The Human Activity Radar Challenge

Benchmarking based on the 'Radar signatures of human activities' dataset from Glasgow University

Yang, Shufan; Le Kernec, Julien; Romain, Olivier; Fioranelli, Francesco; Cadart, Pierre; Fix, Jeremy; Ren, Chengfang; Manfredi, Giovanni; Letertre, Thierry; More Authors

DOI
10.1109/JBHI.2023.3240895

Publication date
2023

Document Version
Final published version

Published in
IEEE Journal of Biomedical and Health Informatics

The *Human Activity Radar Challenge:* Benchmarking Based on the 'Radar Signatures of Human Activities' Dataset From Glasgow University

Shufan Yang, Julien Le Kernec, *Graduate Student Member, IEEE*, Olivier Romain, Francesco Fioranelli, Pierre Cadart, Jérémy Fix, Chenfang Ren, *Member, IEEE*, Giovanni Manfredi, Thierry Letertre, Israel David Hinostroza Sáenz, Jifa Zhang, Huaiyuan Liang, Xiangrong Wang, *Senior Member, IEEE*, Gang Li, Zhaoxi Chen, Kang Liu, Xiaolong Chen, *Senior Member, IEEE*, Jiefang Li, Xing Wu, *Member, IEEE*, Yichang Chen, and Tian Jin, *Member, IEEE*

***Abstract*—Radar is an extremely valuable sensing technology for detecting moving targets and measuring their range, velocity, and angular positions. When people are monitored at home, radar is more likely to be accepted by end-users, as they already use WiFi, is perceived as privacy-preserving compared to cameras, and does not require user compliance as wearable sensors do. Furthermore, it is not affected by lighting conditions nor requires artificial lights that could cause discomfort in the home environment. So, radar-based human activities classification in the context of assisted living can empower an aging society to live at home independently longer. However, challenges remain as to the formulation of the most effective algorithms for radar-based human activities classification and their validation. To promote the exploration and cross-evaluation of different algorithms, our dataset released in 2019 was used to benchmark various classification approaches. The challenge was open from February 2020 to December 2020. A total of 23 organizations worldwide, forming 12 teams from academia and industry, participated in the inaugural Radar Challenge, and submitted 188 valid entries to the challenge. This paper presents an overview and evaluation of the approaches used for all primary contributions in this inaugural challenge. The proposed algorithms are summarized, and the main parameters affecting their performances are analyzed.**

***Index Terms*—Human activity classification, radar, machine learning, convolutional neural networks.**

Manuscript received 6 May 2022; revised 29 November 2022; accepted 24 January 2023. Date of publication 30 January 2023; date of current version 5 April 2023. This work was supported in part by the British Council under Grants 515095884 and 514739586, in part by Campus France 44764WK-PHC Alliance France-UK, in part by the UK EPSRC under Grant INSHEP EP/R041679/1, and in part by the Dutch Research Council NWO (grant RAD-ART). This challenge was supported by the IET Electromagnetics Network. *(Corresponding author: Julien Le Kernec.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/JBHI.2023.3240895

I. Introduction

WHY is radar considered a viable technology for ambient assisted living and monitoring of human activities and well-being? Radar technology is rising at the forefront in fields such as commerce, defense, and security. Indeed, the advancement allowed by the integration of radar in the automotive industry has driven radar technology in more and more civilian applications (medical, human-machine interactions, food security, smart environments, assisted living) as it now has a small form factor, is low-power, works with solid-state integrated chips, and is reconfigurable using software-defined architectures. Due to the common deployment of wifi in home environments, schools, hospitals and other public infrastructures, we know that electromagnetic waves are very efficient indoors and penetrate through walls and non-metal obstacles which alleviate obscuring problems encountered in vision/optical-based technologies. Radar is also a well-established technology and with low-powers of the order of Wi-Fi emissions, it is a safe, affordable and dependable technology. Radar can be seen as a sensor in a suite of sensors (video, infrared, acoustic, pressure, and wearable sensors) in the context of human activity recognition. Radar is a contactless technology that can operate in any lighting condition and as well as haze. The paradigm shift of "aging in place" is being pushed by governments in developed countries ad the number of people over the age of 65 will be greater than 1 billion by 2030. This shift is necessary to provide elderly people and their families a sense of security and a way to monitor their health while remaining in a comfortable and familiar setting. Current assistive technologies such as wearables allow human activity recognition and the detection of critical events such as falls but they are battery operated and may not be suitable for cognitively impaired people as they would need to remember to wear them and recharge them to remain operational. Radar however does not rely on the subject's compliance and is a non-obstructive contactless human activity recognition technology that can alert

emergency services to critical events for a timely response to ensure the health and welfare of the subjects under observation. This technology can also be further expanded to create smart environments where the elderly interacts through gestures with the environments and control appliances.

Compared with wearable devices, video, or ambient sensor technologies, radar does not record plain images of the subjects or environments and does not require the users to wear, carry, or interact with additional devices. Radar can provide rich information and is perceived as privacy-preserving [1], while avoiding potential issues of users' acceptance and compliance. Compared to passive sensing using Wi-Fi emissions or their channel state information (CSI), modern radar technology provides an additional degree of freedom in choosing the waveform and beam-forming used in assisted living applications. For all these reasons, radar has gained popularity as a sensing technology to support innovative/personalized healthcare services, including monitoring of daily activity patterns, analysis of gait parameters and imbalances, and prompt detection of critical events such as falls [2], [3], [4], [5], [6], [7], [8].

Indoor human activities can be mainly grouped into three types based on the extent of movements: action-based, motion-based and interaction-based. Action-based activities only require a unidirectional action such as walking and drinking water. Motion-based activities require connected activities with multi-types of movements. Interaction-based activities require two or more individuals or objects involved in the activities. In this challenge, we set up classification tasks for action-based activities commonly performed indoors.

Automatic human activities classification with radar has attracted considerable attention over the past few years, resulting in dozens of different algorithms for this purpose. Many algorithms exploited features that were specifically designed for the activities to be classified [9], [10], [11]. For example, physical parameters are directly estimated from the micro-Doppler signatures, such as (but not limited to) the center of mass and bandwidth, or by means of transformations such as Singular Value Decomposition (SVD) and Discrete Cosine Transform (DCT), amongst others. Similar techniques were also applied to the study of human gait [12], [13] with the objective of detecting abnormal gait (e.g., limping, imbalances) or extracting parameters such as average speed and stride length that have important medical significance [14]. Furthermore, time-frequency distributions such as Short Time Fourier Transform (STFT) or Wavelet techniques were used to characterize the micro-Doppler signatures of the gaits, from which features and valuable information were then extracted. Deep learning techniques have also been proposed, whereby the feature extraction step is performed by neural networks that take the spectrograms directly as input or other radar data representations such as range-time plots or 3D radar cubes [4], [7], [15]. These networks can be convolutional neural networks and convolutional auto-encoders [16], [17], which interpret the radar data as images, or recurrent networks such as LSTM (Long Short-Term Memory) [18], [19], which interpret the radar data as a temporal sequence of values, or architectures based on their combinations.

However, many of these algorithms were typically implemented and evaluated on relatively small, proprietary datasets. These datasets are largely different from each other with respect to activity types, input data format, radar characteristics, the geometry of the data collection, and classification algorithms, specifically both the extracted features and the selected classification algorithms. Therefore, it has been difficult to compare the performance of the proposed radar-based classification methods on a common benchmark, unlike in other research fields such as image or audio processing, where common datasets are routinely used as the benchmarks for proposed algorithms.

