Real-Time Lightweight Deep Learning Models for Human Activity Recognition Using FMCW Radar

Fahad Ayaz[1], Basim Alhumaily[1], Sajjad Hussain[1], Muhammad Ali Imran[1], and Ahmed Zoha[1]

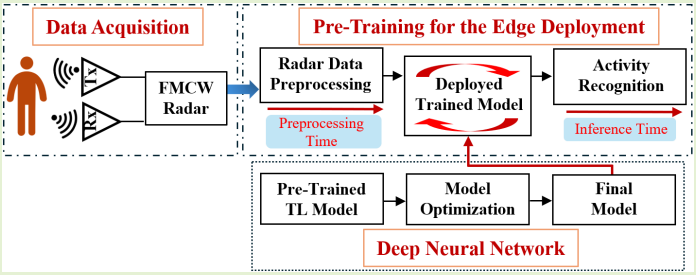
[1]University of Glasgow

July 14, 2024

Real-Time Lightweight Deep Learning Models for Human Activity Recognition Using FMCW Radar

Fahad Ayaz, *Student Member, IEEE*, Basim Alhumaily[1,2], Sajjad Hussain, *Senior Member, IEEE*, Muhammad Ali Imran, *Fellow, IEEE*, and Ahmed Zoha, *Senior Member, IEEE*

***Abstract*— Radar-based Human Activity Recognition (HAR) has attracted much attention in various fields such as smart security, medical monitoring, and human-computer interaction. Integrating Convolutional Neural Networks (CNNs) with radar spectrum techniques for HAR is becoming increasingly popular. However, traditional network models usually have a large number of parameters and require long training and inference times, making them less suitable for real-time applications. To address these issues, this study proposes a lightweight CNN model based on Frequency-Modulated Continuous Wave (FMCW) radar, designed for edge devices for efficient real-time monitoring. We compare three different 2D domain radar data preprocessing techniques - Time Range (TR), Short-Time Fourier Transform (STFT), and Smoothed Pseudo-Wigner-Ville Distribution (SPWVD) - along with four state-of-the-art neural networks. Our approach achieves high accuracy in HAR classification and effectively addresses the challenges posed by limited radar data through Transfer Learning (TL), demonstrating the potential for real-time applications. After evaluating 12 configurations of CNN models and preprocessing methods, we found that MobileNetV2 with STFT was the most efficient and lightweight, with STFT taking only 220 ms to generate a spectrogram sample. This combination achieved an inference time of only 2.57 ms per sample and a recognition accuracy of 96.30%, setting a new benchmark for real-time intelligent systems on edge devices.**



***Index Terms*— Human Activity Classification, FMCW Radar, micro-Doppler Spectrograms, Efficient and Lightweight Network, Transfer Learning.**

I. Introduction

HUMAN activity recognition (HAR) is essential for identifying body movements and is becoming increasingly important in applications such as security, healthcare, assisted living, and sports [1]. It plays a vital role in monitoring the elderly, especially those living alone at home or in care facilities, as it can reduce the risk of accidental falls that can lead to injury or even death. The World Health Organization (WHO) [2], has highlighted the serious consequences of such falls, which not only pose an immediate health threat but also increase healthcare costs, placing tremendous pressure on healthcare providers, insurers, and families.

In the United Kingdom (UK), falls cost the National Health Service (NHS) over £2.3 billion per annum [3], highlighting the economic impact and importance of enabling up to 90% of older people to remain independent and in their own homes for as long as possible [4]. Therefore, this drives the demand for cost-effective alternatives to traditional assisted living and nursing home facilities. Recognizing the importance of HAR systems for independent living among older adults, the literature describes various data acquisition methods, categorized as wearable and non-wearable devices. Wearable devices [5] such as accelerometers, gyroscopes, and magnetometers are widely used. Accelerometers, in particular, detect falls by measuring changes in acceleration along the vertical axis, as falls typically result in noticeable acceleration of different parts of the body. However, despite their affordability, wearable devices require continuous wear that can cause discomfort. In addition, these devices often include a manual button, which can complicate their use in older adults, as they may not be able to activate them when needed because they may be immobilized during a fall.

Non-wearable sensors, which leverage vision-based and radio frequency (RF) technologies, are becoming increasingly popular due to their contactless nature. Vision sensors employ high-resolution cameras and advanced computer vision techniques, but face privacy issues, light sensitivity, and camera limitations [6]. Continuous surveillance may cause some elderly patients to reject such monitoring. In contrast, RF-based sensors, including technologies such as WiFi [7] and radar [8], are becoming increasingly popular. For example, WiFi has been extensively studied in applications, such as WmFall [9]. It typically requires separate transmitter and receiver devices, is susceptible to interference, and has higher preprocessing

Fahad Ayaz, Basim Alhumaily[1], Sajjad Hussain, Muhammad Ali Imran, and Ahmed Zoha are with the University of Glasgow, James Watt School of Engineering, Glasgow, G12 8QQ, UK. (email: {f.ayaz.1 and b.alhumaily.1}@research.gla.ac.uk, {Sajjad.Hussain, Muhammad.Imran and Ahmed.Zoha}@glasgow.ac.uk).

Basim Alhumaily[2] is also with the Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah 52571, Saudi Arabia. (e-mail: b.alhumaily@qu.edu.sa).

requirements for deep learning (DL) models [10].

