Continuous Human Activity Recognition With Distributed Radar Sensor Networks and CNN–RNN Architectures

Zhu, Simin; Guendel, Ronny Gerhard; Yarovoy, Alexander; Fioranelli, Francesco

DOI
10.1109/TGRS.2022.3189746

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Geoscience and Remote Sensing

Citation (APA)
Zhu, S., Guendel, R. G., Yarovoy, A., & Fioranelli, F. (2022). Continuous Human Activity Recognition With Distributed Radar Sensor Networks and CNN–RNN Architectures. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-15. Article 5115215. https://doi.org/10.1109/TGRS.2022.3189746

Continuous Human Activity Recognition With Distributed Radar Sensor Networks and CNN–RNN Architectures

Simin Zhu, Ronny Gerhard Guendel, *Graduate Student Member, IEEE*, Alexander Yarovoy, *Fellow, IEEE*, and Francesco Fioranelli, *Senior Member, IEEE*

***Abstract*— Unconstrained human activities recognition with a radar network is considered. A hybrid classifier combining both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for spatial–temporal pattern extraction is proposed. The 2-D CNNs (2D-CNNs) are first applied to the radar data to perform spatial feature extraction on the input spectrograms. Subsequently, gated recurrent units with bidirectional implementations are used to capture the long- and short-term temporal dependencies in the feature maps generated by the 2D-CNNs. Three NN-based data fusion methods were explored and compared with utilize the rich information provided by the different radar nodes. The performance of the proposed classifier was validated rigorously using the K-fold cross-validation (CV) and leave-one-person-out (L1PO) methods. Unlike competitive research, the dataset with continuous human activities with seamless interactivity transitions that can occur at any time and unconstrained moving trajectories of the participants has been collected and used for evaluation purposes. Classification accuracy of about 90.8% is achieved for nine-class human activity recognition (HAR) by the proposed classifier with the halfway fusion method.**

***Index Terms*— Deep learning (DL), distributed radar, human activity recognition (HAR), micro-Doppler signatures, radar sensor network (RSN).**

I. Introduction

OVER the past decades, radar systems have gained significant attention for short-range indoor scenarios, for example to localize and track moving humans [1]–[3], recognize various human behaviors [4]–[6], and even monitor their vital signs [7]–[9]. Compared with other sensors, such as inertial measurement units (IMUs) or RF tags, radar sensors can capture human motion without the participants carrying any device on them. Furthermore, unlike LiDAR and camera sensors, radar is robust against environmental light conditions. Moreover, it can measure range, velocity, and angular information simultaneously for each independent target. Last but not least, radar may alleviate end-users' privacy concerns in sensitive environments like the lavatory and bedroom.

Among various applications mentioned above, radar-based human activity recognition (HAR) has high societal importance. For example, in elderly care, HAR systems can be applied to detect life-threatening activities like "falling" [10]. Commonly, HAR is achieved by exploiting the features of human motions embedded in the Doppler-time (DT) domain (or spectrogram) [11]–[13]. Additionally, it is possible to use other radar data representations, for instance, the range-time (RT) domain [14], the range-Doppler (RD) domain [15], or the cadence velocity diagram (CVD) [16]. Nevertheless, much work still focuses on using spectrograms for HAR as they can capture the characteristics of individual body parts movement [17].

One of the first works that shows the feasibility of radar-based HAR has been presented in [18]. It uses the support vector machine (SVM) [19] as the classifier trained on a set of handcrafted features extracted from spectrograms. Subsequently, Kim *et al.* [20] proposed an alternative HAR system using deep learning (DL) [21] approaches to remove the requirement for feature extraction and selection steps. In their work, the proposed classifier is based on the convolutional neural network (CNN), which conducts an automatic feature extraction process and allows a hierarchical decomposition of the input data.

CNN treats the input spectrograms as optical images. It is well known that CNN exploits local correlations in the input data by restricting the receptive field of the hidden neurons in the network to be local [22], [23]. However, the Doppler spectrogram of human motion contains not only local correlations but also temporal dependencies [24]. To capture both, it is desired for the output neurons in the CNNs to have a sizeable receptive field so that no critical information will be lost [25].

Modern CNN architectures [26], [27] consider stacking many convolutional layers and using the subsampling technique to increase the size of the receptive field. However, deep CNNs (DCNNs) are notoriously data-hungry because of the significant amount of trainable parameters to be estimated [28]. Besides, to use CNNs for HAR, the input spectrogram is often assumed to contain only one type of activity [29]–[31].

Manuscript received 15 December 2021; revised 20 March 2022 and 22 May 2022; accepted 3 July 2022. Date of publication 11 July 2022; date of current version 18 July 2022. This work was supported by the Dutch Research Council NWO under project RAD-ART, Radar-aware Activity Recognition with Innovative Temporal Networks. *(Corresponding author: Simin Zhu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by TU Delft HREC under Approval ID 1387.

The authors are with the Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: s.zhu-2@tudelft.nl).

Digital Object Identifier 10.1109/TGRS.2022.3189746

To address the above-mentioned limitations in using CNNs for HAR, the recurrent neural network (RNN) and its variants [32] were proposed as alternatives [33]. RNNs are well-suited for processing temporal series due to their loop structure [34]. For example, an HAR system that considers the continuous nature of human activities is presented in [35]. This system uses the long short-term memory (LSTM) [36] network with its bidirectional implementation [37] as the classifier. Dissimilar to the previous works, in which different human activities were separated artificially during the data collection stage, the result shows that RNNs can be applied to process continuous sequences of human movements with seamless interactivity transitions. However, one downside of RNNs compared with CNNs is that they fail to capture the local structure of the input data due to their fully connected topology.

Apart from the limitations of specific network architectures, many research works investigated radar-based HAR tasks in a constrained fashion. More specifically, the directions of the target's movement are limited within a trajectory to mitigate the influence of unfavorable aspect angles on the system performance [38]. This is a simplification of realistic scenarios, where the targets can move freely with an arbitrary trajectory, and the HAR problem becomes unconstrained and more challenging [39]. In addition to the unconstrained trajectories, performed activities in real life are also continuous, with flexible duration and natural transitions between activities.

Multinode radar systems with different internodal data fusion methods were introduced to reduce the dependence of the classification accuracy on the aspect angle [40]. In general, such a radar sensor network (RSN) consists of several spatially distributed radar nodes, working in a multistatic [41] or distributed monostatic [42] framework. Compared with using a single radar for HAR, there is a higher chance that at least one radar node in the RSN can capture the human movement from a favorable aspect angle. With RSN, it is important to investigate the impact of different radar geometries and data fusion techniques. A simulation framework was proposed in [43] to benchmark the activity recognition performance in relation to various radar deployment geometries, and different data fusion strategies to handle the unconstrained HAR problem in a distributed RSN were presented in [44].

Despite these initial studies, to the best of our knowledge, very few works in the literature have investigated radar-based unconstrained HAR with classifiers combining both CNNs and RNNs for spatial–temporal pattern extraction and verified such approaches on experimental data from a distributed RSN. From this perspective, the main contributions of this article are as follows.

1) The proposed classifier provides an end-to-end solution for automatic HAR in an RSN. It exploits a hybrid CNN-RNN architecture to leverage the merits of both NNs. The 2-D CNNs (2D-CNNs) are first applied to perform spatial feature extraction on the input spectrograms. Subsequently, gated recurrent units (GRUs) [45] with bidirectional implementations are used to capture the long- and short-term temporal dependencies in the feature maps generated by the 2D-CNNs. Unlike several previous works which only explored one type of NNs, the proposed classifier can capture the spatial–temporal characteristics of human motion to classify continuous human activities with seamless interactivity transitions.
2) Three types of NN-based data fusion methods are explored, i.e., the *early fusion* (signal-level fusion), *late fusion* (high-level fusion), and *halfway fusion* (intermediate-level fusion), in order to exploit the multiperspective views and data provided by the RSN for unconstrained HAR. Additionally, a weight-sharing (WS) technique is implemented across all feature extraction channels. This leads to a lightweighted NN with a fewer number of trainable parameters and better classification performance.
3) Unlike some previous works, this article considers a more realistic and challenging HAR scenario. In particular, during the data collection stage, the participants were allowed to move freely in the measurement area while performing continuous activities. Hence, each radar recording contains multiple human activities with natural interactivity transitions. In total, nine types of activities are collected and performed by 14 participants. For model evaluation, the K-fold cross-validation (CV) method and the leave-one-person-out (L1PO) method are implemented together to utilize the limited dataset more efficiently and estimate the model performance more rigorously. The result shows that maximum classification accuracy of 90.8% is reached on the unseen test dataset.
4) Considering realistic HAR scenarios, the performance of the proposed classifier is further assessed in three practical aspects, namely, the system generalization capability, the influence of imperfect tracking, and the system scalability. For the generalization capability, it is shown that the proposed classifier can generalize well on the unseen dataset of a new participant with a classification accuracy of 85.1% averaged over 14 participants. Regarding the impact of imperfect tracking on the HAR problem, it is shown that the proposed classifier can maintain a stable classification performance even when the missed detection rate is set to as high as 2E-1 (essentially equivalent to the loss of several time bins of spectrograms for classification). Finally, the proposed hybrid CNN-RNN classifier is shown to be able to process a variable number of radar inputs after network training. In other words, there is no alignment limitation if the number of actually deployed radar nodes is different from those considered in training, showing that the classifier has up/down scalability.