To address the lack of a shared dataset, the University of Glasgow (UoG) released 'the Radar signatures of human activities' dataset [20], [21] in 2019 with the descriptors in the read me file and sample code to process the database (UoG dataset for further references). In 2020, the Radar Challenge "Human Activity Classification with Radar" was organized at the IET International Radar Conference 2020. Because of the COVID-19 global pandemic, the event was postponed to 2021 before being canceled altogether. Despite those disruptions, the challenge carried on as an online event and attracted contributions from several international institutions and researchers.

In this paper, we aim to present the results of this challenge benchmarked on a common dataset. This is not a review paper but an analysis of the submissions for the international radar challenge 2020 and put them in the context of the state of the art. We will review the main methods/models and findings from the works submitted to this challenge. Statistical information on the most successful methods is also provided, and those models are explained in more detail. Other sensing modalities can be used for activity monitoring, and other classification algorithms and neural networks could be employed too. However, the scope of this paper is not to provide a comprehensive review of all possible methods but to discuss in one place all the different proposed algorithms for the UoG dataset [20], [21] at the challenge event. Furthermore, while the main focus of this paper is on the analysis of training methods and classification models, we also present other aspects such as pre-/post-processing steps, training hyperparameters, data augmentation, choice of loss function, and model ensemble methods.

II. The Dataset

This section provides a brief description of the dataset released for the radar challenge. The details of this dataset can be found in [20], [21]. The radar signatures were recorded using a COTS FMCW radar from Ancortek model 580-B. The centre frequency of the chirp was 5800 MHz and the signal bandwidth 400 MHz. The pulse repetition frequency was set to 1 kHz and the transmitted output power +18 dBm. The antennas at the transmitter and receiver were identical −17 dBi Yagi. The raw radar data is composed of de-chirped complex beat frequency samples for each recorded file, with 128 In-phase and Quadrature (I & Q) samples per sweep or chirp. These are the so-called 'raw' data, as no pre-processing has been applied yet to the data.

Similar to common cross-evaluation approaches for other data classification challenges, we proposed using these data for

TABLE I
SUMMARY OF THE DATASETS [21]

Dataset	No. of samples	Participants age
Training dataset	1754	21-98 yrs
Seen dataset for validation	100	60-98 yrs
Unseen dataset for validation	100	24-40 yrs

TABLE II
SUMMARY OF THE TRAINING/SEEN DATASET [21]

ID	Activity description	number of samples	data length
A1	Walking back and forth	312	10s
A2	Sitting down on a chair	312	5s
A3	Standing up from a chair	311	5s
A4	Picking up an object	311	5s
A5	Drinking water	310	5s
A6	Falling	198	5s

system training and two stages of evaluation: *seen dataset* and *unseen dataset*, as shown in Table I. The *seen dataset* is a subset of the training data set. The *unseen dataset* is a dataset that was collected from four additional subjects and that was not available to participants in the classification challenge at the stage of training machine learning algorithms. The motivation was that testing results from the *unseen dataset* would indicate how well the classifier or network could generalize the model as well as avoid overfitting. Furthermore, by comparing the prediction results from the evaluation of *seen data* vs. *unseen data*, we wanted to see where models could perform poorly in feature selection and classification for some of the activities.

As indicated in Table I, it is worth highlighting that the dataset contained data from 8 different indoor environments: UoG laboratory room; UoG common room; UoG MAST lab; Glasgow NG homes (3 rooms); Age UK West Cumbria (2 rooms). Hence, the radar signatures presented different background clutter from static furniture and walls, as well as different distances between the radar and the subjects. This is an important point, as in the majority of the open literature on radar-based human activities classification, data are generally recorded in just one environment, typically in controlled laboratory conditions.

Furthermore, the recordings involved 72 subjects of which 26 were female, which is substantial compared to the typical size of radar-based proprietary datasets in human activity classification literature. The age of these subjects ranged from 21 to 98 years, therefore offering the possibility of analyzing differences in gait and activity signatures compared with datasets that only have younger subjects [4], [7], [15]. While 1754 samples is still rather small compared to datasets for other sensing technologies, to the best of our knowledge the UoG dataset is a good addition to the research landscape of the radar community in this application area. The number of samples per activity are detailed in Table II. The first step of the challenge consisted in classifying seen data from elderly volunteers from 60–98 years of age including only activities A1–A5 as the ederly did not perform the fall activity (A6) for safety and ethical reasons. There were 20 of each for the 100 sample "Seen data" for validation used in step 1 of the challenge. Step 2 in the challenge consisted in 100 samples taken from 108 recordings with 18 repetitions of each activities A1 to A6 (details in Table III).

TABLE III
SUMMARY OF THE UNSEEN DATASET [21]

ID	Activity description	number of samples	data length
A1	Walking back and forth	18[1]	15-30s
A2	Sitting down on a chair	18[1]	10s
A3	Standing up from a chair	18[1]	10s
A4	Picking up an object	18[1]	10s
A5	Drinking water	18[1]	10s
A6	Falling	18[1]	10s

[1]Volunteer 1 Female executed 3 times each activity, Volunteer 2 Male executed 6 times each activity, Volunteer 3 Male executed 3 times each activity, Volunteer 4 Male executed 6 times each activity

Fig. 1 shows an example of the spectrograms (time-velocity patterns) for a sequence of 6 activities. They are walking, standing, sitting down, picking up an object, drinking, and a simulated fall.

III. CLASSIFICATION CHALLENGE TASK DESCRIPTION

The majority of the current research in radar-based human activity recognition is mainly tested on data from a small cohort composed of graduate students or postdocs between the age of 20-30 in the laboratory or in controlled conditions. The data collected for this challenge was organised for a more general-purpose to validate human activities classification in more realistic environments, including clutter from common pieces of furniture in the surroundings. For the evaluation, which ran from 02/2020 to 12/2020, a total of 12 finalist teams made 188 valid submissions out of the many submitted by participants of the challenge who did not submit a paper to the conference. Note that the participants processed the data locally and submitted only the output of their predictions to the radar challenge website for scoring and analysis purposes. The leader board was used as a portal on the grand-challenge.ord website to collect the output of the predictions and rank the entries in order of performance against the 'seen' blind evaluation in phase 1 and the 'unseen' blind evaluation in phase 2.

Fig. 2 displays a heat map representing the number of participating teams per country, with all information in terms of country and affiliation of each team as reported by the participants. Fig. 3 shows the number of valid submissions to the leader board each week within the evaluation period of the challenge for the *seen data*. These submissions were made only for the 1st round of the *'seen'* data prior to reasing the *'unseen'* to the teams who submitted papers to the conference. The term submissions subsequently will refer to the submissions from the finalist teams who submitted a conference paper explaining their method and results [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38]. The *unseen data* evaluation results can be found at the following link https://humanactivityclassificationwithradar.grand-challenge.org/evaluation/challenge/leaderboard/.