On the other hand, radar sensors offer a non-intrusive alternative that is well suited for home or indoor monitoring and is capable of operating under low-light conditions without compromising privacy. These characteristics make radar an excellent choice for HAR, as it is capable of identifying, detecting, and classifying human activities. The functionality of radar-based HAR systems relies heavily on preprocessing radar echoes into 2D echo signals, which involves complex data transformations necessary for effective analysis. To develop an efficient system, we employ techniques from three 2D radar domains: the Time Range (TR) domain [8] and two Time-Doppler (TD) domains, namely the Short-Time Fourier Transform (STFT) [11] and Smoothed Pseudo-Wigner–Ville Distribution (SPWVD) [11]. Each domain provides unique insights and has different advantages and limitations. The TR domain is fast but lacks Doppler information and is more susceptible to noise, which affects the accuracy of the motion analysis. The STFT technique provides valuable Doppler shift data, but cannot provide time-varying range information. On the other hand, the SPWVD method was selected for its high-resolution time-frequency analysis, which is essential for capturing the complex dynamics in human motion.

By exploring these domains, we can leverage their strengths to improve the performance of the HAR systems. This integration ensures that the data input into the DL model is optimally prepared to learn and recognize activity-specific features, thereby achieving a balance between the processing efficiency and performance accuracy. Preprocessing radar data into a 2D spectrum leads to a major feature extraction stage, which has historically been limited by the manual selection required by traditional machine learning (ML) methods. In addition, basic statistical features [12], such as the mean and variance, are insufficient for capturing the complex patterns present in the above mentioned 2D domain methods.

Convolutional Neural Networks (CNNs) [13], have significantly improved the ability to autonomously learn and distinguish complex data patterns. CNNs are particularly effective in processing image data, including spectrograms generated by radar, because they can perform both feature extraction and classification. This dual functionality has revolutionized fields such as image recognition and computer vision. A groundbreaking study conducted by [14] introduced the first innovative CNN model for document recognition. However, it was pioneering work by [15] that brought CNN algorithms into the spotlight. Their CNN architecture achieved an impressive top-1 error rate of 37.5% at the ImageNet challenge [16] in 2012. As a result of this significant breakthrough, several CNN architectures have been developed, including VGG-Net [17], Mobile-Net [18] and Res-Net [19]. Due to their excellent performance in image classification, these architectures have also been applied to the processing of radar spectrograms for HAR and fall detection. However, training a CNN from scratch requires a large amount of data, which is often rare in specialized applications, such as radar-based HAR, resulting in overfitting or underfitting. To overcome this challenge, we employed Transfer Learning (TL) approach [20]–[22].

TL allows the use of pre-trained models on large datasets, thus enabling the adaptation of existing CNNs to new tasks. This strategy significantly reduces the need for large amounts of data and accelerates the training process. Therefore, our study leverages well-known CNN architectures, such as VGG-16, VGG-19, ResNet-50, and MobileNetV2, which were chosen for their demonstrated effectiveness in image-based learning tasks [23], [24]. We fine-tuned our training samples on these four pre-trained architectures to optimize the HAR system for high accuracy and fast real-time prediction, which is essential in critical applications such as fall detection [25].

These models were chosen because of their unique architectural advantages and proven performance in image-based learning tasks, which are well-suited for processing radar-generated spectrograms. The VGG model, which is known for its deep architecture, provides powerful feature-extraction capabilities. ResNet-50 introduces residual learning to solve the gradient vanishing problem, thereby facilitating the training of deeper networks. MobileNetV2 uses depth-wise separable convolutions, which improve the computational efficiency by processing each input channel separately and then combining the feature maps with 1×1 convolutions. In addition, MobileNetV2 integrates reverse residual connections and a modified residual link to learn more complex features while maintaining efficiency.

In this study, we explored the combination of lightweight DL models with radar-generated 2D spectra to create efficient HAR systems. We analyzed the performance of various preprocessing techniques and CNN architectures, focusing on their application in edge-computing scenarios with limited computational resources. By optimizing these techniques for real-time processing, we aim to enhance the capabilities of HAR systems, thereby making them more accessible and effective in real-world environments. Our contribution specifically focuses on exploring three different preprocessing techniques and four CNN models, resulting in 12 unique data preprocessing model pairs. We examined their recognition accuracy and efficiency with the goal of identifying the most effective combination for deployment on low-power edge devices.

Our contributions can be summarized as follows:

- **Evaluation of Radar 2D Domain Techniques:** We empirically evaluated two TD domain techniques (STFT and SPWVD) in conjunction with TR analysis. We quantified their computational efforts in real-time HAR systems.
- **Optimizing models with Transfer Learning (TL):** We evaluated the performance of various CNN architectures, including VGG-16, VGG-19, ResNet-50, and MobileNetV2, to improve the accuracy of the proposed HAR system using TL methods.
- **Real-time Performance Analysis of Model Domain Combinations:** We conducted a comprehensive analysis of 12 CNN domain technique combinations, focusing on real-time performance to optimize the balance between accuracy and computational efficiency (training and inference times). The analysis is also extended to performance metrics beyond accuracy, such as recall, precision, and F1 score, which are critical for evaluating the effectiveness in real-world applications.

This paper is structured as follows: Section II presents an in-

depth description of the radar-based HAR approach, covering the radar technology, dataset, preprocessing techniques, and CNN architecture used. Section III presents a comparative evaluation of different model preprocessing combinations. Finally, Section IV summarizes the main findings and contributions of this study, and suggests possibilities for future research.

II. Proposed HAR System

The system model for this study is shown in Fig. 1, outlining a comprehensive process that starts with data acquisition using an FMCW radar. It proceeds with signal processing to generate three different domain representations, which are then sequentially applied to the four different CNN models. The following subsections explain each component of the system model in detail.

A. Data Acquisition

This study used a dataset from the James Watt School of Engineering at the University of Glasgow, UK, which includes a wide range of everyday human movement activities [26], as described in Table I. The selected dataset is notable for its extensive use in recent academic work [27], ensuring that our research is consistent with the current trends in the field.