The rest of the article is structured as follows. First, the experimental setup and the used dataset are explained in Section II. Next, the design details related to the proposed HAR system are presented in Section III. Then, the system performance is evaluated using the experimental data in Section IV. After that, three practical aspects related to

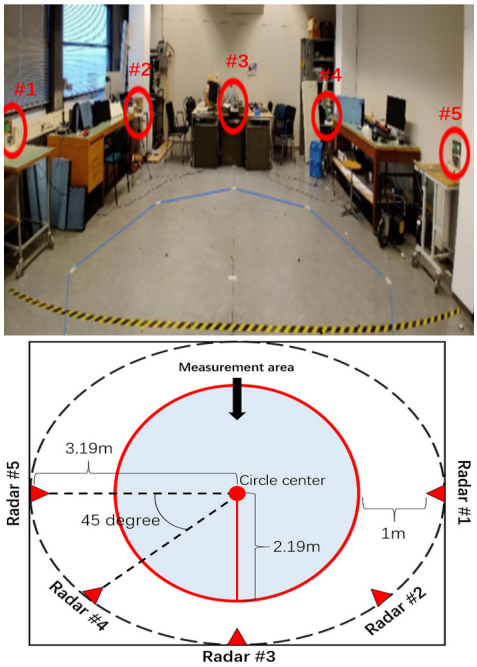

Fig. 1. Layout of the distributed RSN at TU Delft MS3 Laboratory [44]. It consists of five identical pulsed UWB radars placed circularly around the measurement area. The measurement area is a circle with radius equals to 2.19 m.

realistic HAR scenarios are inspected in Section V. Finally, conclusions and potential future directions are provided in Section VI.

II. Experimental Setup and Dataset Description

This section presents the experimental setup of the homogeneous RSN [44] and an overview of the used dataset [46].

A. Experimental Setup

Fig. 1 shows the layout of the distributed RSN, in which the top picture illustrates the position of the radars at TU Delft MS3 Laboratory, Delft, The Netherlands, and the bottom one provides the schematic of the radar deployment geometry.

In the RSN, five identical pulsed radio ultrawideband (UWB) [47] radar sensors were used. Compared with other types of radar sensors, for example, the popular frequency-modulated continuous-wave (FMCW) radar, the UWB radar is also suitable for indoor sensing [48]. Also, due to its large operational bandwidth, it can provide extraordinarily high range resolution and localization capability. Furthermore, UWB radar is robust against the multipath and fading effect and has the advantage of low power consumption, compact installation size, and affordable price. The UWB radar used in this article is provided by Humatics (PulsON P410). It operates at a center frequency of 4.3- with 2.2-GHz bandwidth. The pulse repetition frequency (PRF) of the radar is 122 Hz, or 8.2 ms in terms of pulse repetition interval (PRI), with a resulting maximum unambiguous velocity of around ± 2.13 m/s. Each radar has two omnidirectional broadband antennas, and they work in a monostatic mode with a computer synchronizing the simultaneous operation of all nodes.

B. Dataset Description

The dataset used in this work is publicly available; readers interested in more detail and visualizations can refer to [46]. The human movements considered in the dataset are designed to simulate unconstrained human activities of daily living (ADL). The pursuit of unconstrained HAR is crucial, not only because of the continuous nature of human motions but also for the future development of home monitoring systems for older and possibly frail residents [10]. Overall, nine types of activities are performed (labeled from 1 to 9) and 14 participants joined the data collection process (labeled from A to N). To have continuous human activities with seamless interactivity transitions, each participant is asked to perform a combination of the designed activities inside the measurement area. In total, seven combined sequences of the designed human movements were constructed, and each combination was performed four times by every participant. These combined sequences include the following.

1) "Walking" (label 1) versus "Stationary" (label 2).
2) "Sitting-down" (label 3) versus "Standing-up" (label 4).
3) "Bending from sitting" (label 5) repeated in a sequence.
4) "Bending from standing" (label 6) repeated in a sequence.
5) "Falling-from-walking" (label 7) versus "Standing-up-after-falling" (label 8).
6) "Falling-from-standing' (label 9) versus "Standing-up-after-falling" (label 8).
7) "Mix of all nine activities" (contains labels from 1 to 9).

While performing different activities, participants were free to determine transition points between two consecutive activities, and these transitions could occur multiple times during the 2-min (120-s) recording time of each sequence. This is expected to increase the diversity of the transitions between activities in the collected data, as opposed to data collection approaches that may require participant time or some degree of control over transitions [42]. Moreover, participants can choose their own trajectory for translational activities, such as "walking." Likewise, for in-place motions, such as "Sitting-down," participants can change their body orientation, as well as the aspect angle to the lines of sight of the radar nodes.

An example of the recorded data for one of the "Mix of all nine activities" sequence is shown in Fig. 2.

III. Methodology

This section provides the design details of the proposed HAR system.

A. System Overview

Fig. 3 provides an overview of the proposed HAR system. This consists of five blocks, including: 1) the data preprocessing block; 2) the CNN block; 3) the data fusion block; 4) the RNN block; and 5) the fully connected NN (FCNN) block.

First, the data preprocessing block transforms the raw radar data into the input data (spectrograms). Next, the CNN blocks extract spatial features from the spectrograms and generate intermediate feature maps. Then, the data fusion block combines the maps from five convolutional channels and selects

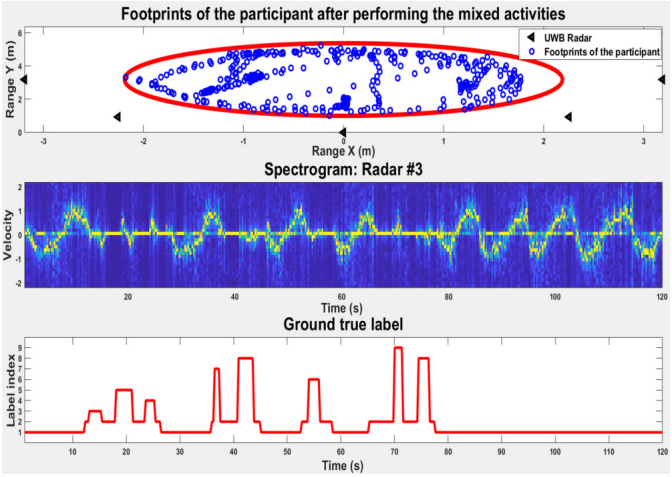

Fig. 2. Example of the recorded data with nine human activities performed continuously, with seamless interactivity transitions. The top figure shows the footprints of the participant during his movement in the measurement area. The middle figure presents the spectrogram extracted from radar #3. The bottom figure provides the ground truth labels for the target's movements.

the most prominent features. After that, the RNN block is used to capture the temporal characteristics from the feature map after fusion. Finally, the FCNN block maps the high-level feature representations to the final predictions.

B. Data Preprocessing Block

The main objective of the data preprocessing block is to process the raw UWB radar data and format it as input data for model training, validation, and testing. This consists of three steps: 1) data alignment; 2) target localization; and 3) feature extraction. In this section, we summarize these three preprocessing steps. However, readers interested in additional details of specific processing methods for UWB radar waveforms used in this work are referred to [49], [50].

The UWB radars in the proposed RSN work as monostatic nodes, each with a local oscillator. Since small inaccuracies in the central synchronization process may cause a slightly different start and end sampling times, it is important to first align the data from the five radar nodes so that the detection of human motion can be fused later. As the global acquisition start and end timestamps are provided for each radar, data alignment can be accomplished by simply selecting the collected data of the five radars within the latest acquisition start time and the earliest acquisition end time.

Generally, knowing the target's position is a necessary condition for extracting its Doppler signatures [51]. However, integrating tracking algorithms with the HAR task will directly increase the system complexity. Furthermore, as discussed in Section V-B, imperfect tracking can influence the quality of the generated input data. Therefore, this article uses an alternative solution that considers the Doppler effect in the fast-frequency/slow-time domain [49].

The alternative method is similar to the phase-based demodulation method described in [50]. It first implements the fast Fourier transform (FFT) on the collected RT map along the fast-time dimension. FFT converts the received pulses from the fast-time/slow-time domain to the fast-frequency/slow-time domain, where "fast-frequency" represents the frequency spectrum of each individual UWB pulse. Assuming that the pulses generated by the UWB radar are identical and that the bandwidth and the center frequency of the pulses are known, tracking moving targets can be simplified by locating the frequency bins that contain the phase variations caused by human motion. This method is specifically used in UWB radar, allowing only one target to appear in the sensing area.

In the feature extraction stage, the short-time Fourier transform (STFT) is implemented to process the extracted frequency bins and generate the spectrogram. Since the complete spectrogram contains human motions of 2 min, it is further segmented into several small, nonoverlapping segments. The final input data to the proposed NN has a size of 32 by 30, where 32 is the Doppler axis and 30 is the time dimension (262.4-ms interval between two timestamps). It should be noted that this article does not provide detail analysis of the input size, focusing on the network structure and fusion capabilities.

Finally, it is worth mentioning that the data preprocessing discussed here aims to remove the dependencies between the tracking and classification problems. However, since the ultimate goal for many HAR tasks is joint tracking and classification, it might be rewarding as future work to replace the preprocessing block with a multiple target tracking (MTT) system with subsequent feature extraction, as illustrated in Section V-B.

C. CNN Block

In this work, the stacked 2D-CNN is used to perform a stepwise feature extraction process on the input data. This can automatically generalize spatial features without "manual feature engineering." In addition, one advantage of using CNNs is that they are invariant to the translation of frequency changes [52]. This is a crucial feature because the Doppler characteristics of human activities can translate into the time and frequency dimensions.

Fig. 4 shows the design details of the CNN block. Each input source from a radar node is processed by a CNN block. A total of five CNN blocks with the same architecture are used. The CNN block contains three 2D-CNN layers for extracting spatial features and one 1×1 convolutional layer (also known as network-in-network (NIN) [53]) for depth reduction. Since the input data came from five identical UWB radars, a WS technique [54], [55] is implemented on the five CNN blocks. This WS approach allows the five blocks to have the same architecture and share the same weights, thereby reducing the number of model parameters and mitigating potential overfitting problems. It should be noted that the WS approach does not imply that the data from the five radar nodes are of equal importance to the final prediction (this is actually decided in a later data fusion block), but the WS uses the same parameters for the initial feature extraction part.

