original draft preparation, H.S.; writing—review and editing, L.S.; visualization, H.S.; supervision, L.S.; project administration, L.S.; funding acquisition, L.S. All authors have read and legally accepted the final version of the article published in the journal.

AI Exclusion: AI tools are strictly prohibited from being listed as authors.

ACKNOWLEDGEMENT

This study was supported by Scientific and Technological Research Council of Türkiye (TUBITAK) under Grant Number 125E391. The authors thank TUBITAK for their support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abdu, F. J., Zhang, Y., & Deng, Z. (2022). Activity classification based on feature fusion of FMCW radar human motion micro-Doppler signatures. *IEEE Sensors Journal*, *22*(9), 8648-8662. https://doi.org/10.1109/JSEN.2022.3156762

Avinash, B. S., Aryan, R., Yadav, A., Kumar Gupta, V., & Kumar, B. (2025). Hierarchical Multiscale CNN With Frequency-Aware Attention for Enhanced HAR. *IEEE Sensors Journal*, *25*(16), 31175-31184. https://doi.org/10.1109/JSEN.2025.3586069

Biswas, S., Manavi Alam, A., & Gurbuz, A. C. (2024). HRSpecNET: A deep learning-based high-resolution radar micro-Doppler signature reconstruction for improved HAR classification. *IEEE Transactions on Radar Systems*, *2*, 484-497. https://doi.org/10.1109/TRS.2024.3396172

Faisal, K. N., Mir, H. S., & Sharma, R. R. (2023). Human activity recognition from FMCW radar signals utilizing cross-terms free WVD. *IEEE Internet of Things Journal*, *11*(8), 14383-14394. https://doi.org/10.1109/JIOT.2023.3344100

Gong, S., Shi, H., Yan, X., Fang, Y., Paul, A., Wu, Z., & Long, W. (2023). Human activity recognition with FMCW radar using few-shot learning. *IEEE Sensors Journal*, *23*(15), 17815-17824. https://doi.org/10.1109/JSEN.2023.3290565

Hakobyan, G., & Yang, B. (2019). High-performance automotive radar: A review of signal processing algorithms and modulation schemes. *IEEE Signal Processing Magazine*, *36*(5), 32-44. https://doi.org/10.1109/MSP.2019.2911722

Hasch, J., Topak, E., Schnabel, R., Zwick, T., Weigel, R., & Waldschmidt, C. (2012). Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band. *IEEE Transactions on Microwave Theory and Techniques*, *60*(3), 845-860. https://doi.org/10.1109/TMTT.2011.2178427

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (27-30 June 2016, pp. 770-778), Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.90

Dang, V. N., Hoang, N. C., Le, M. T., Nguyen, K., & Nguyen, Q. C. (2026). Deep Learning-based Human Activity Recognition with FMCW Radar: A Review. *IEEE Sensors Journal*, *26*(4), 5302-5326. https://doi.org/10.1109/JSEN.2026.3651578

Kim, W. Y., & Seo, D. H. (2022). Radar-based human activity recognition combining range–time–Doppler maps and range-distributed-convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-11. https://doi.org/10.1109/TGRS.2022.3162833

Kumawat, H. C., & Raj, A. A. B. (2022). SP-WVD with adaptive-filter-bank-supported RF sensor for low RCS targets' nonlinear micro-Doppler signature/pattern imaging system. *Sensors*, *22*(3), 1186. https://doi.org/10.3390/s22031186

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. In: F. Pereira, C. J. Burges, L. Bottou, & K. Weinberger (Eds.), Advances in Neural Information Processing Systems 25 (NIPS 2012).

Miao, F., Huang, Y., Lu, Z., Ohtsuki, T., Gui, G., & Sari, H. (2025). Wi-Fi sensing techniques for human activity recognition: Brief survey, potential challenges, and research directions. *ACM Computing Surveys*, *57*(5), 107. https://doi.org/10.1145/3705893

Mishra, K. K., & Pachori, R. B. (2025). FMCW radar-based human activity recognition based on higher-order synchrosqueezing transform. *IEEE Sensors Journal*, *25*(15), 28934-28941. https://doi.org/10.1109/JSEN.2025.3579423

Patole, S. M., Torlak, M., Wang, D., & Ali, M. (2017). Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine*, *34*(2), 22-35. https://doi.org/10.1109/MSP.2016.2628914

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). M*obileNetV2: Inverted residuals and linear bottlenecks*. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (18-23 June 2018, pp. 4510-4520), Salt Lake City, UT, USA. https://doi.org/10.1109/CVPR.2018.00474

Sarna, M. L. A., Hossain, M. R., & Islam, M. A. (2024). Comparative analysis of stft and wavelet transform in time-frequency analysis of non-stationary signals. *International Journal of Novel Research in Engineering and Science*, *11*(1), 72-78. https://doi.org/10.5281/zenodo.11230337

Ullmann, I., Guendel, R. G., Kruse, N. C., Fioranelli, F., & Yarovoy, A. (2023). A survey on radar-based continuous human activity recognition. *IEEE Journal of Microwaves*, *3*(3), 938-950. https://doi.org/10.1109/JMW.2023.3264494

Ünal, L., Pakfiliz, A.G. (2022). LPI Radar Signal Detection Based on Autocorrelation Function and Wigner-Ville Distribution. *Review of Computer Engineering Studies*, *9*(4), 125-135. https://doi.org/10.18280/rces.090401

Wang, C.-H., Zong, Z.-Y., Yin, X.-Y., Li, K., & Zuo, Y.-H. (2025). A novel local maximum synchrosqueezing W transform for reservoir characterization. *Petroleum Science*, *23*(4), 1842-1859. https://doi.org/10.1016/j.petsci.2025.11.034

Wu, Y., Fioranelli, F., & Gao, C. (2025). RadMamba: Efficient Human Activity Recognition Through a Radar-Based Micro-Doppler-Oriented Mamba State-Space Model. *IEEE Transactions on Radar Systems*, 4, 261-272. https://doi.org/10.1109/TRS.2025.3648848

Yin, W., Shi, L.-F., & Shi, Y. (2024). Indoor human action recognition based on millimeter-wave radar micro-doppler signature. *Measurement*, *235*, 114939. https://doi.org/10.1016/j.measurement.2024.114939

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, *22*(4), 1476. https://doi.org/10.3390/s22041476