The dataset was collected using an FMCW radar operating at a 5.8 GHz carrier frequency and 400 MHz chirp bandwidth, and involved 81 volunteers of varying ages. The FMCW radar

TABLE I: Human Activity Samples

No.	Activity Description	No. of Samples	Data Length
A1	Walking	312	10 s
A2	Sitting	311	5 s
A3	Standing	311	5 s
A4	Picking an object	309	5 s
A5	Drinking	311	5 s
A6	Fall	197	5 s

can capture detailed range and Doppler frequency information of a target, making it a valuable tool in HAR systems. This radar operates by transmitting a *chirp* signal (a modulated continuous wave) reflected from objects within the radar's line of sight. The fundamental representation of a radar signal reflecting towards a target is given by [28]:

$$s(t) = e^{j(2\pi f_c t + \pi \frac{B}{T} t^2)} \tag{1}$$

where B is the bandwidth of the chirp, T is the chirp duration, and $\frac{B}{T}$ is the frequency modulation slope of the chirp. The received signal, a delayed and attenuated version of the transmitted signal, is given as:

$$r(t) = e^{j(2\pi f_c (t-t_d) + \pi \frac{B}{T} (t-t_d)^2)} \tag{2}$$

where $t_d = \frac{2R}{c}$ is the time delay, R is the range to the target, and c is the speed of light. The reflected signal captured by the receiving antenna is mixed with the transmitted signal to produce an intermediate frequency (IF) signal, which is then converted into a digital signal using an analog-to-digital converter (ADC) in the radar equipment for further analysis. The beat frequency signal, called the raw radar data matrix, was generated using digital signal processing (DSP). It is typically expressed in complex form as:

$$IF(t) = e^{j\left(4\pi \frac{BR}{cT} t + \frac{4\pi R}{\lambda}\right)} \tag{3}$$

or equivalently:

$$IF(t) = e^{j(f_b t + \phi_b)} \tag{4}$$

where f_b is the beat frequency and ϕ_b is the phase difference.

The result is then demodulated to baseband to produce the In-phase (I) and Quadrature (Q) components. The frequency of the beat signal is proportional to the target's range, and its phase is related to the target's velocity (Doppler information). Analyzing the I and Q signals allows the determination of both the range and velocity of the target. The IF signal is organized in such a way that its rows and columns correspond to slow and fast time variables, respectively, where 'fast time' refers to the time of a single sweep, whereas 'slow time' spans across multiple sweeps [8].

B. Data Preprocessing

The IF signal that contains I and Q signals, undergoes several steps to improve its quality and accurately represent human activity information, as shown in Fig. 1. First, these signals were converted into a complex digital format and reshaped into a 2D matrix to align the data for subsequent processing before applying the Fast Fourier Transform (FFT).

The process begins by applying the Blackman windowing technique to minimize spectral leakage, followed by conducting an FFT on the fast-time frames, or chirps, to extract range information over time, known as Range-FFT. A filter called a Moving Target Indicator (MTI) detects only moving targets and effectively removes any clutter or stationary objects from the radar signal. The filter was designed as a fourth-order Butterworth high-pass filter with a cut-off frequency of 0.0075 Hz. After filtering with the MTI filter, the range bins are extracted to generate a range profile or the TR domain. The general formula for range-FFT is defined as:

$$X[k] = \sum_{n=0}^{N-1} s[n] e^{-j\frac{2\pi}{N} kn} \tag{5}$$

where $X(k)$ represents the FFT of each k frequency bin, and N is the number of samples. The TR domain reflects the time-varying range information between the radar and target. Finally, the STFT and SPWVD techniques were applied to the selected range bins to generate TD spectrograms, as detailed in the subsequent subsections and illustrated in Fig. 1.

1) STFT based TD spectrogram: To obtain a TD representation of the activity, we employed the STFT technique [11]. In particular, the STFT employs a Hann window of size 200, with the FFT incorporating 800 sampling points, a zero padding factor of 4 and a 95% overlap between consecutive frames (i.e., 190 samples). This approach balances frequency and time information, resulting in a 2D image, as shown in Fig. 2. The mathematical representation of the STFT applied in our study is given by [10]:

$$STFTs(m,f) = \sum_{k=-\infty}^{+\infty} s(k,n) w(m-k) e^{(-j2\pi kf)} \tag{7}$$

where $w(\cdot)$ represents the window function. From this formulation, it is evident that the time-frequency resolution of the STFT spectrogram is affected by the choice of the window function. A longer window length improved the frequency resolution, whereas a shorter window length improved the time

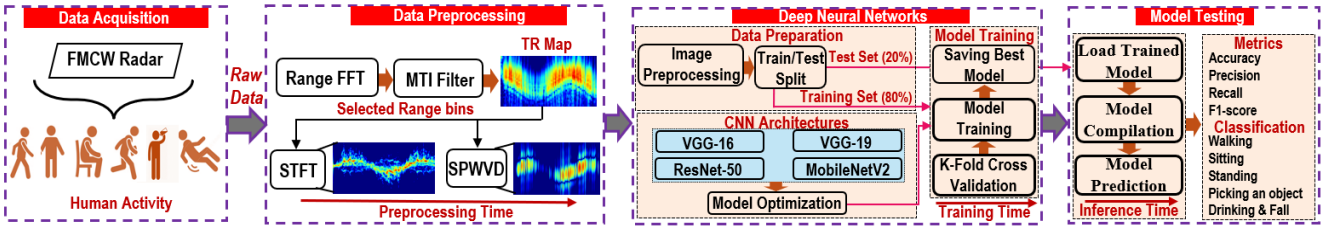


Fig. 1: Proposed radar-based HAR system depicting the workflow from data acquisition to 2D spectrum generation, along with state-of-the-art Neural Networks.

resolution. Therefore, the STFT spectrogram faces a trade-off and cannot achieve a high resolution in both the time and frequency domains.