In the CNN block, each 2D-CNN layer is usually followed by a batch normalization (BN) [56] layer, a nonlinear activation layer, and a pooling layer. BN is used to standardize the

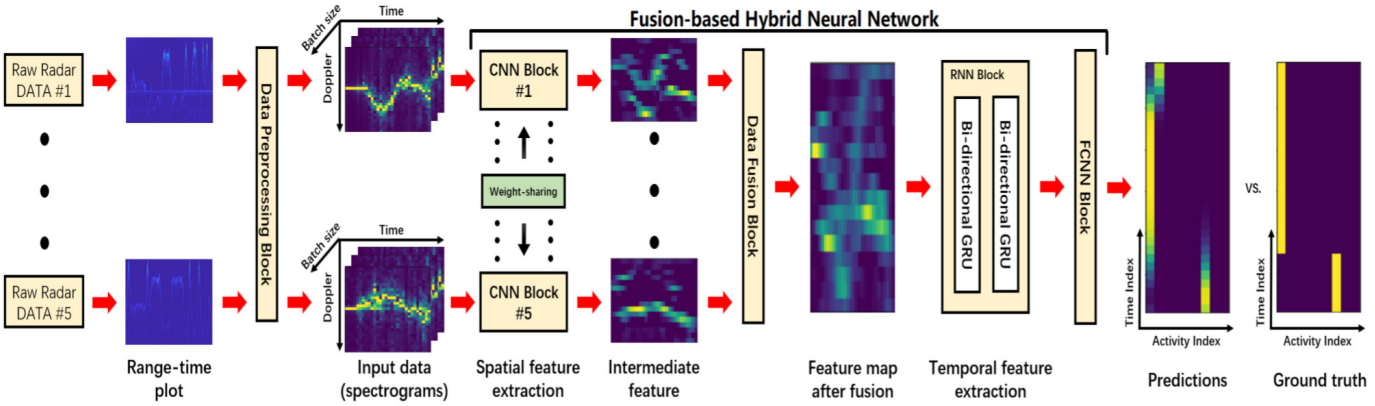

Fig. 3. Architecture overview of the proposed HAR system. On the leftmost are the raw radar inputs from five UWB radar sensors. The inputs to the CNN block are the spectrograms. The outputs of the HAR system are predictions, which provide information about the performed activities.

feature maps generated by the 2D-CNN layer. It implements normalization and scaling steps on the feature maps and readjusts the mean and standard deviation through two trainable weights. As shown in [57], BN can speed up the model training process. In addition, it can reduce generalization errors and alleviate the problem of internal covariate shifts [58].

After BN, nonlinearity is introduced in the NN by adding a nonlinear activation layer. This enables the network to learn complex relationships between input and output. It converts the learned linear mapping into a nonlinear form for propagation in the hidden layer, or predictions in the output layer [59]. In the proposed CNN block, the rectified linear activation function (ReLU) [60] is used to perform an element-level nonlinear transformation on the data passed from the BN layer.

In addition to the BN and activation layer, the first 2D-CNN layer has an additional layer, the pooling layer. This can provide additional translational invariance to the CNN block [61]. Besides, it offers an inexpensive method to quickly increase the receptive field of the CNN block and reduces the computational cost. In this work, maximum pooling (*MaxPool*) is added after the first 2D-CNN layer, and it only applies down-sampling in the frequency dimension [58], [62].

Finally, the output of each CNN block is an intermediate-level feature map, which can be further processed by subsequent processing blocks.

D. Data Fusion Block

Fig. 5 illustrates the implementation of the data fusion block. It uses the *halfway fusion* method to combine the rich information of human motion observed by the five radar nodes. This method first concatenates the intermediate feature maps provided by the five CNN blocks to form a data cube. Then, a channelwise MaxPool is performed to select the most representative features. After that, the data cube is compressed into a feature map again and sent to the RNN block.

There are several advantages of using the *halfway fusion* method. First, it can be used to fuse data from heterogeneous sensor systems. It also supports an end-to-end data fusion pipeline without the need for multistage model training. However, the *halfway fusion* method is often hard to train and has

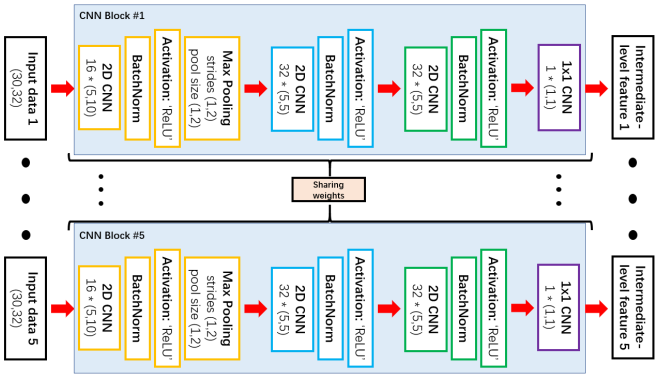

Fig. 4. Architecture of the five CNN blocks. Each of them consists of three 2D-CNN layers for feature extraction and one 1×1 convolutional layer for channel reduction. In addition, the weighted-sharing technique is applied to the five CNN blocks.

many training parameters. This is because the prediction error needs to be back-propagated through multiple CNN blocks to adjust their weights. In this work, the use of homogeneous RSN and WS technique alleviate the shortcomings of the *halfway fusion* method.

In addition to the *halfway fusion* method, two popular data fusion strategies, *early fusion* and *late fusion* are also considered and compared in this work. As the name suggests, *early fusion* combines the multisensor data in the early stages of processing. In this article, the *early fusion* method concatenates the input spectrograms into a cube and then feeds them to the CNN block. Different from *halfway fusion*, the NN with *early fusion* has lower complexity because it only needs one CNN block. However, if the input data have different formats, such as fusion of optical and audio inputs, or even radar data of different formats and dimensions, the *early fusion* method is not applicable.

Contrary to *early fusion*, *late fusion* merges the data in the later stages of processing. The input spectrogram from each radar node is first processed independently by the proposed CNN and RNN blocks. After that, the generated high-level feature maps are connected, and the final prediction is achieved through the FCNN block. As a result, the *late fusion* method

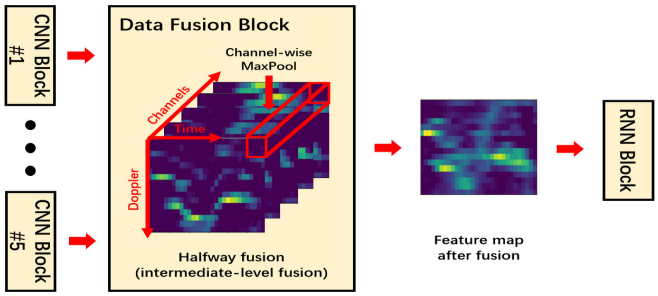

Fig. 5. Illustration of the data fusion block. It first concatenates the feature maps generated by the five CNN blocks. Then, it applies a channelwise MaxPool to select the most representative features.

allows the fusing of data from heterogeneous sensors. In addition, it permits single-modal pretraining. However, the *late fusion* method is time-consuming because it requires multiple stages of pretraining.

In short, the main difference between the three NN-based data fusion methods lies in the way and location where the data sources are fused. Because of this, they allow feature interaction between the five input channels at different levels.

E. RNN Block

Although the CNN block can take advantage of the local correlation from the input spectrogram in the time and frequency dimensions, DCNNs are usually required to capture long-term temporal dependencies. However, DCNNs are difficult to design and require a lot of training data. Therefore, the RNN block is introduced after the data fusion block to directly model the signal in time.

Three types of RNNs are studied in this work, including: 1) the simple RNN [32]; 2) LSTM [36]; and 3) GRU [45]. The simple RNN is similar to a single neuron but has cycles in the time dimension. It calculates the hidden state vector of the next timestamp and the output vector of the current timestamp based on the previous hidden state and current sequence input. The loop continues and the hidden state propagates until the last data sequence is processed. Therefore, for the output vector of a given timestamp, the simple RNN uses not only the current input but also the previous hidden state containing the information of all past input. It should be noted that the propagation of hidden states over time is a key characteristic of RNN because continuous human activities are time-dependent.

However, due to the vanishing gradient problem [63], simple RNNs are not good at capturing long-term dependencies. LSTM and GRU are two variants of simple RNN, which can effectively handle long-term temporal correlation. LSTM and GRU have a loop architecture similar to the simple RNN. They process the input data in turn and propagate the hidden state vector forward in time. Unlike simple RNNs, LSTM and GRU introduce several gates, each of which regulates the data flow into and out of the network and learn to keep only relevant information. The special gate architecture of LSTM and GRU alleviates the problem of gradient disappearance and improves the training speed. As for differences between LSTM and GRU, first of all, they have similar architectures, and both use gates to control the update of the hidden state vector. However, GRU has fewer gates and trainable parameters. Hence, it runs faster than the LSTM during model training. Nevertheless, LSTM and GRU are often used in research, because it is difficult to say which is more effective for specific problems and data.

It should be mentioned that the RNNs discussed earlier are unidirectional. In other words, the output of these RNNs at a given timestamp is based on the current and previous model inputs. However, human activities are time-dependent in both front and rear directions, such as the continuous movement of "walking," "falling-from-walking," and "standing-up-after-falling." For example, to classify a "fall," not only the Doppler signatures of the past "walking" motion will help, but the future "standing-up-after-falling" will also help. In order to solve this problem, this article considers the bidirectional implementation [37] of the RNNs discussed earlier. Compared with the unidirectional RNNs, the bidirectional RNNs have an additional reverse loop layer. The forward layer takes the input sequence forward in time, and the reverse layer takes the input sequence backward in time. Then, the output vectors from the forward and reverse layers are concatenated. Therefore, the output of the bidirectional RNNs at a given timestamp is based on the past, present, and future inputs.