Instead of using the F1 score, scoring was done using the following method to define the leader table ranking. The classification task considered the combined prediction accuracy. This included $mAP_{prediction}$, the prediction accuracy of the same subject for one of the six activities, and $mAP_{generalisation}$, which was the prediction accuracy of multiple actions of the same class

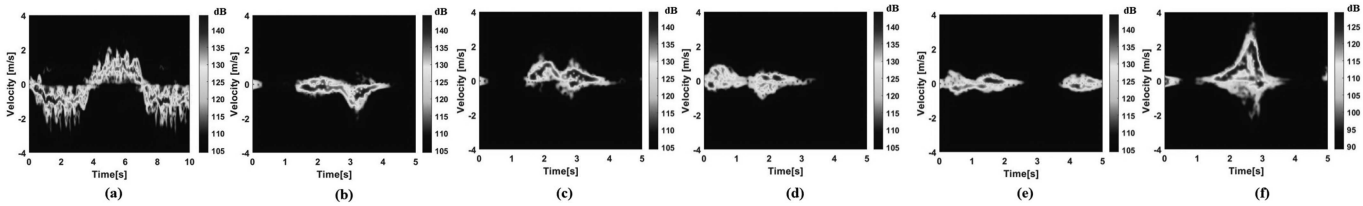

Fig. 1. Radar spectrograms of five activities [a) A1, b) A2, c) A3, d) A4, e) A5] performed by a 68-year-old male participant, and 1 activity performed by a 24-year old student [f) A6] for safety reasons the elderly are not performing the falling activity.

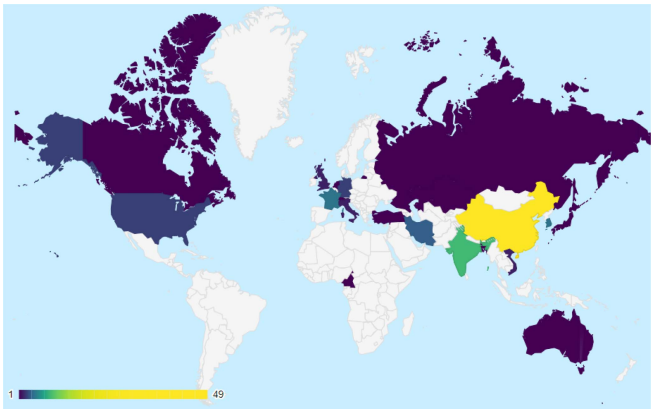

Fig. 2. Heat map of the world showing the number of participants per country.

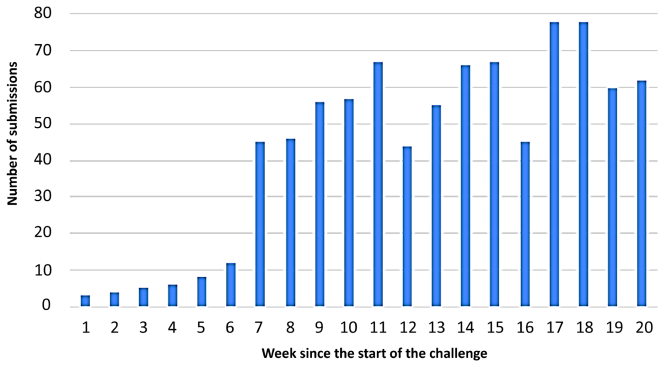

Fig. 3. Submission statistics (valid submissions per week) for the first human activity classification challenge with radar - the submissions include the finalist submissions and all the other submissions from people who did not submit to the conference.

conducted by various subjects, as shown in equation 1 adapted from [39].

$$S_{leadertable} = \alpha \cdot mAP_{prediction} + \beta \cdot mAP_{generalisation} \tag{1}$$

$$mAP_{prediction} = \frac{1}{|pc|}\Sigma_{c\in pc}\left(\frac{TP(c)}{(TP(c)+FP(c))}\right) \tag{2}$$

$$mAP_{generalisation} = \frac{1}{|gc|}\Sigma_{c\in gc}\left(\frac{TP(c)}{(TP(c)+FP(c))}\right) \tag{3}$$

α and β are arbitrary parameters to give more or less importance to the two *mAP* components in the calculation of the final score. In this case, to emphasize the performance in the classification of the different activities, it was chosen to set β at 0.6 and α at 0.4. pc is the total number of predication classes for seen dataset and gc is total number of generalisation class for unseen classes. $TP(c)$ is for true positives for class c; $FP(c)$ is for false positives for class c.

IV. Analysis of the Results

This section presents key results from the finalist submissions to this challenge in terms of performance, pre-processing, and classification methods reported in Section IV-A. The submitted papers for this challenge are labeled and referenced as [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38]. By finalist submissions, we mean that 12 teams submitted results for the *seen data* and *unseen data*, as well as a paper submission for the online hosting conference. In Section IV-B, we will give some insight into some of the pre-processing strategies to prepare the data for the machine learning algorithms. Among 12 papers that were submitted to this challenge as part of the IET radar conference, convolutional neural network (CNN) methods dominated the best performing models and will be discussed in Section IV-C. There was also a logic-based supervised learning model among the submissions, whose results were also outstanding in comparison to the CNN-based models in the leader table submission [29]. This logic-based supervised learning and its combination with ensemble methods will be discussed in Section IV-D.

A. Challenge Outcomes

The evaluation results reported from the validated 12 submissions can be found in Table IV.

Among these reported results, submission [34] only reported the best run for individual action accuracy using 10-fold cross-validation. Submission [35] only reported the best run for average classification accuracy.

The leader table of *unseen data* and *seen data* evaluation results are shown in Fig. 4. The finalists achieved high accuracy $>$ 90% for the *seen data*, which is very promising for the challenge of human activities classification, given the results reported in prior studies ranging from 65% to 99% seen in the open literature from various datasets [4], [7], [15]. It is worth emphasizing here that the 1st place holders for the *unseen data* [27], [35], [36], [37] have used various forms of

TABLE IV
THE EVALUATION RESULTS FROM 12 SUBMISSION TEAMS

Paper ID	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34][1]	[35][2]	[36]	[37]	[38]
Cross-validation	10-fold	10-fold	10-fold	10-fold	10-fold	N/A	3-fold	10-fold	N/A	N/A	5-fold	N/A
Walking	100	99	100	100	100	100	100	100	100	99.7	99	100
Standing	97.1	98.7	95.5	98.4	94	100	100	100	100	95.99	97	100
Sitting down	88.9	97.4	95.2	98.5	100	88	100	100	100	98.75	98	100
Picking an object	90.9	87.1	76.9	95.2	83	69	97.82	82	100	94.42	89	90
Drinking	91.4	86.4	84.6	94.4	87	100	97.82	91.3	100	90.56	88	93
Simulated falling	100	99.5	100	90.2	100	100	100	100	100	93.2	99	100
Avg *Seen Data*	94.27	94.3	92	95.4	95.43	94	99.27	93.9	100	95.4	94.8	97
Seen data ranking[3]	8	7	11	5	4	9	2	10	1	5	6	3
Unseen Data	100	93	92	94	95	94	94	94	100	100	100	84
Unseen data ranking[4]	1	4	5	3	2	3	3	3	1	1	1	6

The blue color indicates that the best run is shown instead of the average accuracy because the data is missing
[1]Individual action accuracy best run, average accuracy 10-fold cross-validation [2]Individual action and average accuracy – best run [3]average accuracy from the *seen data* dataset [4]best submission submitted to the challenge website on the *unseen data* dataset