2) SPWVD based TD Spectrogram: Since STFT is limited by non-independent time and frequency windows, SPWVD provides a more sophisticated approach to TD analysis by utilizing independent windows. The SPWVD is defined as follows [10]:

$$\mathrm{SPWVD}_s(m,f) = \sum_{k=-\infty}^{+\infty} \sum_{\tau=-\infty}^{+\infty} s\left(k+\frac{\tau}{2},n\right) s^*\left(k-\frac{\tau}{2},n\right) \times h(\tau) w(k-m) e^{-j2\pi\tau f} \quad (8)$$

In the above equation, $w(\cdot)$ and $h(\cdot)$ shows the time window and the frequency window functions. In this study, we utilized SPWVD employing Kaiser and Hann for time and frequency window functions with lengths of 25 and 15, respectively, to improve the resolution and clarity of the spectrograms, as shown in Fig. 2. One of the objectives of this study is to evaluate the applicability of the SPWVD as a spectrogram technique and examine its characteristics to determine the potential benefits for efficient HAR systems. Therefore, integrating these three domains, utilizing their respective strengths, and addressing their challenges are essential for improving the performance of the HAR system. This comprehensive approach ensures that the data fed into the CNN architecture are well suited for learning and recognizing activity-specific features, thereby achieving a balance between processing efficiency and performance accuracy.

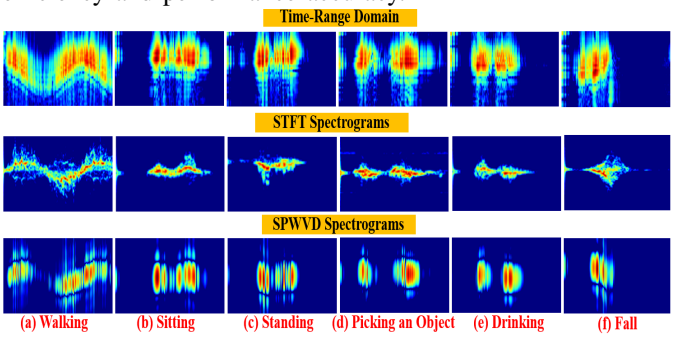


Fig. 2: 2D images of six activities resulting from TR, STFT and SPWVD techniques.

C. Deep Neural Networks (DNNs)

Following the discussion in the previous section II-B, we preprocessed the acquired data and utilized a DNN, as shown in Fig. 1. This process involves three steps: preparing data for input to the CNN, selecting and optimizing the model, and training the data on the selected model.

1) Data Preparation: The first step of the DNN is preprocessing to standardize all the images in the dataset. Each image was resized to a uniform size of 224 × 224 pixels to satisfy the input size requirement of the selected CNN model. This preprocessing step also includes image normalization and subtraction of the mean RGB value based on the training set as well as other necessary transformations. To ensure a strong learning environment, the processed images were labelled and randomly shuffled. This process was designed to ensure a diverse and representative distribution of data within the dataset. In terms of data allocation, we adopted an 80-20 split; 80% of the data were used for training, and the remaining 20% were used for testing purposes. An important aspect of our dataset management method is the stratified partitioning approach, which ensures that the class distribution remains consistent in the training and test sets.

2) CNN Architectures: The basis of our approach is to apply the TL method. TL is a powerful technique in DL, in which a model developed for one task is reused as the starting point for a model for a second task. It is particularly useful in scenarios with limited labeled data, such as radar datasets for HAR. Hence, it is not necessary to train a CNN from scratch. Moreover, it plays a key role in alleviating overfitting and enhancing the generalization process. This approach involves leveraging pre-trained weights from widespread datasets, such as ImageNet, which is particularly useful in addressing the class imbalance problem in the dataset used. In this study, we investigated the effectiveness of four prominent CNN architectures: VGG-16, VGG-19, ResNet-50, and MobileNetV2, as shown in Fig. 1.

Visual Geometry Group (VGG): VGG-16 and VGG-19 are CNN architectures developed by the Visual Geometry Group (VGG) at the University of Oxford, UK. The VGG-16 architecture contains 16 layers and is known for its simplicity and performance in image recognition tasks. The model achieved a top-5 test accuracy of nearly 92.7% on ImageNet. It replaces filters with large kernel sizes with several 3×3 kernel-size filters, providing a significant improvement over the AlexNet model [17]. The same group extended the VGG-16 to VGG-19. The numbers '16' and '19' denote the weight layers. Although similar in structure to VGG-16, the added layers provide deeper features that may improve the recognition performance for complex scenes in radar-based HAR datasets.

ResNet-50: The ResNet architecture had multiple configurations, each with a different number of layers. In our

study, we chose to implement ResNet-50, a variant of ResNet equipped with 50 neural network layers [18]. The ResNet-50 model stands out for its ability to handle CNN tasks without performance degradation, which is a common problem when scaling CNN structures. With its 50-layer framework, ResNet-50 excels at recognizing complex patterns and performs well across a range of recognition tasks. This deep architectural capability is particularly beneficial for detecting subtle human activities in radar data.

MobileNetV2: MobileNetV2 is a successor to the original MobileNet [16] and is a CNN variant designed specifically for mobile and embedded vision applications. This architecture ensures that the network is not only lightweight, but also has a lower inference latency, which is crucial for critical applications such as fall detection. MobileNetV2 uses two unique block structures: a residual block with stride 1 to maintain dimensionality, and another block with stride 2 to reduce space [29]. Despite its lightweight, MobileNetV2 is expected to maintain a competitive performance compared to models such as VGGs or ResNet.