F. FCNN Block

The FCNN block appears at the end of the proposed NN and is used to make the final prediction. It consists of three dense layers and one dropout layer [64]. For the dense layer, it usually contains multiple neurons. Each neuron in a dense layer is connected to all neurons in its adjacent layer. Due to its special architecture, the dense layer can learn how to build a mapping between input and output.

In this work, the input data of the FCNN block is the high-level feature map generated by the RNN block. In order to predict the human activity at each time step, the FCNN block is distributed in the time dimension. In other words, the FCNN block processes the vectors in the feature map one at a time. Finally, the Softmax activation function is used to calculate the probability vector of human activity changes over time.

G. Implementation Details

In this work, the proposed classification model is developed using the TensorFlow platform [65]. All model training, validation, and testing experiments are performed on a single-core Intel Xeon CPU at 2.2 GHz and a Tesla T4 GPU running with CUDA [66].

The proposed classifier is trained from scratch. The mini-batch gradient descent method [67] optimized by adaptive moment estimation (Adam) [68] is applied to model training. The training data is equally divided into batches with a batch size of 32. During training, the initial learning rate is set to 1E-3, and once learning stalls, a callback function is used to reduce the learning rate. Although the model can be trained for up to 100 epochs, the early stopping strategy [69] is used to avoid over-training the neural network.

TABLE I

COMPARISON OF VALIDATION RESULTS OF THE PROPOSED CLASSIFIER WITH A DIFFERENT NUMBER OF CNN AND RNN LAYERS

Depth of CNN	Validation Acc. (%)	std	Parameters
1	84.9%	1.72	32K
2	84.9%	1.94	45K
3	**87.1%**	**1.11**	**71K**
4	86.2%	1.06	96K
5	86.3%	1.35	122K
Depth of RNN	Validation Acc. (%)	std	Parameters
1	84.0%	1.61	32K
2	**87.1%**	**1.11**	**71K**
3	87.0%	1.02	90K
4	86.8%	1.07	108K

Thanks to the lightweight network architecture, the total training time for tuning the 71 K trainable parameters is about 2 min, and the average runtime latency for computing one prediction is about 78.9 ms (or 128.6 ms using only the CPU). Considering that the input data (spectrogram) provided to the classifier has an update rate of 262.4 ms (equivalent to 32 radar pulses), the time margin when using a single-core CPU is about 133.8 ms. Although the real-time implementation of the proposed classifier is not the goal of this article, this time margin seems to suggest that the proposed method is real-time implementable given the prediction rate reasonably required for indoor human monitoring.

IV. RESULTS AND DISCUSSION

This section presents the evaluation results of the proposed classifier for various model configurations.

For the train-validation-test dataset split, the K-fold CV and L1PO methods are used in combination. First, the L1PO method randomly holds out the data of one participant among the 14 as the testing data. Next, the fivefold CV method partitions the remaining data into the training and validation sets. Then, the classifier and its hyperparameters are trained and adjusted based on the training and validation data. Finally, the optimized classifier is assessed using the testing data.

The default model architecture is provided in Section III and, unless specifically mentioned, is applied to all experiments presented here.

A. Hyperparameter Tuning

The depth of the CNN and RNN blocks are two crucial hyperparameters in the hybrid CNN-RNN architecture. Table I shows the validation accuracy of the classifier as a function of the number of CNN and RNN layers.

Based on these results, deepening the CNN and RNN block can improve the system performance in terms of validation accuracy and standard deviation (std). This is because a deeper CNN and RNN architecture enables the classifier to learn more complicated spatial and temporal characteristics. However, further increasing the depth of both NNs beyond a certain threshold will not continue to benefit the performance. This is because the more complex the model the higher the number of model parameters, potentially leading to the overfitting problem, especially when the available data for model training is limited, which is a typical condition for radar-based experimental HAR. Finally, it is worth mentioning that similar trends have also been observed in [61], [70], [71] for speech recognition tasks using the CNN-RNN architecture.

TABLE II

COMPARISON RESULTS OF THE PROPOSED CLASSIFIER WITH/WITHOUT USING THE WS TECHNIQUE

Weight-sharing (WS)	Validation Acc. (%)	std	Parameters
With WS	**87.1%**	**1.11**	**71K**
Without WS	83.7%	0.89	229K

B. Impact of the Weight-Sharing Method

In radar-based HAR tasks, the available dataset for system development and training of algorithms often has a small size due to the demanding efforts of data collection and labeling. Thus, the complexity of the proposed NN becomes a critical factor to consider since a complex NN can easily overfit the limited training data and learn irrelevant information.

To address this issue, a WS method is proposed and applied to the five convolutional blocks. Table II shows the impact of the WS method on classification performance and model complexity in terms of a number of parameters. With this technique, the number of trainable parameters has reduced significantly from 229 K to 71 K. Moreover, the model achieves a higher classification accuracy. Despite the advantages, it should be noted that implementing the WS method requires a "homogeneous" sensor network consisting of identical sensor nodes.

C. Impact of Different Types of RNNs

In the proposed hybrid architecture, RNNs are responsible for capturing the temporal dependencies within the input data. To find the best-suited RNN, Fig. 6 shows the comparison results of the proposed classifier with six different RNN architectures. More specifically, the simple RNN is compared with its two variants, i.e., GRU and LSTM. All three types of RNNs are also implemented with their bidirectional implementation for comparison.

As shown in the results, all RNNs with a bidirectional architecture outperform the corresponding unidirectional RNNs. This indicates that both forward and backward temporal characteristics are beneficial for radar-based HAR and should be captured. However, the simple RNN does not gain as much improvement as LSTM and GRU with the bidirectional architecture. This shows that the long-term dependencies, which the simple RNN fails to capture due to the vanishing gradient problem, are also crucial features for improving the model performance in this application.

D. Impact of Different Data Fusion Methods

For a distributed RSN, it is necessary to consider various types of data fusion methods to leverage the best performances from the combination of information from different radar nodes. Fig. 7 presents the comparison results of the proposed

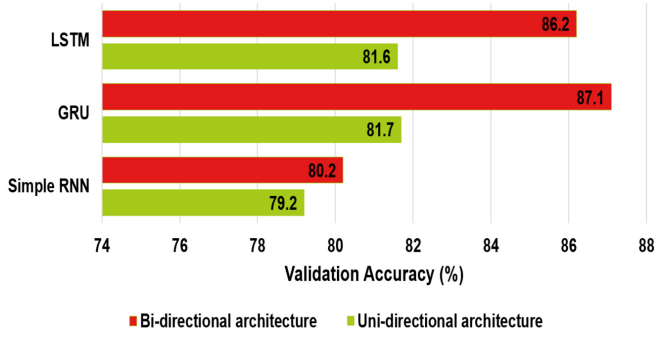

Fig. 6. Comparison results of the proposed classifier with three popular RNN architectures and their bidirectional implementations.

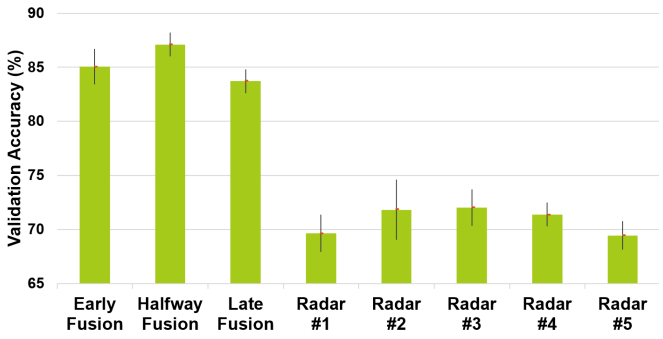

Fig. 7. Comparison results (accuracy bars and std interval) of the proposed classifier with three types of data fusion strategies: *early fusion*, *late fusion*, and *halfway fusion*. The three fusion methods are further compared to the case when only the data from a single radar is used (denoted "*Radar #N*," *N* indicates the radar index).

classifier with three NN-based fusion strategies, i.e., *early fusion*, *late fusion*, and *halfway fusion*. Moreover, the system performance with data fusion is compared with the case when only the data from a single radar is used (i.e., "No Fusion").

Based on the results, it is evident that all three fusion techniques outperform the "No Fusion" case. This demonstrates that using a distributed RSN with data fusion methods for solving unconstrained HAR problems is superior to using just a single-radar setting. Moreover, we can see that the system performances for the single-radar case are similar to each other. This is related to the fact that the participant's movements inside the measurement area are nearly random, and the chances for each radar node to have a good or bad aspect angle are approximately equal. This randomness can also be observed from the example of movements' footprints in Fig. 2.

Apart from that, it is noted that the *halfway fusion* method outperforms its two counterparts with regard to classification accuracy and std. This indicates that the *halfway fusion*, which allows fusing intermediate-level convolutional features, is the best choice of fusion scheme for the problem at hand. It provides a balance between combining fine DT details and high-level feature representations.

Finally, the relationship between the recognition accuracy and the number of radars is shown in Fig. 8. It is clear that as more radar sensors are used together, the recognition rate improves. Also, if we compare the use of a single radar with

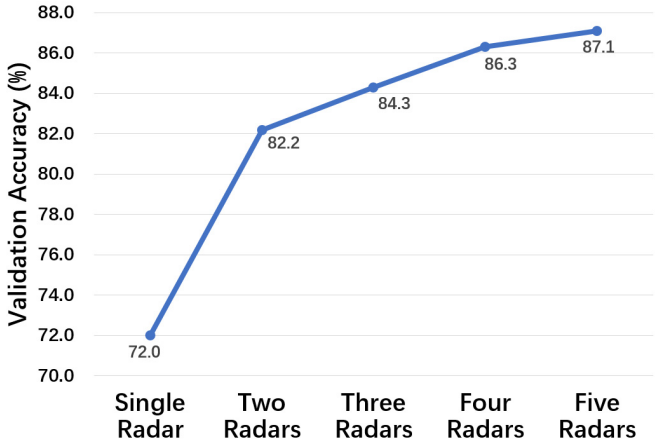

Fig. 8. Relationship between recognition accuracy and the number of radars. Due to a large number of possible radar combinations, only one combination is chosen for each number of nodes: "Single Radar" uses radar #3; "Two Radars" uses radars #2 and #4; "Three Radars" uses radars #1, #3, and #5; and "Four Radars" uses radars except radar #5. For experiments related to fusing multiple radars, the *halfway fusion* method is used.