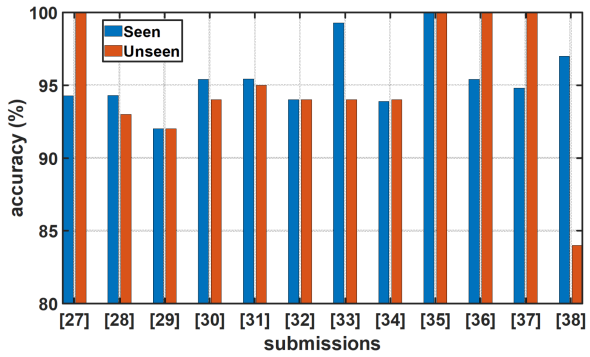

Fig. 4. The results for *unseen data* and *seen data* evaluation scores from the leader board submissions.

ensemble strategies. All those methods used cross-entropy loss functions, while 80% adopted learning rates of 0.001. [31] used in the *unseen data* evaluation challenges a variety of deep networks derived from computer vision to leverage the pre-training of these networks to boost performances. MobileNetV2, a lightweight network with 2.23M hyperparameters, achieved the highest performance amongst their selection of tested models on the *seen data* with 95.43% accuracy overall. The model from submission G0164 [34] and G0121 [30] both used slightly modified ensemble methods, which achieved the joint 4th place. Furthermore, Fig. 4 shows that the performance trends of the top 3 teams are generally similar, which results in substantial gains in classification performance either in unseen data or seen data.

The summary of the techniques used by these 12 submissions can be found in Table V. Network architectures and ensemble methods from these validated submissions are discussed in the following sections.

As illustrated in Table V, RT (range-time), RD (range-Doppler), and mD (micro-Doppler signature) were generally used for training neural networks, except for two teams who proposed to use the 3D data cube. One approach used a range-Doppler-amplitude time surface to generate a "point-of-cloud" and the spectrogram phase [27]. A threshold-based iso-surface extraction method was then used to compress the "point cloud" [27]. As the authors reported, the time domain complex radar data (in-quadrature and in-phase) are processed through the 2D Fourier transform with a sliding window to obtain a series of range-Doppler images. The sequence of range-Doppler images captures time-domain information as well. Since these images describe the evolution of the energy distribution in the range-time domain, the interval between two successive range-Doppler images is the time step of the sliding window. So the neural network input data form a 3D format with range, Doppler, and the phase of a micro-Doppler signature at a given time. Guo et al. reported a comparison between a simple convolutional neural network and their proposed Phase-Net [27]. With traditional Doppler-Time spectrograms as input, this CNN had almost the same network structure as their proposed complex field-based fusion network (CFFN). The proposed "point-of-cloud" inputs showed superior prediction accuracy over activities like walking, standing up, sitting down, and falling down, whereas there is a severe confusion between picking up an object and drinking water.

An alternative approach used range-time maps and micro-Doppler spectrograms [37]. Liu et al. proposed a feature extraction method based on motion pattern from singular value decomposition (SVD) on limbs and torso [36]. The micro-Doppler features of limbs and torsos were extracted into principal Doppler patterns and principal human activity timing patterns.

Finally, a transfer learning method was used as part of pre-processing to provide weights initialization. It is interesting to notice that transfer learning did not boost the classification performance in most cases. However, it provided fast training, as reported in other works in the literature [25].

Fig. 5 shows a box-plot chart from the performance of the unseen data evaluation submissions. It is interesting to see that the overall performance for classification in data collected from volunteer 1 is worse than for the data collected from the other volunteers. This may be related to the fact that volunteer 1 is female and the other volunteers in the unseen data are all males. Furthermore, also the larger training dataset is collected from a majority of male participants. This suggests the importance of considering gender balance when planning data collections on which data-driven approaches will be based. While, to the best

TABLE V
SUMMARY OF RADAR SIGNAL PRE-PROCESSING, INPUT FORMATS AND CLASSIFICATION TECHNIQUES FROM 12 VALIDATED SUBMISSIONS FOR THE FIRST RADAR CHALLENGE. EACH COLUMN REFERS TO A DIFFERENT SUBMISSION WITH RELATED PUBLICATIONS REFERENCED IN THE TEXT*

Ref	**[27]**	**[28]**	**[29]**	**[30]**	**[31]**	**[32]**	**[33]**	**[34]**	**[35]**	**[36]**	**[37]**	**[38]**
Leaderboard ranking	1	4	5	3	2	N/A	3	3	1	1	1	6
Raw radar data	FFT	FFT	FFT	FFT	FFT	FFT	FFT	WT denoise SOI FFT	FFT	FFT	FFT (Ham)	FFT
RT	MTI1, STFT (95, 200, 4x, rect)	MTI1, ROI, STFT (95, 200, 4x, Ham)	MTI1, ROI, STFT (95, 200, 4x, Ham)	MTI1, ROI, STFT (95, 200, 4x, Ham)	MTI1, STFT (95, 200, 4x, rect)	MTI1, STFT (95, 200, 4x, rect)	MTI1, ROI, STFT (95, 200, 4x, rect)	STFT (97.6, 500, 4x, Han)	MTI2, adaptive ROI, STFT (95, 200, 4x, rect)	MTI2, STFT (95, 256, 4x, rect)	MTI2, ROI, STFT (50, 128, 4x, Ham)	MTI2, ROI, STFT (95, 200, 4x, rect)
mD	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	ROI	ROI	ROI	N/A
RD	Summing CFAR	Summing	Summing	Summing	Summing	Summing	Summing	Summing	Summing	OS CFAR	Summing	Summing
Train	90%	90%	90%	90%	80%	97%	80%	80%	100%	N/A	80%	95%
Valid	10%	10%	10%	10%	10%	N/A	30%	10%	N/A	N/A	20%	N/A
Test	10%	10%	10%	10%	10%	3%	20%	10%	100%	N/A	20%	5%
Convergence (epochs)	50	400	N/A	N/A	4k	N/A	50	60	N/A	N/A	N/A	N/A
Method	Pointnet phase CNN, deep fusion, voting of 4 models	GRU	SVM quadratic kernel, Chi-squared feature	SVM/LR. kernel, feature	Mobile net v2, TL	VGG19 0.7 pruning	Data Aug. by varying the ROI, VGG16	CNN with deep fusion	CNN with leaky ReLU, weighted average of 6 models fed with 6 variations data	Faster RCNN	CNN+ LSTM	DTW+ 3-NN
Data domain	*RD(A): CFAR, Iso surface extr. Point cloud. *mD(ϕ) : Gray scale.	*mD(A) Gray scale	*RT(A) *mD(A) *CVD(A)	*mD(A) *CVD(A)	*mD(A)	*mD(A)	*mD(A): resized 100x100	*RT(A) mD(A): Resized 128x128	*mD(A): Resized 250x250	*mD(A): crop ±6m/s, resized 981x800 352x384 800x600	*RT(A): resized 76x20 *mD (A): resized 76x32	*mD(A): 175X60, intensity scaling 256 levels

[1]RT: range-time, RD: range-Doppler, mD: micro-Doppler signature, CVD: cadence-velocity diagram, A: amplitude, ϕ: phase, MTI1: 4th-order Butterworth Infinite impulse response, MTI2: two-pulse canceller, STFT: short-time Fourier transform (overlapping factor in %, window size in ms, zero-padding factor, window), CFAR: constant false alarm rate, OS-CFAR: order statistic CFAR, ROI: regions of interest, SOI: signals of interest, WT: wavelet transform, rect: rectangular window, Ham: Hamming window, Han: Hanning window. GRU: gated Recurrent Units, DTW: dynamic time warping, 3-NN: K = 3 Nearest Neighbour, LR: Linear regression.