Model Optimization: When optimizing the CNN models, a grid search was used to determine the best hyperparameters that achieved a high classification accuracy for all classes. To mitigate overfitting and minimize loss, batch normalization and dropout layers are strategically placed before the flattening layer. The models contained a dense layer with 512 neurons for VGG-16 and 1024 neurons for VGG-19 and ResNet-50. The layers were initialized using *ReLU* activation and *he_normal* kernel initialization function. To enhance the generalization, an additional dropout layer was added after the dense layer. Table II, provides a comprehensive summary of the parameters for each model.

For MobileNetV2, a similar grid search approach determined the ideal number of neurons in the dense layers and was tailored for each radar preprocessing technique with consistent hyperparameters in the VGG and ResNet-50 models. The modifications to the optimizer and the learning rate are presented in Table II. The number of neurons was adjusted based on the unique characteristics of each 2D spectrum type: 2048 neurons were used for the noisy TR domain, 512 neurons were used for the high-definition SPWVD, and 1024 neurons were used for the STFT, which is known for its medium clarity and complexity. These adjustments were designed to prevent overfitting and optimize the model performance for the unique characteristics of each spectrogram. All CNN models adopted *categorical cross-entropy* as the loss function, prioritizing accuracy optimization.

TABLE II: Pre-trained CNN hyperparameters

Parameters	VGG-16	VGG-19	ResNet-50	MobileNetV2
Batch size	32	32	32	32
Dropout	0.5	0.2	0.2	0.2
Learning rate	2e-3	2e-3	2e-3	4e-4
Optimizer	SGD	SGD	SGD	Adam
Decay	-	1e-6	1e-6	-
Momentum	0.9	0.9	0.9	-
Epochs/Fold	25	25	25	25

3) Model Training: To optimize the training robustness, we used a stratified 15 k-fold cross validation (CV) technique. This method divides the training data into 15 subsets, each of which maintains an even class distribution. In each fold, 14 subsets were used for training, whereas one subset (10%) was used for validation. For each fold, we monitored the duration of the training and validation accuracy. The training process consisted of fitting the model to the training data using the validation split and saving the model weights with the highest validation accuracy across all folds.

III. Results and Discussion

A. Runtime Environment

In our study, we performed data preprocessing on raw radar data and CNN training using Python toolkits on a GPU-accelerated PC. The system was equipped with an 11th Generation Intel® Core™ i7-11700 processor with 8 cores and 16 threads enabled by hyper-threading technology and a base frequency of 2.50 GHz. It also equiped with 16 GB of RAM and an NVIDIA GeForce RTX 3060 Ti graphics card with 8 GB of memory. To generate radar spectra, we used Python libraries such as Scipy for signal processing, time-frequency for spectrum analysis, and fftpack for FFT execution. For CNN model training, we used the Keras and TensorFlow frameworks, taking advantage of the multicore and multithreaded CPU capabilities of the workstation and the parallel processing power of the GPU.

The experimental results, including inference time and test accuracy, are obtained by averaging five runs with random initialization to ensure that the measurements are stable and reliable.

B. Model Generalization on Multiple Spectrograms

Evaluating the generalization ability of HAR systems is essential, particularly for limited radar-based human activity datasets. Owing to a lack of data, CNN classifiers tend to overfit. Therefore, it is crucial to evaluate the performance of new unseen radar data. The evaluation results of our HAR models are shown in Fig. 3, showing consistent results, with the test accuracy showing the smallest variance across the 12 model-domain pairs (MDPs) (named M1, M2, M3, etc.), as shown in Table III. These findings indicate that the models have strong generalization capabilities, with test accuracies ranging from 92.88% to 98.01%, confirming their effectiveness on the new data. Model M10 achieved the highest test accuracy of 98.01%. On the other hand, although models M2 and M4 achieved perfect or near-perfect average training accuracies of 100% and 99.42%, respectively, they had lower test accuracies of 94.30% and 92.88%, indicating that there is room for improvement. This difference highlights the importance of thorough testing of unseen data to accurately determine how well a model adapts to new input. The results emphasize the importance of using CV and external test datasets to evaluate the model generalization in real-world scenarios. Furthermore, choosing the best MDP requires a balanced evaluation of its performance, based on its generalization ability and computational efficiency.

TABLE III: Real-Time Performance Comparison of Proposed HAR Models. P, R and F1 shows Precision, Recall and F1-score

MDPs	Domains	Models	Training Time/epoch (s)	Inference Time/sample (ms)	Accuracy (%)	P	R	F1
M1	TR	VGG-16	3.40	7.16	95.73	0.9576	0.9573	0.9572
M2	TR	VGG-19	3.77	8.11	94.30	0.9436	0.9430	0.9429
M3	TR	ResNet-50	2.77	3.80	93.73	0.9373	0.9373	0.9368
M4	TR	MobileNetV2	1.79	2.78	92.88	0.9307	0.9288	0.9284
M5	STFT	VGG-16	3.38	7.10	96.87	0.9697	0.9687	0.9687
M6	STFT	VGG-19	4.38	6.90	96.01	0.9639	0.9624	0.9624
M7	STFT	ResNet-50	2.74	3.54	97.15	0.9721	0.9731	0.9721
M8	STFT	MobileNetV2	1.49	2.57	96.30	0.9635	0.9651	0.9642
M9	SPWVD	VGG-16	3.50	7.02	97.44	0.9764	0.9744	0.9745
M10	SPWVD	VGG-19	3.76	6.88	98.01	0.9803	0.9801	0.9801
M11	SPWVD	ResNet-50	2.73	3.99	97.15	0.9720	0.9715	0.9715
M12	SPWVD	MobileNetV2	1.34	2.76	96.01	0.9629	0.9580	0.9600