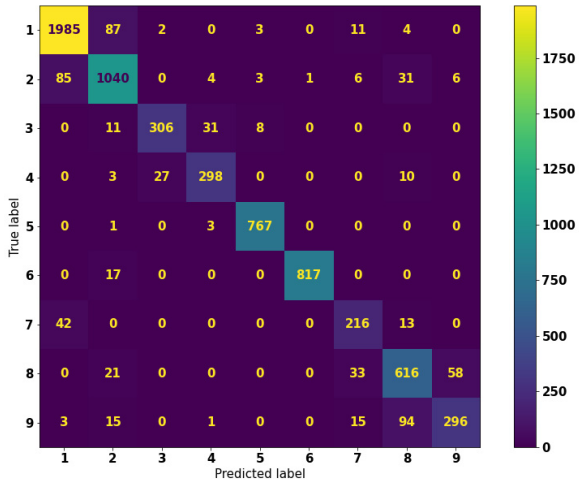

Fig. 9. Confusion matrix of the proposed classifier for nine-class HAR obtained using the testing data. The rows in the confusion matrix represent the actual labels, whereas the columns represent the predicted labels.

the use of two radars, the performance gain is most significant, but it becomes progressively more difficult to considerably increase performance while adding more sensors ("law of diminishing returns"). Beyond the simple number of radar sensors used together, it may be an interesting work to study the effects of different geometries of these nodes [43], but this is beyond the scope of this article.

E. Performance Evaluation of the Optimized Network

After hyperparameter tuning and searching for the optimal NN architecture, the performance of the optimized classifier is studied in this section.

Fig. 9 provides the confusion matrix of the proposed classifier to check if there are misclassifications in distinguishing two different classes. Fig. 9 shows that the proposed classifier performs generally well in making correct predictions for

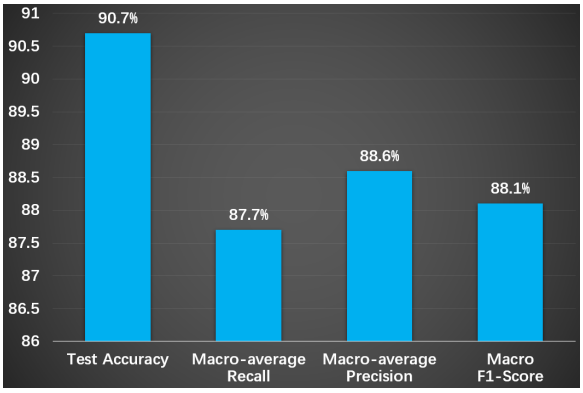

Fig. 10. Performance of the proposed classifier evaluated using accuracy and three additional metrics. All results are measured using the testing data with the L1PO method.

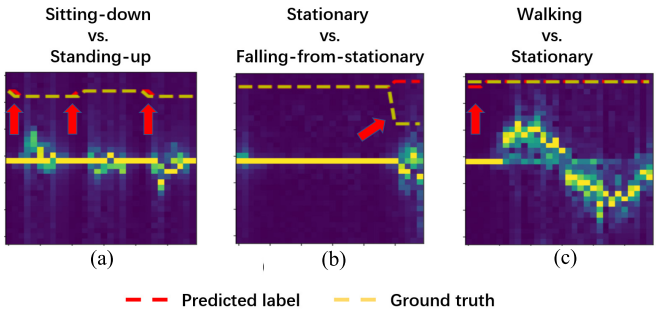

Fig. 11. Three primary causes of the prediction errors in the proposed classifier. From left to right: (a) transition error, (b) boundary error, and (c) labeling error.

each type of human activity. However, three error patterns are observed. They are: 1) "Walking" versus. "Stationary" (label 1 and 2); 2) "Walking" versus "Falling-from-walking" (label 1 and 7); and 3) "Standing-up-from-falling" versus "Falling-from-stationary" (label 8 and 9). Further analysis of the error patterns is discussed in Section IV-F.

In Sections IV-A–IV-D, the accuracy metric is the primary evaluation metric used to measure the performance of different network configurations and hyperparameters, which is a popular choice widely used in the literature. However, for multiclass recognition tasks, such as the one at hand, the accuracy metric cannot reflect the model performance for each class and can hide significant classification errors for minority classes if the dataset is imbalanced. In this case, the dataset is imbalanced as there is a majority of walking compared with in-place activities and relatively fewer falling cases than other activities.

Fig. 10 provides the classification accuracy of the optimized classifier and three additional metrics, including: 1) macro-average recall; 2) macro-average precision; and 3) macro F1-score. The macro-average recall and precision reflect the averaged recall and precision performance across different classes. In contrast, the macro F1-score aggregates the macro-average recall and precision to indicate the model's overall performance. As shown in the results, even though the collected data is imbalanced for each class, the proposed classifier is able to learn complex feature patterns using limited data and achieve acceptable performance on the additional metrics (i.e., 88.1% macro F1-score), without having significant errors in the less frequent classes.

F. Error Analysis

To further improve the performance of the proposed classifier, it is necessary to understand the reasons behind the errors that the classifier made. Hence, this subsection analyzes the error patterns based on an inspection of the testing data and the prediction results.

Fig. 11 illustrates three primary error patterns generated by the proposed classifier. Among them, the most common error is the *transition error* [shown in Fig. 11(a)]. It might happen when the target translates its activity from the current type to another but the classifier is confused in determining the exact transition point in time. It should be noted that this potential confusion exists already in the labeled ground-truth data, as it is difficult (even for the participants themselves) to define exactly the transition instant between two consecutive activity types. Nevertheless, one would like to have the prediction of the classifier as close as possible to the transition instant indicated in the ground truth.

The second type of error is the *boundary error*. This often happens at the two boundaries of the input spectrogram, i.e., the starting and ending points in time. The boundary error can have a high incidence rate when the target changes its activity category at the boundaries [as illustrated in Fig. 11(b)]. Based on the conclusions from Section IV-C, it is known that both the backward and forward temporal correlations are beneficial for model prediction. However, there is not enough forward or backward information for the classifier to exploit at the boundaries of the spectrogram.

The last type of mistake is the *labeling error*. These are unavoidable human errors that can happen at any time due to mislabeling. As shown in Fig. 11(c), the participant performed two continuous activities, "Stationary" and "Walking," and this is already empirically noticeable by visual inspection in the spectrogram. However, the ground truth label indicates that the target conducted "Walking" only, though the classifier correctly recognized the "Stationary" action.

V. Further Exploration of Practical Aspects

This section provides further explorations of the system performance from three practical aspects in relation to realistic HAR problems.

A. System Scalability

One practical issue in dealing with realistic HAR tasks is the required alignment in the number of radar nodes between the development (e.g., controlled conditions in laboratory) and the deployment environment (e.g., private spaces in homes or offices). The alignment may be hard to achieve because the number of radars in the deployment location can be limited by conditions such as cost and space. Therefore, it is desirable that the designed HAR system has scalability so that it can handle a variable number of radar outputs.

To test this aspect, the proposed classifier was trained and tested with an unmatched, different number of radars. The

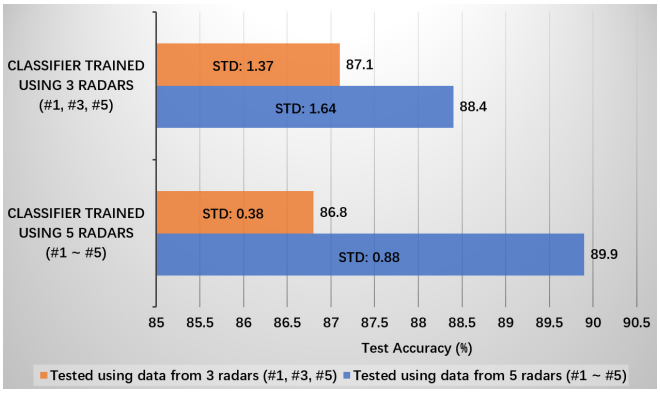

Fig. 12. Test accuracy of the proposed classifier for an unmatched number of radars in the training and testing data. Here, the classifier was trained and tested on a different number of radars. This aims to simulate the case when different radar nodes were deployed after the model training. The blue bars indicate that the classifier was tested using five radars, whereas the orange means only three radars were used for testing (radar index *#1*, *#3*, and *#5*).

results are shown in Fig. 12. In Fig. 12, two experiments were conducted. The first experiment measures the upscaling capability. The classifier is first trained using three radars (radar index #1, #3, and #5). After model training, the classification accuracy is measured using the testing data collected from all five radars. Based on the results, although the classifier is trained with only three radars, it is able to capture relevant information from five radars and improve the test accuracy.

On the contrary, the second experiment inspects the downscaling capability of the classifier. Unlike before, the classifier is trained using all five radars but tested with three radars (radar index #1, #3, and #5). The result shows that although the test accuracy is reduced from 89.9% to 86.8% by using fewer radars, the accuracy is similar (86.8% versus 87.1%) and the std is smaller (0.38 versus 1.37) compared with the case when the classifier is trained and tested with three radars.