* Note that the training/validation/test split for the different techniques presented may vary which may have some statistical influence on the results for the training accuracy for the *'seen'* data, however it does not influence the inference for the *'unseen'* data set.

of the authors' knowledge, detailed studies on gender-related differences in the radar signatures of human activities do not exist in the open literature, several studies hypothesized the fact that such differences occur and can be perceived with modern radar sensors, as it is possible for example with cameras [40]. Specifically, [41] is among the first papers to claim that differences in movements of male and female can be perceived by radar, whereas [42] presents some evidence based on quantitative gait parameters extracted from radar signatures of male and female volunteers.

B. Pre-processing

The participants came up with similar strategies for pre-processing, with a classic processing chain to generate micro-Doppler signatures. This followed range compression using a Fast Fourier Transform (FFT) on the beat frequencies (i.e., raw data straight from the ADC) to obtain the range profiles. Then accumulating the range profiles over a time period to process on every range bin a short-time Fourier transform (STFT) to obtain a range-Doppler image that is then summed for every Doppler bin to obtain a slice of the micro-Doppler image and then repeated to obtain the full spectrograms. There were some minor variations in this chain, as detailed in Table V.

Denoising: From the raw data, [34] used a wavelet transform approach to first denoise the signal before reconstructing it using Daubechy's db10 wavelet. This had the benefit of both reducing the noise by retaining only the decompositions containing the signals of interest and improving the mD signature as it presented fewer breaks compared to a classic approach.

Range compression: All contributions used FFT to generate the range profiles; only [37] used a Hamming window on the raw data to smooth the range profiles at the cost of widening the range peaks.

Moving target indicator (MTI) approaches were used in all submissions except [34]. MTI is used to remove stationary

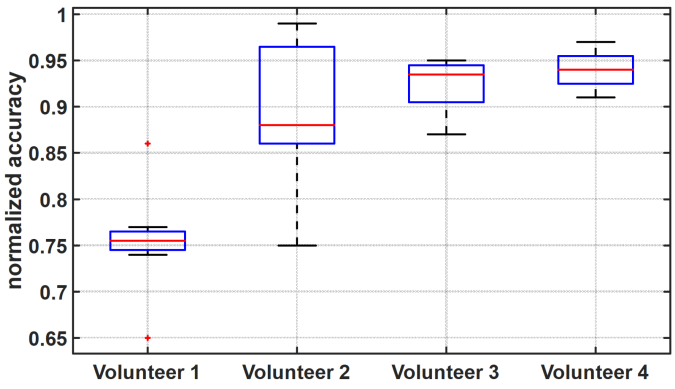

Fig. 5. The leader table submission classification accuracy for unseen validation data from each volunteer. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' marker symbol.

clutter from the environment to improve the contrast between static and moving targets. [27], [28], [29], [30], [31], [32], [33] used a 4th-order Butterworth high-pass filter with an Infinite Impulse Response (IIR) with a cut-off at 0.075 Hz. IIR has a lower implementation cost than FIR filters, and has low latency. However they have non-linear phase characteristics and can create instabilities due to the infinite impulse response. Knowing that the phase information in recovering Doppler information is critical, the non-linearity of this implementation can affect the quality of the signatures. An alternative MTI implementation in [35], [36], [37], [38] is a 2-pulse canceller that subtracts the pulses in the time domain and filters out static clutter. It creates a filter in Doppler speeds for the clutter but also blind speeds at multiples of the pulse repetition frequency (PRF) of the radar. Luckily for this application, a human target indoors is not likely to attain speeds in excess of 45km/h in the C-band. Even though the Doppler of the foot can be 2-3 times that of the centroid of a human for a brisk walk or run, the aliasing caused by such speed is very unlikely to happen indoor, especially in the context of assisted living where patients are older adults or patients suffering from a condition such as chronic obstructive pulmonary disease.

Region of Interest selection: Before moving to the range-Doppler domain, some submissions [27], [28], [29], [30], [33], [35], [37], [38] selected a region of interest from the range-time domain to reduce the noise in the following domain by only incorporating range bins with sufficient energy present before applying the next pre-processing step, namely, the STFT. [28], [29], [30], [33], [37], [38] set the range bin interval to be used manually. This means they looked into the database and empirically determined the region of interest. Note [33] used multiple intervals to create diversity for training. The others used adaptive methods that are more pragmatic for real scenarios. [35] determines a threshold based on the average energy in that range bin and only keeping values active above that threshold. [27], [36] use constant false alarm rate (CFAR) and order-statistic (OS) CFAR, respectively, to determine whether a target is present or not by testing the cells around the cell under test.

Spectrogram generation: The range-Doppler (rD) domain is now formed using FFT with various parameters for the overlapping ranging from 50% to 95%, the number of range profiles used to generate the rD image ranging from 200 to 500 ms, different windowing functions (rectangular, Hamming, Hanning), and a zero-padding factor of 4x for all contributions. The setting of these parameters influences the classification results, as shown in [43] and its optimization for an activity or a set of activities is an important parameter. The second aspect to consider is computational load: when an FFT operation is performed with a number different from a power of 2, then 3 FFTs are required to produce the DFT results, whereas if a power of 2 is selected, then the FFT-radix2 algorithm can be used directly, thus considerably reducing the computational load. A smaller overlap to generate the rD domain will also reduce the computational load with longer strides in the data. Even if this will make the results harder to understand for humans, machine learning algorithms can still make sense of it better.

From the rD domain, the range cells are added coherently or incoherently for each Doppler bin to create one slice of the micro-Doppler (mD) domain. This is repeated until a mD signature is obtained. [35] and [38] were the only teams who determined ROIs from the mD domain. [35] aimed to remove the noise from mD signature by adjusting the dynamic range of the amplitudes to -40 dB to 0 dB, as the noise did not offer salient information for classification. [37] and [36] only kept a limited range of Doppler bins around the zero-frequency for classification. This was done by essentially cropping the signatures with a lower and upper limit containing the signatures to help improve classification accuracy and allow the machine learning algorithm to only learn from content-rich data.

Radar data domain for classification: It is worth noting that all contributions used the mD signature for classification. The majority used the amplitude only, which means they can sum the amplitudes incoherently to recover the amplitude for the mD signature. However, [27] used the phase information of the mD signature. The RT domain amplitude is used in [29], [34], [37] for classification leveraging the range information for classification, as it may help to distinguish from mD signatures that are similar but would have a larger range spread, for example. [29] and [30] utilize the CVD which is obtained by applying an FFT on the mD signatures to determine if there are frequencies repeating themselves across the Doppler bins. This would indicate periodicity in a motion, such as walking that is cyclic, as opposed to a fall that is an aperiodic motion. Finally, [27] used an rD surface. First, a CFAR algorithm was used to set a threshold to extract the surface in 2D, and then they were collated together into a 3D surface for classification, yielding a range-Doppler time iso-surface.