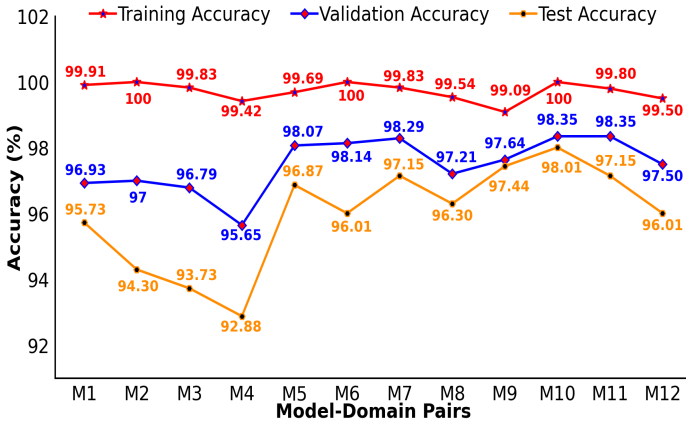


Fig. 3: Generalization capability of the proposed model-domain combinations.

C. Real-Time Performance Comparison of Proposed HAR Models

In this study, we compared the performance metrics of the 12 MDPs, as described in the above section, and the results are listed in Table III. The metrics used include 'training time/epoch' and 'inference time/sample', measuring the model training time per epoch and the time required to predict a test sample. For multiclass classification in HAR systems, we calculated several metrics to evaluate the effectiveness of the CNNs. These metrics go beyond accuracy and include the precision, recall, and F1-score, that provide a detailed evaluation of the classification performance by considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Based on these metrics, models M1, M7, and M10 were selected from the 12 MDPs listed in Table III. M1, M7, and M10 achieved the highest test accuracies in their respective domains (TR, STFT, and SPWVD) when used as inputs, thereby emphasizing their significance in radar-based HAR system. These MDPs also showed competitive results in terms of training and inference times, making them suitable for real-time systems. To evaluate the detection performance of each activity, confusion matrices were analyzed. Fig. 4a, 4b, and 4c shows the confusion matrices for models M1, M7, and M10, respectively. Notably, model M7 identified A6, representing fall activity, with 100% accuracy, whereas models M1 and M10 achieved 97.50% accuracy. All three models detected the A1 class, which represents walking activity, with 100% accuracy. Furthermore, the three best models (M4, M8, and M12) were selected in terms of inference time, which is a key metric for edge devices that require fast activity prediction, as explained in detail in the next subsection.

D. Computational Efficient and LightWeight HAR Model

Computational efficiency is critical for real-time radar-based HAR system, particularly for resource-constrained edge devices. Therefore, it is crucial to develop lightweight model that can quickly and accurately recognize human activities while minimizing the inference latency. This section examines the computational efficiency using time metrics, such as training time and inference latency, to evaluate the suitability of various models for real-world deployment, as detailed in Table III.

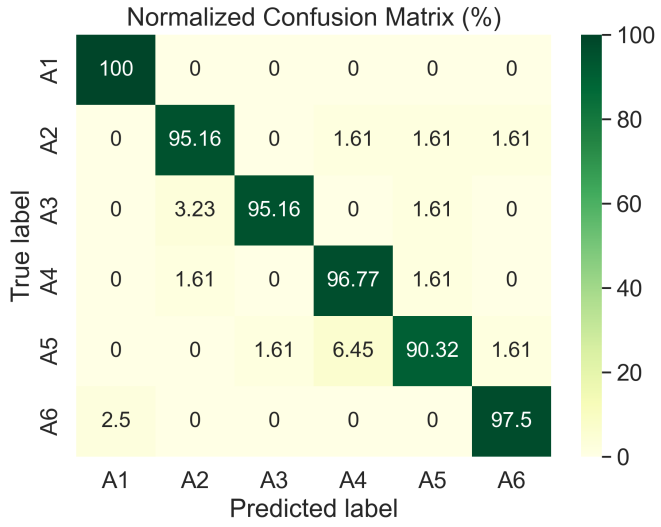
The inference time, or inference latency [30], is defined as the time between initiating a prediction request and receiving the prediction output from the test model. This metric is very important for evaluating the performance and efficiency of a model, particularly in applications that require real-time processing on edge devices or standalone systems. The inference time directly affects the user experience and applicability of the model in time-sensitive scenarios such as fall detection.

To measure the inference time ($T_{\text{inference}}$) of the model, we adopted the simple method from [30], focusing on the time required to perform a single inference cycle on the test set. Specifically, the inference time is calculated as follows:

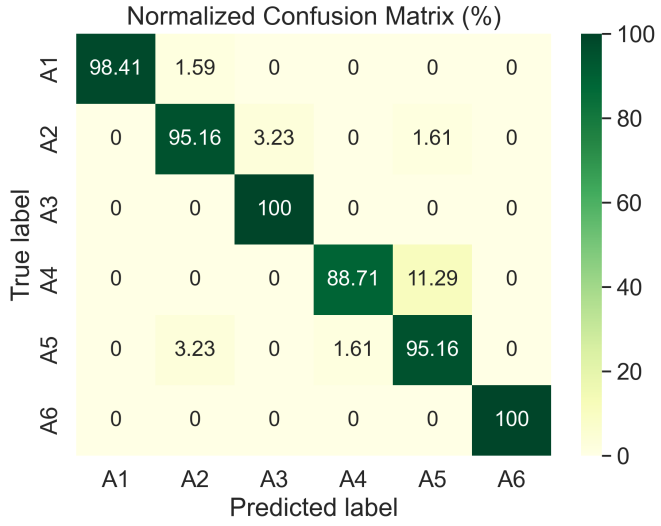
$$T_{\text{inference}} = T_{\text{end}} - T_{\text{start}} \tag{12}$$

Where T_{start} represents the timestamp when the inference request is issued and T_{end} represents the timestamp when the inference result is obtained.