The reason behind the success of the up- and downscalability of the proposed classifier is the use of a hybrid CNN-RNN architecture and the *halfway fusion* method. In the CNN-RNN architecture, CNNs are specialized in capturing various motion patterns in the input data. After model training, the CNN block can be used as a feature extractor and applied to generalize middle-level convolutional features from a variable number of inputs. Moreover, the *halfway fusion* block itself does not contain any trainable parameters but simply combines feature maps from different convolution channels.

Finally, it is important to note that the experiments conducted in this section are simplified and idealized. In real-world scenarios, for example, radars might be placed in very different geometrical settings. However, it suffices to prove the capability of the proposed classifier in terms of up/down scalability, which has been rarely discussed in previous works.

B. Impact on Classification of Imperfect Tracking

Although many works did not consider tracking tasks while investigating the HAR problem, being able to detect and track the target in the indoor environment is a prerequisite for realistic activity recognition, especially for multisubject activity recognition tasks. Human tracking using UWB radar sensors [1] is a challenging assignment that is often affected by ubiquitous false alarms and a high misdetection rate. Thus, it is important to consider the impact of imperfect tracking on the proposed HAR system.

TABLE III

NUMBER OF ZERO COLUMN VECTORS GENERATED FOR EACH RADAR CHANNEL ACCORDING TO A GIVEN MISSED DETECTION RATE (P_{MD}). THE TOTAL NUMBER OF COLLECTED DOPPLER VECTORS IN THE TESTING DATASET FOR EACH RADAR CHANNEL IS 6990

P_{MD}	Rad. #1	Rad. #2	Rad. #3	Rad. #4	Rad. #5
1E-3	10	6	9	7	6
5E-3	39	27	40	29	30
1E-2	76	59	77	61	63
5E-2	362	325	366	330	334
1E-1	717	664	722	672	677
2E-1	1424	1348	1431	1360	1367
3E-1	2128	2036	2137	2050	2060

Fig. 13 provides an example of the extracted spectrograms from the proposed RSN based on the tracking results generated by a decentralized tracker [3]. This example shows that the direct implications of an imperfect filtering process within the tracking algorithm are global and local misdetections. Worse than the local misdetections at an individual radar node, global misdetections imply that all radar sensors fail to track the target in the scene. As a consequence of these problems, the extracted spectrograms can be incomplete and filled with zero-column vectors at certain time bins.

In this article, a test is performed to quantitatively analyze the influence of the zero columns in the spectrograms on the system performance. In the planned test, zero column vectors are randomly added to the testing dataset according to a Bernoulli process, parameterized by the missed detection rate (P_{MD}). This would model the effect of imperfect tracking. Table III shows the number of zero vectors generated for each radar node under different P_{MD} values. To empirically visualize how the zero vectors influence the quality of the input data, Fig. 14 provides the comparison between the input spectrograms when P_{MD} equals 0 and 2E-1. Finally, it is noteworthy that the zero vectors are only added to the testing data, with the assumption that the training data were collected in a more controlled scenario where imperfect tracking is not present, unlike in the testing scenarios. It is possible that mixing zero vectors also in the training data may add extra regularization effects to the network learning and improve the system's robustness, but this is left for future work.

Fig. 15 shows the classification accuracy of the proposed classifier measured for different P_{MD}. Furthermore, two depth reduction strategies are compared, the average pooling (*AvgPool*) and *MaxPool* methods, within the *halfway fusion* framework. Based on the results, the proposed classifier shows a robust performance against misdetections as it can maintain a test accuracy of approximately 83.6% even when P_{MD} is set to 1E-1 (i.e., nearly one in ten-time bins of the spectrograms are zeroed). Additional details on the problem of multitarget tracking in indoor human activity classification are also reported in [72].

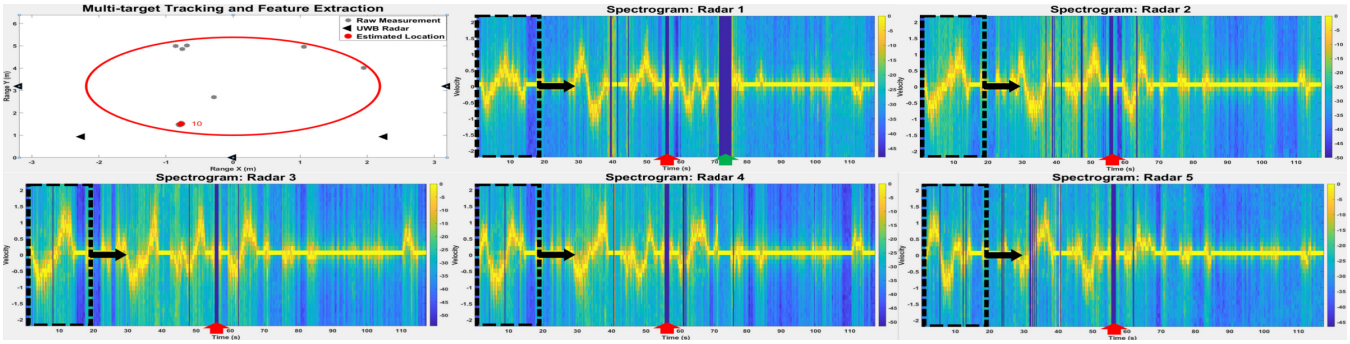

Fig. 13. Example of the generated spectrograms using the proposed RSN based on the tracking results. The black rectangles which shift along time are equivalent to the inputs to the proposed classifier. The red arrows pinpoint the time when all radars lost track of the target (global misdetection), whereas the green arrow denotes the moment when the target is partially missed by the RSN (local misdetection).

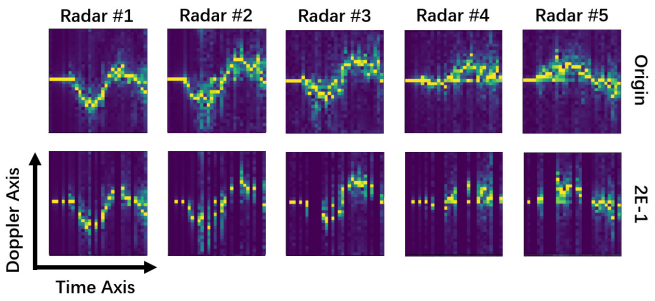

Fig. 14. Visual illustration of the zero column vectors added to the input spectrograms in case of missed detections. The original input data in the top row ($P_{\mathrm{MD}} = 0$) is compared with the case when P_{MD} is set to 2E-1.

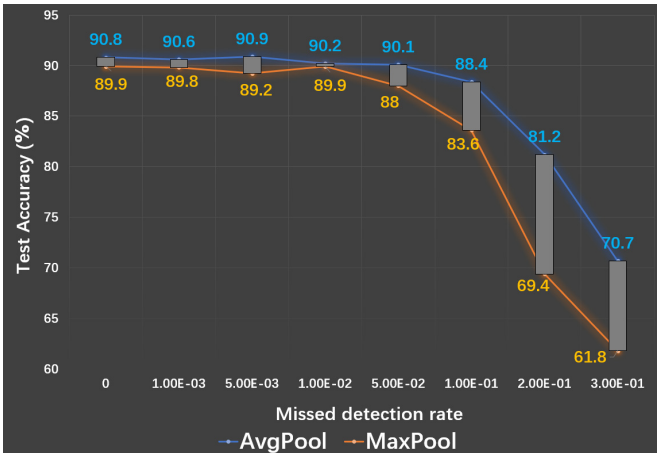

Fig. 15. Classification accuracy of the proposed classifier evaluated under different missed detection rates (ranging from 0 to 3E-1). In addition, two feature combining methods, *AvgPool* and *MaxPool*, for fusing the feature maps generated by the convolutional feature extractor are compared.

Finally, a further improvement in classification accuracy and system robustness is observed when an *AvgPool* layer is applied to compress the convoluted feature maps. Unlike the *MaxPool* layer, which only selects the most prominent middle-level features and discards the others [73], the *AvgPool* layer takes the average value of the features and retains more robustness in case of missing information happening at different nodes because of the imperfect tracking. As shown in the results, the proposed classifier with the *AvgPool*-based

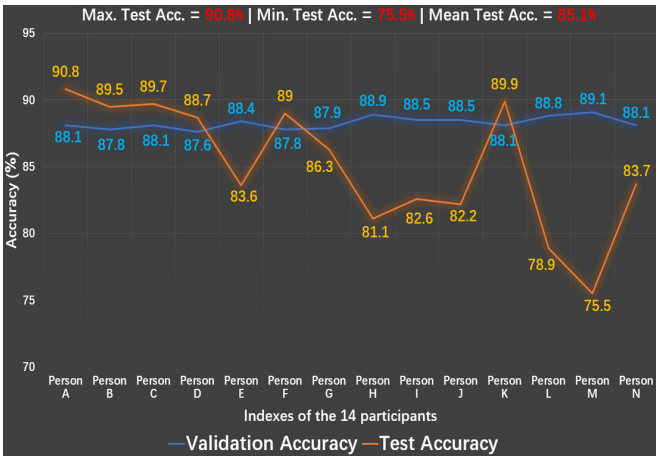

Fig. 16. Generalization capability of the proposed classifier measured over 14 participants. The validation and test accuracy are calculated using the fivefold CV method and the L1PO method, respectively.

halfway fusion method can withstand a P_{MD} of 2E-1 and achieve approximately 81.2% test accuracy.

C. Generalization Capability for Unknown Participants

Also, very important is the generalization capability when evaluating an HAR system. This is because radar-based applications often have a limited dataset, and an NN-based classifier can easily overfit the training data. Therefore, it is necessary to know how the proposed classifier performs on the dataset sampled from an unknown participant never seen before.

Fig. 16 provides the evaluation result of the proposed classifier. To measure its generalization capability, the L1PO method is implemented across all 14 participants. More explicitly, each person has been selected as the testing data once, while the remaining data are used for model training and validation. With the fivefold CV method applied, for every test instance, the model is trained, validated, and tested five times, respectively, before averaging the results for each participant.