C. Neural Network Architectures

The popular backbone networks used in this challenge were VGG [22], AlexNet [23], and mobileNet [24]. VGGNet increased the number of layers to 16 or 19 layers compared with

the conventional 8 layers depth in AlexNet. Another typical deep learning network is mobileNet, a deep learning network architecture with around 50 layers. Apart from the higher number of layers providing more depth in this network, mobileNet also uses a recurrent neural network (RNN) structure. Three of the six activities can be recognized almost at 100% when adopting mobileNet [31], namely fall, walking, and standing up. The human activities of standing up and sitting down can also be recognized very accurately and quickly. However, the "pick up an object" and "drink water" activities are very difficult to distinguish, with confusion occurring between them for the majority of the submissions, with the exception of [35] that got 100% on all activities and [31] that confused sitting down with picking up an object.

Furthermore, the RNNs used in this challenge were Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) [28]. This team reported an average accuracy of 94.3% when using the gated recurrent unit, and 93.9% for the LSTM network. They also reported a difficulty in distinguishing the activities of "picking up an object from the ground" and "drinking water from a glass" as well. An advantage of using LSTM or GRU networks is that they can use an arbitrary length of the sample in time for the inference, as opposed to vision-based networks that require resizing the samples to tiles of a specific size, e.g., 128x128 for a 'home-brew' CNN network. This particular feature can be used for data augmentation as this means that samples can be truncated to random sizes as long as they are not under 0.5 s as the performances decrease in that case. This method yielded the best performances reported in that submission.

The techniques of data augmentation have shown notable performance improvements, which can be found in [33]. The authors reported that the use of data augmentation methods achieved 99% test accuracy during training and produced a good generalization of the model. They analyzed the effects of two data augmentation methods: selecting range bins before STFT; and varying the amplitude range of the spectrogram display after STFT before being converted to images. Their results had demonstrated that data augmentation with range bins before STFT was more efficient in identifying human motion activities compared with data augmentation using spectrogram amplitude variations. Chen et al. [37] used random cropping and scaling for data augmentation, which would work in the context of radar given that the cropping does not clip the signature and that scaling is kept within physical realms. However, rotations and shearing, as can be found in some vision-based techniques for images [44] are not suitable for radar classifications, as a 20° rotation on a signature is not going to happen in a real system and shearing will create a skewness in the data distribution that would not happen either in real-systems. However, adding noise is a good way of augmenting data. This will create random variations that will reflect the signal-to-noise ratio (SNR) of the system that can happen from a lower quality receiver with a higher noise figure, as well as for a further target away from the receiver, or maybe occluded behind an obstacle.

Zhang et al. [38] is also worthy of notice, as this team implemented a dynamic time warping (DTW) technique. The DTW algorithm was originally used to match sequence data that are of different lengths. In essence, this corresponds to comparing samples that might be scaled in time to one another and estimating a distance to other samples present in the database to determine which class it belongs to. The disadvantage is that it requires more operations as the size of the dataset increases.

TABLE VI
NETWORK PARAMETER NUMBERS FOR TYPICAL NEURAL NETWORK ARCHITECTURES

Network Architecture	hyperparameters (M)[1]	average accuracy
VGG-19 [32]	20.02	93%
MobileNetV2 [31]	4 (estimated)	95.43%
CNN (5 layers) [56]	1.72	89%
VGG-19 pruning 0.7 [32]	1.8	94%
Double CNNs [37]	0.0856	94.8%

[1]M: Million

TABLE VII
AN EXAMPLE OF AN ENSEMBLE METHOD USED FOR HUMAN ACTIVITIES

Range time (128 x 128)	Spectrogram (128 x 128)
*Conv Kernel	*Conv Kernel
BatchNorm X 5	BatchNorm X 5
GlobalAvg,Dropout(0.5)	GlobalAvg, Dropout(0.5)
Fully Connected (128), RelU Dropout(0.5)	
Fully connected (6), Softmax	
*Conv Kernel included two convolution layers and a max pool layer	

The number of parameters indicates the complexity of neural network architectures. It is interesting to see that the network's performance was not linearly correlated with the number of parameters. Table VI lists the number of parameters reported in the main paper submissions. All these networks achieved an average classification accuracy greater than 89%. In [32], the authors demonstrated the VGG19 network could achieve higher performances with pruning (0.7) gaining 1% over the full VGG19 network for less than one-tenth of the size. Pruning allows reducing the number of hyperparameters from 20M down to 1.8M in this case.

D. Notable Ensemble Methods

It is worthy to note that ensemble methods are the most popular methods among the top-ranked teams in the challenge. The team from Centrale-Supelec [34] reported an ensemble method for neural-network-based supervised learning, as shown in Table VII. In this method, the range-time and the spectrogram convolutional networks were identical and formed branches for a later deep feature fusion. The last convolutional network layers were global average pooling layers. Prior to this layer, the inner representations are tiles stacked in $8 \times 8 \times 64$. Two reasons justify the use of global average pooling: 1) feature reduction for the following fully connected layer and 2) improve generalization by generating more invariant features. During training, the validation accuracy reached up to 96.6%.

A second ensemble method was proposed, where the architecture used a CNN to capture features from range-time domain and Doppler-time domain inputs [37]. Those automatically captured features were used to train an LSTM network (as shown in Fig. 6).

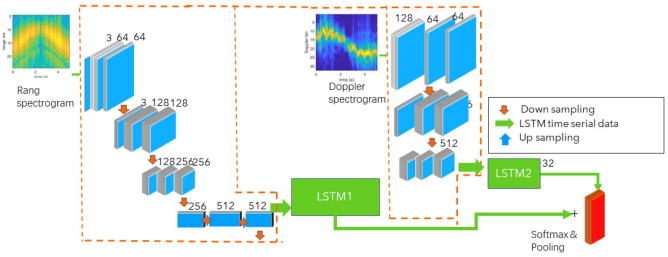

Fig. 6. The architecture of a combined approach of LSTM with 2 CNNs adapted from [37].

The third notable ensemble model was based on six CNNs [35]. By initializing the parameters of these models and setting the number of training iterations, the authors used six trained CNNs that correspond to the six variations of the pre-processing. A weighted voting step was then used to produce the prediction score for unseen data. Thanks to this weighted voting system, this team was the first to report 100% predictions in the unseen evaluation leader table.

V. Discussion and Perspectives

In this paper, the results from the inaugural challenge of radar-based human classification organized as part of the IET International Radar Conference 2020 were presented. The paper discussed the proposed top-performing classification models across the leader board submissions and their key results on the common database used for benchmarking.

A. Insight Into the Training Dataset

Most of the teams used micro-Doppler spectrograms for classification, a data domain for which the six activities in the dataset could generally be distinguished clearly. For example, "sitting down" produced negative Doppler, "standing up" positive Doppler, and "drinking water" 'almost random' Doppler as the arm movement to grab the glass was not tied to a specific direction. Leveraging these differences from the predefined directions at which the activities were recorded, even simple classifiers (e.g., a 5-layer CNN) could achieve an average of over 90% in classification accuracy. However, the two activities, "pick up an object" and "drink water," were the hardest to classify for most teams and proposed methods. Besides the varying direction of the arm movement for the "drinking" activity, one should also notice that these two are not single actions but more interaction activities consisting of several actions in sequences. For example, "drinking" included stretching the arm, grabbing the glass, flexing the arm back, and in many cases stretching the arm again to put back the glass after one or more sips of water. "Bending" meant moving the upper body forward towards the radar (similar to some extent to standing up) and then moving back up with the object in the hands (similar to moving back when sitting and also to bringing the glass to the mouth for drinking).