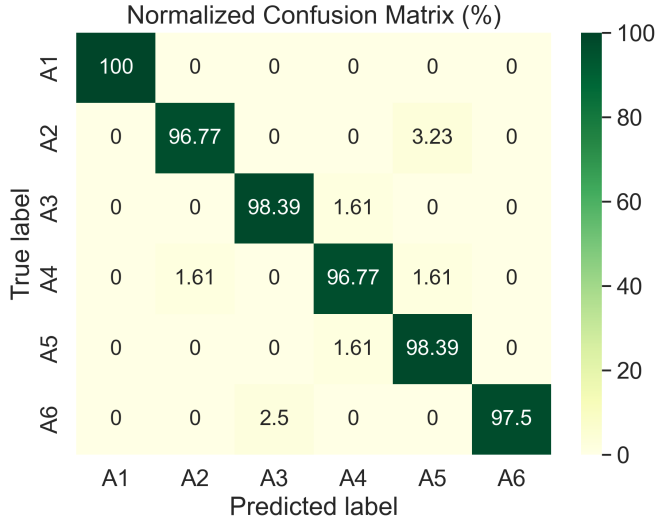
TR Domain: For the TR domain, the preprocessing time for processing the input raw radar data and visualizing an image representing the range over time as shown in Fig. 2, is only 0.035 s, which is very low compared to other



(a) Model M1



(b) Model M7



(c) Model M10

Fig. 4: Confusion Matrices for the Best Performing Models.

techniques. However, when inputted to a CNN model, it results in lower accuracy and higher computational cost. Among all the networks, MobileNetV2 exhibits the best training and inference efficiency, with a training time of 1.79 s/epoch, an inference time of 2.78 ms/sample, and a recognition accuracy of 92.88%. In contrast, the other three models are known for their higher accuracy but with increased time measurements.

STFT-based TD Spectrogram: The STFT-based spectrogram shows the change in frequency over time (Doppler shift), as shown in Fig. 2 and takes only 0.22 s to preprocess and generate a spectrogram using the STFT method. When this spectrogram is used as an input feature to the network, it provides a good balance between the performance and efficiency. For example, MobileNetV2 had a training time of 1.49 s/epoch and an inference time of 2.57 ms/sample, with a test accuracy of 96.30%. In contrast, VGG-16 and ResNet-50 achieved higher recognition accuracies of 96.87% and 97.15%, respectively, but had longer training and inference times, indicating higher resource usage. On the other hand, VGG-19 has a longer prediction time of 6.90 ms/sample, making it less suitable for real-time systems.

SPWVD-based TD Spectrogram: Despite the lengthy preprocessing time of 52.58 s to generate a spectrogram using the SPWVD method, MobileNetV2 had a lower training time of 1.34 s/epoch and an inference time of 2.76 ms/sample, when SPWVD was used as input compared to the TR and STFT domains. Due to the higher preprocessing time, SPWVD is not suitable for real-time systems that require rapid preprocessing and prediction response, from data acquisition to model prediction.

From the aforementioned discussions, it is clear that the best network suitable for a real-time radar-based HAR system deployed on edge devices is MobileNetV2 in combination with STFT as a feature. The MDP provides a balance between performance accuracy and computational efficiency, indicating its suitability for edge devices that require rapid activity recognition in real-time systems.

E. Comparison of Model M8 with State-of-the-art models

A detailed comparative analysis is presented in Table IV. All models used STFT-based spectrogram inputs with a resolution of 224 × 224 pixels. The CNN [31], model trained from scratch achieved 95.44% accuracy, but the inference time per sample was 5.14 ms, which is almost twice that of our proposed MobileNetV2 model. The CNN + LSTM [8], model had the shortest training time per epoch of 1.12 s, but the accuracy was only 84.90%, and the inference time was as high as 6.04 ms/sample, almost three times that of our proposed model. The Bi-LSTM [32], model achieved a competitive accuracy of 95.16% with an inference time of 2.77 ms/sample, lagging behind the MobileNetV2 model in both training and inference time.

The proposed MobileNetV2 model, leveraging TL, achieved the highest accuracy of 96.30% and the best inference time of 2.57 ms/sample, making it suitable for real-time applications. This demonstrates that MobileNetV2 not only surpasses the accuracy of other state-of-the-art models trained from scratch but also significantly reduces the inference time, providing

an efficient option for real-time processing. This comparison highlights the adaptability and potential of our model for a wide range of future applications, thereby setting a new benchmark for HAR systems in terms of both performance and efficiency.

TABLE IV: Time metrics and accuracy comparison of proposed lightweight model and alternative approaches

Model	Training Time/epoch (s)	Inference Time/sample (ms)	Accuracy (%)
CNN [31]	1.88	5.14	95.44
CNN+LSTM [8]	1.12	6.04	84.90
Bi-LSTM [32]	1.70	2.77	95.16
MobileNetV2	**1.49**	**2.57**	**96.30**

IV. Conclusion

In this study, we applied three preprocessing techniques such as: TR, STFT and SPWVD, as inputs to CNN models for HAR and evaluated their computational efficiency for edge deployment, resulting in 12 different combinations of model-preprocessing pairs. These combinations include VGG-16, VGG-19, ResNet-50, and MobileNetV2 architectures. Among them, the combination of MobileNetV2 with STFT (model M8) showed balanced performance, setting a new benchmark for state-of-the-art radar-based HAR system. This result emphasizes the importance of thorough evaluation of the entire process chain. The effectiveness of model M8 highlights its ability to support more advanced edge device models, which are typically associated with TinyML. Our work not only contributes to current methodologies but also lays the foundation for integrating more complex models into low-power, real-time edge systems.