The result shows that, although the test accuracy fluctuates across participants, the proposed classifier is able to generalize well using only a very small part of all possible motion patterns. The classifier scores an average test accuracy of 85.1% on the data from participants it has never seen before. However, one might notice that the classifier is less accurate

TABLE IV
PERFORMANCE COMPARISON BETWEEN REPRESENTATIVE METHODS IN THE LITERATURE AND THE PROPOSED APPROACH

Works	Sensors	Size	Fusion	Features	Motion Diversity	Scalability	Mis-detection	Classification Performance
SVM ([18])	CW Doppler radar	/	No	/	Separated and constrained human motions	Unexplored	Unexplored	92.8% accuracy measured using the hold-out method with handcrafted features for 7-class HAR task
CNN ([74])	SFCW radar	About 135M	No sensor fusion	Spatial features	Separated and constrained human motions	Unexplored	Unexplored	94.9% accuracy reported using the hold-out method for 8-class HAR task
RNN ([42])	One FMCW + three UWB	Not given	Yes	Temporal features	Semi-continuous and semi-constrained human motions	Unexplored	Unexplored	92.7% average accuracy measured using the L1PO method with handcrafted features for 7-class HAR task
CNN-RNN ([75])	Planar FMCW radar	About 10M	No sensor fusion	Spatio-temporal features	Separated and constrained human motions	Unexplored	Unexplored	97.7% accuracy measured using 5-fold CV with 0.5 sec aggregation time for 7-class HAR task
CNN-RNN ([76])	Linear FMCW radar	Not given	No sensor fusion	Spatio-temporal features	Separated and semi-constrained human motions	Unexplored	Unexplored	92.0% accuracy reported using the hold-out method for 3-class HAR task
CNN-RNN ([77])	CW Doppler radar	205K	No sensor fusion	Spatio-temporal features	Separated and constrained human motions	Unexplored	Unexplored	98.3% accuracy measured by 5-fold CV for 7-class HAR and 76.2% accuracy achieved by this method on our dataset with L1PO method
CNN-RNN (**ours**)	Five UWB radars	71K	Yes (three types)	Spatio-temporal features	Continuous and unconstrained human motions	Up/down Scalable	Only 2.6% accuracy loss at 1E-1	90.8% accuracy measured using the L1PO and 5-fold CV method for 9-class HAR task

in identifying the activities of Person L and M. One hypothetical explanation for this performance drop is that there was a considerable time gap between the data collection for Person L and M and the remaining participants. During such gap time, the layout of the laboratory furniture had to be partially modified due to construction work, and the resulting changes in multipath and clutter are hypothesized to affect the performance of HAR systems.

In general, these results also reflect the importance of the generalization test. Without it, the evaluation of the performance can be severely biased, as the test accuracy ranges from a maximum of 90.8% to a minimum of 75.5%.

VI. CONCLUSION

In this work, a novel hybrid CNN-RNN architecture for HAR is designed for fusing data from an RSN. As summarized in Table IV, the proposed network is validated by a continuous and unconstrained stream of human motion collected using a network of five distributed monostatic UWB radar nodes. This more realistic human motion scenario differs from previous studies that have tended to use artificially separated activities, or at best, continuous activities just from a single-monostatic radar. Notably, during the data collection, the 14 participants were able to have arbitrary moving trajectories within the measurement area, while performing nine ADLs with seamless interactivity transitions.

Also, instead of manually generating handcrafted features, the proposed CNN-RNN classifier performs automatic feature extraction across spatially distributed radar nodes. The stacked 2D-CNNs perform a hierarchical feature decomposition on the input data, whereas the bidirectional GRUs exploit the past and future temporal dependencies within the human motion. Furthermore, three NN-based data fusion methods were explored and compared to utilize the rich information provided by the different nodes of the RSN. Due to the lightweight CNN-RNN architecture and WS technique, the proposed network is small in scale with only 71 K trainable parameters, which reduces the need to collect large radar datasets that are often expensive.

Besides, considering more realistic human activities, three practical aspects for the implementation of the proposed classifier in HAR problems were considered, namely, the system scalability across different numbers of nodes, the impact of imperfect tracking, and the generalization capability for unknown participants. The results showed that the proposed classifier has up/down scalability to handle a variable number of radar inputs after model training. Also, it shows a robust performance against missed detections during the tracking process prior to spectrogram generation, as well as acceptable performances for unseen participants.

For a fair and comprehensive performance comparison in terms of classification accuracy, as shown in the last column of Table IV, one challenge is the broad diversity of approaches proposed in the literature. These tend to consider different sensor types and numbers, a different number of classes and data representations, and different evaluation methods. This makes it difficult to compare on the same benchmarking dataset, also because the source code of some models is not shared. To attempt a possible comparison, the method proposed in [77] for separated (noncontinuous) activities has been reimplemented to handle our dataset. This method and our proposed method have been validated rigorously using a combination of L1P0 and K-fold CV method: the former decides the test data and the latter splits the rest into training and validation. The result shows that fivefold-averaged classification accuracy of about 90.8% is achieved for nine-class HAR by the proposed classifier, whereas the method adapted from [77] yields lower results.

Future work aims to further improve the classifier's performance based on the current results. For example, promising future directions can be investigating the influence of negative classes, trying to address the transition error and boundary error as discussed in Section IV-F, or replacing the current CNN-RNN architecture with deep transformers operating on audio, images, or on point clouds [78]–[80]. Negative classes can be, for example, movements that are not related to the activities of interest to be classified. For the transition error, it is possible to remove the requirement of a precise activity-to-label alignment by using the connectionist temporal classification (CTC) [81] loss. Preliminary result shows that adding the CTC loss to the proposed NN can significantly improve the F1 score of "Falling-from-walking" from 0.71 to 0.92.

Apart from that, different radar data representations for the network input can be investigated. In this work, spectrograms are used to capture the various velocity patterns of different body parts in human movements. However, recent works [17], [82], [83] have shown that combining multidomain information for HAR can also be advantageous, e.g., combining the RD, DT, CVD, and RT information.

ACKNOWLEDGMENT

The authors are grateful to all the volunteers who helped with the data collection.

REFERENCES

[1] M. Chiani, A. Giorgetti, and E. Paolini, "Sensor radar for object tracking," *Proc. IEEE*, vol. 106, no. 6, pp. 1022–1041, Jun. 2018.

[2] T. Sakamoto, T. Sato, P. J. Aubry, and A. G. Yarovoy, "Texture-based automatic separation of echoes from distributed moving targets in UWB radar signals," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 352–361, Jan. 2015.

[3] Y. He, P. Aubry, F. L. Chevalier, and A. Yarovoy, "Decentralised tracking for human target in multistatic ultra-wideband radar," *IET Radar, Sonar Navigat.*, vol. 8, no. 9, pp. 1215–1223, 2014.

[4] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.

[5] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, Jan. 2019.

[6] X. Shi, Y. Li, F. Zhou, and L. Liu, "Human activity recognition based on deep learning method," in *Proc. Int. Conf. Radar (RADAR)*, Aug. 2018, pp. 1–5.

[7] S. A. Shah and F. Fioranelli, "RF sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, no. 11, pp. 26–44, Nov. 2019.

[8] T. Sakamoto, P. J. Aubry, S. Okumura, H. Taki, T. Sato, and A. G. Yarovoy, "Noncontact measurement of the instantaneous heart rate in a multi-person scenario using *X*-band array radar and adaptive array processing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 2, pp. 280–293, Jun. 2018.

[9] F. Fioranelli, J. Le Kernec, and S. A. Shah, "Radar for health care: Recognizing human activities and monitoring vital signs," *IEEE Potentials*, vol. 38, no. 4, pp. 16–23, Jul. 2019.

[10] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. C. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[11] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[12] F. Fioranelli, M. Ritchie, and H. Griffiths, "Classification of unarmed/armed personnel using the NetRAD multistatic radar for micro-Doppler and singular value decomposition features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1933–1937, Sep. 2015.

[13] B. Cağliyan and S. Z. Gürbüz, "Micro-Doppler-based human activity classification using the mote-scale BumbleBee radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2135–2139, Oct. 2015.

[14] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2017, pp. 1519–1523.

[15] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.

[16] S. Björklund, H. Petersson, and G. Hendeby, "Features for micro-Doppler based activity classification," *IET Radar Sonar Navigat.*, vol. 9, no. 9, pp. 1181–1187, 2015.

[17] Z. Li *et al.*, "Multi-domains based human activity classification in radar," in *Proc. IET Int. Radar Conf. (IET IRC)*, 2021, pp. 1744–1749.

[18] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, May 2009.

[19] N. Cristianini *et al.*, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[20] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, Jan. 2016.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Nov. 2015.

[22] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *Handbook Brain Theory Neural Netw.*, vol. 3361, no. 10, p. 1995, Apr. 2015.

[23] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.

[24] H. Jiang, F. Fioranelli, S. Yang, O. Romain, and J. Le Kernec, "Human activity classification using radar signal and RNN networks," *IET Radar, Sonar Navigat.*, vol. 2021, pp. 1595–1599, Jul. 2021.

[25] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4905–4913.

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

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-Doppler spectrograms," in *Proc. IEEE Int. Conf. Comput. Electromagn. (ICCEM)*, Mar. 2018, pp. 1–3.

[29] Y. Yang, C. Hou, Y. Lang, T. Sakamoto, Y. He, and W. Xiang, "Omnidirectional motion classification with monostatic radar system using micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3574–3587, Nov. 2019.

[30] Y. Kim and T. Moon, "Classification of human activity on water through micro-Dopplers using deep convolutional neural networks," *Proc. SPIE*, vol. 9829, May 2016, Art. no. 982917.