In other words, activities labeled as separated classes may present kinematic similarities between each other, resulting in less distinguishable radar signatures. Furthermore, in some cases, especially for some of the older participants, the extent of the bending of the body and stretching of the arm for drinking was accordingly limited to reduce their physical effort. This created intra-class differences between signatures of activities having the same label, hence an additional challenge for automatic classification. This highlights the fact that algorithms need to be trained on the intended target population, as 20–30 year volunteers working in labs do not perform activities in the same way as 70+ years old adults. A potential solution could also be to decompose such actions into atomic elements that constitute a more complex activity, thus developing a semantic representation of atomic actions that then form more complex ones from their combination.

B. Prospective Classification Methods

Recent achievements of CNN-based models on image classification tasks have encouraged scholars to modify such models for many other tasks. This includes radar-based classification tasks, as also extensively seen in the analysis of the methods proposed for this challenge. In the convolutional neural network, as sigmoid is easily leading to the exploding gradient problem, more than half of the teams chose to use the ReLU in CNN-based architectures [33]. Since ReLU has "black spots" of negative hyper-dimensional planes, which sometimes become inactive for all inputs, one team chose LeakyRelU instead [36]. As the number of layers of the neural network deepens, the stochastic gradient cannot propagate down into deeper layers; for this issue, a "highway" system, such as ResNet, was chosen by the team in [30]. Another way of avoiding exploding stochastic gradients is to add batch normalization [33]. However, in recurrent neural networks, adding batch normalization is not stable, so the solution is to add dropout or gate control to/from LSTM and the simplified LSTM (i.e. the GRU). With the latest success of generative adversarial neural networks (GANs), we hope to see some applications of using GANs to improve the unbalanced data classification for some activities [58]. GANs use adversarial methods to train generator networks for synthetic data. Some very interesting literature using physics-aware GAN is emerging for increased fidelity radar signatures [45].

According to the analysis in this paper, of all CNN-based models that have been reported, the most successful results in the challenge leader board were based on an ensemble method. Ensemble learning aims to build a prediction model by combining the strengths of a collection of simpler weaker learners. Ensemble learning can be broken down into two tasks: developing a population of base learners from the training data and then combining them to form the composite predictor. In the previous section's analysis, We observed several examples falling into this paradigm. For example, the ensemble of six trained networks for six activities cast one vote each for the classification [36]. Any dictionary method using weak learner basis functions could be considered an ensemble method. These methods could be Bayesian, regression splines, or non-parametric methods. The general idea is to combine/average the outputs of different classifiers to boost the prediction performances. In the future, we hope

to see automated optimizers that conduct a multi-variate search within weak classifiers and the voting structure for ensemble methods that will maximize performances in accuracy and the other key indicators for the application constraints.

C. Future Work

Compared to statistical based machine learning methods, one of the most challenging problems in the current applications using artificial neural networks is the non-deterministic nature of the non-linear regularization used in activation layers.

Compared with linear counterparts, artificial neural networks must deal with the existence of multiple local minima on the error function surface (i.e., multiple solutions). Linear regression models present one global minimum. However, AI algorithms may find a multitude of local minima that may be satisfactory from the cost function perspective. However, the generalization of those models depending on the selected local minimum will vary widely and need to be evaluated individually. This is why cross-validation is important to ascertain the average behavior of proposed AI models. This creates a generalization problem for human activities classifications. Each human subject has different body shapes and individual walking gait, which is especially evident with multiple activities to be classified simultaneously.

Designing neural networks where all parameters such as initialization weights, network architecture, batch size, and many more are properly set to ensure robustness and generalization for radar-based automatic human activities classification remains an open challenge.

Feature extraction from radar data can leverage the increasing dimensionality of radar signals (e.g., higher range resolution), frontends with multiple input and multiple output capabilities, increasing angular and spatial diversity, and on the various data representations ranging from raw IQ data directly sampled by the radar directly to range-time, range-Doppler, range-azimuth, angle-of-arrival, spectrograms, cadence velocity diagrams, cepstrograms, phase plots [46], [47], cyclostationarity signatures [48], radon transform signatures [49], and other composite views e.g., range-Doppler surfaces [30]. Thus, open research questions remain as to what format or combination of formats are most suitable for the classification process, perhaps exploiting forms of cognition that modify accordingly such processes depending on the specific activities to be classified. To enable better feature extraction, the importance of the role of pre-processing should not be ignored for the final classification accuracy. We saw in [34] that de-noising of the raw radar data improved the overall performance. We also saw different parameters for the STFT (window size/type, overlap) from the different submissions, suggesting that optimization can be found in those parameters or exploring other time-frequency transforms. We also note that participants usually tried in one way or another to identify regions of interest either from observation or using adaptive methods to improve the accuracy of classification. The focus of future research should be on methods to render pre-processing adaptive to be more robust to variations in signal-to-noise ratio and signal intensity. An example of the range of possibilities when extracting features from multiple radar data domains is increasing from the increase in dimensions and the range of features that can be defined. The classification accuracy using range-Doppler-time representations into a volume before extracting features in [27] exhibited a faster convergence, leading to a shorter network training time. This can enable the deployment of deep learning classifiers onto resource-constrained platforms. The effect of the precision of the representation for signal processing and classification (full/half-precision floating point, fixed precision) and the reduction of the general footprint (hardware resources, energy) are open challenges for real-time implementation with works emerging in this area [50], [51], [52].

To ensure consistency in the inferences of the network and their measures, the proposed machine learning algorithms must be stress-tested under different conditions. Various strategies are proposed in [57] to get the best out of neural networks and determine the source of underperformance. Recently, different strategies such as training with a dynamic gradient, using the Levenberg-Marquardt training function, cross-validation, and hold out are evaluated on the same datasets to understand the effect of initialization and the architecture on the key performance indicators. Our challenge proposed a two-stage cross-validation method, which is a step forward to benchmark models in terms of generalization. Further benchmarks for evaluation training kernels will be needed in future work using the classification algorithms across different datasets e.g., [21], [53], [54]. Furthermore, this will provide the basis for reproducible state-of-art baselines for radar-based human activity classification challenges.

VI. Conclusion

We organized the 'Human Activity Radar Challenge', a classification challenge for radar-based human activities where raw radar data (i.e., without any pre-processing) were provided to participants as input to further processing in view of the application of machine learning for classification. This Challenge was open in 2020 and affiliated with the IET International Radar Conference of the same year, later postponed to 2021 and then canceled because of the COVID-19 pandemic.

In this paper, we presented a summary of the main methods proposed in the Challenge, including the task, data, performance metric, results, and performance analyses. The primary objectives were to systematically measure the recent progress in human activities classification in radar, particularly in the machine learning classification domain, and stimulate new ideas and collaborations. This paper's results and analyses indicate great progress in human activities classification in radar data, with the performance for walking as accurate as 100% for the leading methods. This opens the way for further gait analysis to extract parameters for timed-up and go test and balance tests that rely on gait features and speed [55]. Nevertheless, the performance gap in certain activities remains relatively large, at least for motion-based activities. This motivates further research towards developing a more robust technology that can maintain performance across a wide range of activities, also leveraging on the increasingly available public datasets [59] such as [53], [60], [61], [62], [63], [64], [65], [66], [67].