Furthermore, in anticipation of advancements, our future research will focus on integrating neuromorphic federated learning and congestion-aware spiking neural networks to design energy-efficient systems, which is an important aspect not discussed in this study. This strategy aims to improve the real-time performance of radar-based HAR systems and address the trade-off between accuracy and energy efficiency.

References

[1] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, pp. 5281, 2023.

[2] World Health Organization, "Global Report on Falls Prevention in Older Age," World Health Organization Press, Geneva, Switzerland, 2008. Accessed: Oct. 20, 2020. [Online]. Available:Global Report on Falls Prevention in Older Age.

[3] National Institute for Health and Care Excellence (UK), "2019 Surveillance of Falls in Older People: Assessing Risk and Prevention (NICE Guideline CG161)," [Online]. Available: https://www.nice.org.uk.

[4] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring Activities of Daily Living in Smart Homes: Understanding human behavior," *IEEE Signal Processing Mag.*, vol. 33, no. 2, pp. 81-94, Mar. 2016.

[5] I. H. Lopez-Nava and A. Munoz-Melendez, "Wearable inertial sensors for human motion analysis: A review," *IEEE Sensors Jr.*, vol. 16, no. 22, pp. 7821–7834, 2016.

[6] H. Yu, Z. Chen, X. Zhang, X. Chen, F. Zhuang, H. Xiong, and X. Cheng, "FedHAR: Semi-supervised online learning for personalized federated human activity recognition," *IEEE Trans. Mobile Computing*, 2021.

[7] Z. Zhou, C. Wu, Z. Yang, and Y. Liu, "Sensorless sensing with WiFi," *Tsinghua Sci. and Technol.*, vol. 20, no. 1, pp. 1–6, 2015.

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

[9] X. Yang, F. Xiong, Y. Shao, and Q. Niu, "WmFall: WiFi-based multistage fall detection with channel state information," *Int. Jr. of Distributed Sensor Networks*, vol. 14, no. 10, pp. 1550147718805718, 2018.

[10] P. F. Moshiri, R. Shahbazian, M. Nabati, and S. A. Ghorashi, "A CSI-based human activity recognition using deep learning," *Sensors*, vol. 21, no. 21, pp. 7225, 2021.

[11] L. Tang, Y. Jia, Y. Qian, S. Yi, and P. Yuan, "Human Activity Recognition Based on Mixed CNN With Radar Multi-Spectrogram," *IEEE Sensors Jr.*, vol. 21, no. 22, pp. 25950-25962, 2021.

[12] A. H. Victoria and G. Maragatham, "Activity recognition of FMCW radar human signatures using tower convolutional neural networks," *Wireless Networks*, pp. 1–17, 2021.

[13] H. Arab, I. Ghaffari, L. Chioukh, S. O. Tatu, and S. Dufour, "A convolutional neural network for human motion recognition and classification using a millimeter-wave Doppler radar," *IEEE Sensors Jr.*, vol. 22, no. 5, pp. 4494–4502, 2022.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[16] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248-255.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[20] Z. Ardalan and V. Subbian, "Transfer learning approaches for neuroimaging analysis: a scoping review," *Frontiers in artificial intelligence*, vol. 5, pp. 780405, 2022.

[21] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, 2016.

[22] L. R. Triani, N. Ahmadi, and T. Adiono, "Human Activity Recognition Based on FMCW Radar Using CNN and Transfer Learning," in *Proc. APSIPA ASC*, 2023, pp. 248–253.

[23] F. J. Abdu, Y. Zhang, and Z. Deng, "Activity Classification Based on Feature Fusion of FMCW Radar Human Motion Micro-Doppler Signatures," *IEEE Sensors Jr.*, vol. 22, no. 9, pp. 8648-8662, 2022.

[24] M. Chakraborty, H. C. Kumawat, and S. V. Dhavale, "Application of DNN for radar micro-doppler signature-based human suspicious activity recognition," *Pattern Recognition Letter*, vol. 162, pp. 1–6, 2022.

[25] X. Zhang, J. Tian, and D. Hao, "A lightweight network model for human activity classification based on pre-trained mobilenetv2," in *Proc. IET Int. Radar Conf. (IET IRC 2020)*, 2020, pp. 1483-1487.

[26] F. Fioranelli *et al.*, "Radar signatures of human activities," Univ. Glasgow, Glasgow, U.K., 2019. [Online]. Available: https://researchdata.gla.ac.uk/848/.

[27] S. Yang, et al., "The Human Activity Radar Challenge: Benchmarking Based on the 'Radar Signatures of Human Activities' Dataset From Glasgow University," *IEEE Jr. Biomedical and Health Informatics*, vol. 27, 2023.

[28] F. Ayaz, B. Alhumaily, S. Hussain, L. Mohjazi, M. A. Imran, and A. Zoha, "Integrating Millimeter-Wave FMCW Radar for Investigating Multi-Height Vital Sign Monitoring," in *Proc. 2024 IEEE WCNC*, UAE, Apr. 21-24, 2024.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[30] S. Takano, "Chapter 2 - traditional microarchitectures," in *Thinking machines*, S. Takano, Ed. Academic Press, 2021, pp. 19–47.

[31] K. Papadopoulos and M. Jelali, "A Comparative Study on Recent Progress of Machine Learning-Based Human Activity Recognition with Radar," *Applied Sciences*, vol. 13, no. 23, pp. 12728, 2023.

[32] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with Bi-LSTM networks," *IEEE Sensors Jr.*, vol. 20, no. 22, pp. 13607–13619, 2020.