[31] Y. He, Y. Yang, Y. Lang, D. Huang, X. Jing, and C. Hou, "Deep learning based human activity classification in radar micro-Doppler image," in *Proc. 15th Eur. Radar Conf. (EuRAD)*, Sep. 2018, pp. 230–233.

[32] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.

[33] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digit. Signal Process.*, vol. 87, pp. 125–131, Apr. 2019.

[34] X.-Y. Zhang, G.-S. Xie, C.-L. Liu, and Y. Bengio, "End-to-end online writer identification with recurrent neural network," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 2, pp. 285–292, Apr. 2017.

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

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[38] D. Tahmoush and J. Silvious, "Radar micro-Doppler for long range front-view gait recognition," in *Proc. 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–6.
[39] B. Vandersmissen *et al.*, "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.
[40] F. Fioranelli, M. Ritchie, and H. Griffiths, "Aspect angle dependence and multistatic data fusion for micro-Doppler classification of armed/unarmed personnel," *IET Radar, Sonar Navigat.*, vol. 9, no. 9, pp. 1231–1239, Dec. 2015.
[41] F. Fioranelli, M. Ritchie, and H. Griffiths, "Multistatic human micro-Doppler classification of armed/unarmed personnel," *IET Radar, Sonar Navigat.*, vol. 9, no. 7, pp. 857–865, Aug. 2015.
[42] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7590–7603, Mar. 2021.
[43] B. Zhou *et al.*, "Simulation framework for activity recognition and benchmarking in different radar geometries," *IET Radar, Sonar Navigat.*, vol. 15, no. 4, pp. 390–401, Apr. 2021.
[44] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Continuous human activity recognition for arbitrary directions with distributed radars," in *Proc. IEEE Radar Conf.*, May 2021, pp. 1–6.
[45] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[46] R. G. Guendel, M. Unterhorst, F. Fioranelli, and A. Yarovoy. (Nov. 2021). *Dataset of Continuous Human Activities Performed in Arbitrary Directions Collected With a Distributed Radar Network of Five Nodes*. [Online]. Available: https://data.4tu.nl/articles/dataset/Dataset_of_continuous_human_activities_performed_in_arbitrary_directions_collected_with_a_distributed_radar_network_of_five_nodes/16691500
[47] D. Benedetto, *Understanding Ultra Wide Band Radio Fundamentals*. London, U.K.: Pearson, 2008.
[48] D. Wang, S. Yoo, and S. H. Cho, "Experimental comparison of IR-UWB radar and FMCW radar for vital signs," *Sensors*, vol. 20, no. 22, p. 6695, Nov. 2020.
[49] Y. He, "Human target tracking in multistatic ultra-wideband radar," Ph.D. dissertation, Dept. Microelectron., Delft Univ. Technol., Delft, The Netherlands, Apr. 2014.
[50] L. Ren and A. E. Fathy, "Noncontact heartbeat detection using UWB impulse Doppler radar," in *Proc. USNC-URSI Radio Sci. Meeting*, Jul. 2015, pp. 1–3.
[51] T. Wagner, R. Feger, and A. Stelzer, "Radar signal processing for jointly estimating tracks and micro-Doppler signatures," *IEEE Access*, vol. 5, pp. 1220–1238, 2017.
[52] W. Song and J. Cai, "End-to-end deep neural network for automatic speech recognition," Standford, CA, USA, Tech. Rep. CS224D, 2015.
[53] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[54] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 539–546.
[55] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[57] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2488–2498.
[58] Y. Wang, L. Zhang, B. Zhang, and Z. Li, "End-to-end Mandarin recognition based on convolution input," in *Proc. MATEC Web Conf.*, vol. 214. Les Ulis, France: EDP Sciences, 2018, p. 01004.
[59] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*.
[60] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 1–8.
[61] V. Passricha and R. K. Aggarwal, "A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1261–1274, Mar. 2019.
[62] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," 2017, *arXiv:1701.02720*.
[63] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[64] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[65] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. design Implement. (OSDI)*, 2016, pp. 265–283.
[66] P. Vingelmann and F. H. Fitzek. (2020). *Cuda, Release: 10.2.89*. NVIDIA. [Online]. Available: https://developer.nvidia.com/cuda-toolkit
[67] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[70] W. Wang, X. Yang, and H. Yang, "End-to-end low-resource speech recognition with a deep CNN-LSTM encoder," in *Proc. IEEE 3rd Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Sep. 2020, pp. 158–162.
[71] D. Wang, X. Wang, and S. Lv, "End-to-end Mandarin speech recognition combining CNN and BLSTM," *Symmetry*, vol. 11, no. 5, p. 644, May 2019.
[72] S. Zhu, "Multiple target tracking and human activity recognition based on the IR-UWB radar sensor networks," M.S. thesis, Dept. Microelectron., Delft Univ. Technol., Delft, The Netherlands, Oct. 2021.
[73] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Personnel recognition and gait classification based on multistatic micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 669–673, May 2018.
[74] L. Tang, Y. Jia, Y. Qian, S. Yi, and P. Yuan, "Human activity recognition based on mixed CNN with radar multi-spectrogram," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25950–25962, Nov. 2021.
[75] Y. Kim, I. Alnujaim, and D. Oh, "Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13522–13529, Jun. 2021.
[76] H.-U.-R. Khalid, A. Gorji, A. Bourdoux, S. Pollin, and H. Sahli, "Multi-view CNN-LSTM architecture for radar-based human activity recognition," *IEEE Access*, vol. 10, pp. 24509–24519, 2022.
[77] 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. 24713–24720, 2020.
[78] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[79] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[80] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021.
[81] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.
[82] S. Li, M. Jia, J. L. Kernec, S. Yang, F. Fioranelli, and O. Romain, "Elderly care: Using deep learning for multi-domain activity classification," in *Proc. Int. Conf. UK-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–4.
[83] M. Jia, S. Li, J. L. Kernec, S. Yang, F. Fioranelli, and O. Romain, "Human activity classification with radar signal processing and machine learning," in *Proc. Int. Conf. UK-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–5.

Simin Zhu received the B.Sc. degree in electrical engineering and automation from Central South University, Changsha, China, in 2016, and the master's degree in radar signal processing and machine learning from the Microwave Sensing, Signals and Systems Group, Delft University of Technology, Delft, The Netherlands, in 2021, where he is currently pursuing the Ph.D. degree.

He was a Hardware Engineer with Huawei Technology Company Ltd., Shenzhen, China, for 1.5 years.

Ronny Gerhard Guendel (Graduate Student Member, IEEE) received the Dipl.-Ing. (FH) degree from the West Saxon University of Applied Sciences of Zwickau, Zwickau, Germany, in 2017, and the M.Sc. degree in signal processing and communications from the Center for Advanced Communications, Villanova University, Villanova, PA, USA, in 2019. He is currently pursuing the Ph.D. degree with Delft University of Technology, Delft, The Netherlands.

He worked with the Fraunhofer Institute for Machine Tools and Forming Technology, Chemnitz, Germany. From 2017 to 2018, he joined Villanova University as a Fulbright Scholar in electrical engineering. In 2018, he worked on vehicular wireless communications with the TU Dresden Vodafone Chair Mobile Communications Systems, Dresden, Germany, guided by Prof. Gerhard Fettweis. From 2018 to 2019, he was a Research Assistant with the Center for Advanced Communications, Villanova University, under Dr. Moeness G. Amin, working on continuous activity classification. He works on monitoring continuous human activities in range and time beyond micro-Doppler by using RF-sensing technology.

Alexander Yarovoy (Fellow, IEEE) received the Diploma degree (Hons.) in radiophysics and electronics and the Candidate and Ph.D. degrees in physics and mathematical science (radiophysics) from Kharkiv State University, Kharkiv, Ukraine, in 1984, 1987, and 1994, respectively.

In 1987, he joined as a Researcher with the Department of Radiophysics, Kharkiv State University, where he became a Full Professor in 1997. From 1994 and 1996, he was with the Technical University of Ilmenau, Ilmenau, Germany, as a Visiting Researcher. Since 1999, he has been with Delft University of Technology, Delft, The Netherlands, where he has been leading a Chair of Microwave Sensing, Systems and Signals since 2009. He has authored and coauthored more than 450 scientific or technical papers, six patents, and 14 book chapters. His main research interests include high-resolution radar, microwave imaging, and applied electromagnetics, in particular, ultrawideband antennas.

Dr. Yarovoy served as the Director of the European Microwave Association from 2008 to 2017. He was a recipient of the European Microwave Week Radar Award for the paper that best advances in state-of-the-art radar technology (together with L. P. Ligthart and P. van Genderen) in 2001 and (together with T. Savelyev) in 2012. In 2010, he received the Best Paper Award from the Applied Computational Electromagnetic Society together with D. Caratelli. He served as the General TPC Chair for the 2020 European Microwave Week, the Chair and the TPC Chair for the 5th European Radar Conference, and the Secretary for the 1st European Radar Conference. He also served as the Co-Chair and the TPC Chair for the Xth International Conference on GPR (GPR2004). He served as an Associate Editor for the *International Journal of Microwave and Wireless Technologies* from 2011 till 2018 and a Guest Editor for five special issues of the IEEE TRANSACTIONS and other journals.

Francesco Fioranelli (Senior Member, IEEE) received the Ph.D. degree from Durham University, Durham, U.K., in 2014.

He is currently a tenured Assistant Professor with Delft University of Technology, Delft, The Netherlands. He was a Research Associate with the University College London, London, U.K., from 2014 to 2016, and an Assistant Professor with the University of Glasgow, Glasgow, U.K., from 2016 to 2019. His research interests include the development of radar systems and automatic classification for human signatures analysis in healthcare and security, drones and unmanned aerial vehicles (UAVs) detection and classification, automotive radar, wind farm, and sea clutter.