Future benchmark challenges should consider more comprehensive metrics to define a figure of merit to better gauge the performances of various software especially in terms of deployment and footprint considerations. The metrics that should be considered in the future are convergence, training time, inference time, number of hyperparameters, pre-processing and classification (Giga Operations / second), Accuracy, Sensitivity, Precision and F1 score. This would encompass both the machine learning performance and the pre-processing required to format the data prior to classification for resource-constrained embedded platforms. Furthermore, the teams will need to adhere to prescribed data splits for machine learning and cross-validation to ensure that the comparisons are statistically equivalent for the 'seen' dataset addressing one of the shortcomings of this challenge. Authors will be asked to submit their algorithms as well to run the training time and inference time on the same platform to have fairer performances comparison using the same implementation for an improved benchmarking.

Acknowledgment

Authors' Affiliations

Shufan Yang is with the School of Computing, Edinburgh Napier University, EH11 4BN, Scotland (e-mail: s.yang@napier.ac.uk).

Julien Le Kernec is with the James Watt School of Engineering, University of Glasgow, G128QQ, Scotland (e-mail: Julien.lekernec@glasgow.ac.uk).

Olivier Romain is with the ETIS Lab, University-Cergy-Pontoise, 95011, France (e-mail: olivier.romain@cyu.fr).

Francesco Fioranelli is with the MS3 group, Department of Microelectronics, TU Delft, 2628 CD, The Netherlands (e-mail: f.fioranelli@tudelft.nl).

Pierre Cadart is with the CentraleSupélec, 57000 Metz, France (e-mail: pierre.cadart@gmail.com).

Jérémy Fix is with the Université de Lorraine, CentraleSupélec, CNRS, 57000 Metz Loria, France (e-mail: jeremy.fix@centralesupelec.fr).

Chenfang Ren, Giovanni Manfredi, Thierry Letertre, and Israel David Hinostroza Sáenz are with the SONDRA, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: chengfang.ren@centralesupelec.fr; giovanni.manfredi@centralesupelec.fr; thierry.letertre@centralesupelec.fr; israel.hinostroza@centralesupelec.fr).

Jifa Zhang, Huaiyuan Liang, and Xiangrong Wang are with the School of Electronic and Information Engineering, Beihang University, 100191, China (e-mail: jifazhang@yeah.net; lianghuaiyuan@buaa.edu.cn; xrwang@buaa.edu.cn).

Gang Li and Zhaoxi Chen are with the Department of Electronic Engineering, Tsinghua University, 100084, China (e-mail: gangli@mail.tsinghua.edu.cn; chenzx16@mails.tsinghua.edu.cn).

Kang Liu is with the College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China (e-mail: kang.liu@cjlu.edu.cn).

Xiaolong Chen is with the Marine Target Detection Research Group, Naval Aviation University, Yantai 264001, China (e-mail: cxlcxl1209@163.com).

Jiefang Li and Xing Wu are with the Shanghai Key Laboratory of Multidimensional Information Processing, School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China (e-mail: lijiefang1@163.com; xwu@ee.ecnu.edu.cn).

Yichang Chen is with the Early Warning Academy, Wuhan 430019, China (e-mail: cyc_2007@163.com).

Tian Jin is with the College of Electronic Science and Technology, National University of Defense Technology, 410073, China (e-mail: tianjin@nudt.edu.cn).

References

[1] 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 Process. Mag.*, vol. 33, no. 2, pp. 81–94, Mar. 2016, doi: 10.1109/MSP.2015.2503881.

[2] 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.

[3] C. Li, P. Mak, R. Gómez-García, and Y. Chen, "Guest editorial wireless sensing circuits and systems for healthcare and biomedical applications," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 2, pp. 161–164, Jun. 2018.

[4] J. L. Kernec et al., "Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 29–41, Jul. 2019.

[5] J. A. Nanzer, "A review of microwave wireless techniques for human presence detection and classification," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 5, pp. 1780–1794, May 2017.

[6] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[7] 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.

[8] F. Fioranelli and J. L. Kernec, "Radar sensing for human healthcare: Challenges and results," in *Proc. IEEE Sensors*, 2021, pp. 1–4, doi: 10.1109/SENSORS47087.2021.9639702.

[9] 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.

[10] S. Z. Gürbüz, B. Erol, B. Çağlıyan, and B. Tekeli, "Operational assessment and adaptive selection of micro-Doppler features," *IET Radar, Sonar Navigation*, vol. 9, no. 9, pp. 1196–1204, 2015.

[11] H. Li, A. Shrestha, H. Heidari, J. L. Kernec, and F. Fioranelli, "A multisensory approach for remote health monitoring of older people," *IEEE J. Electromagn., RF, Microw. Med. Biol.*, vol. 2, no. 2, pp. 102–108, Jun. 2018.

[12] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan, "Micro-Doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar Navigation*, vol. 11, no. 1, pp. 107–115, 2017.

[13] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 9, pp. 2629–2640, Sep. 2019.

[14] B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic, "Doppler radar fall activity detection using the wavelet transform," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 865–875, Mar. 2015.

[15] X. Li, Y. He, and X. Jing, "Survey of deep learning-based human activity recognition in radar," *Remote Sens.*, 2019, vol. 11, Art. no. 1068.

[16] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2164–2180, Oct. 2019.

[17] M. S. Seyfioglu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[18] X. Li, Y. He, Y. Yang, Y. Hong, and X. Jing, "LSTM based human activity classification on radar range profile," in *Proc. IEEE Int. Conf. Comput. Electromagn.*, 2019, pp. 1–2, doi: 10.1109/COMPEM.2019.8779144.

[19] A. Shrestha, H. Li, J. L. 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.

[20] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. L. Kernec, "Radar sensing for healthcare," *Electron. Lett.*, vol. 55, no. 19, pp. 1022–1024, 2019.

[21] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. L. Kernec, "Radar signatures of human activities," Univ. Glasgow, 2019, doi: 10.5525/gla.researchdata.848.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[23] K. Alex, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] S. Mark, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520 .

[25] S. Yang, C. Lemke, B. F. Cox, I. P. Newton, I. Näthke, and S. Cochran, "A learning based microultrasound system for the detection of inflammation of the gastrointestinal tract," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 38–47, Jan. 2021.

[26] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997, pp. 1895–1898.

[27] J. Guo et al., "Complex field-based fusion network for human activities classification with radar," in *Proc. IET Int. Radar Conf.*, 2020, pp. 68–73, doi: 10.1049/icp.2021.0572.

[28] H. Jiang, F. Fioranelli, S. Yang, O. Romain, and J. L. Kernec, "Human activity classification using radar signal and RNN networks," in *Proc. IET Int. Radar Conf.*, 2020, pp. 1595–1599, doi: 10.1049/icp.2021.0556.

[29] Z. Li et al., "Multi-domains based human activity classification in radar," in *Proc. IET Int. Radar Conf.*, 2020, pp. 1744–1749, doi: 10.1049/icp.2021.0557.

[30] X. Li, F. Fioranelli, S. Yang, O. Romain, and J. L. Kernec, "Radar-based hierarchical human activity classification," *IET Int. Radar Conf.*, 2020, pp. 1373–1379, doi: 10.1049/icp.2021.0566.

[31] Z. Xiaolong, J. Tian, and D. Hao, "A lightweight network model for human activity classifiction based on pre-trained mobilenetv2," in *Proc. IET Int. Radar Conf.*, 2020, pp. 1483–1487, doi: 10.1049/icp.2021.0595.

[32] D. Hao, J. Tian, D. Yongpeng, and X. Zhuo, "A compact human activity classification model based on transfer learned network pruning," in *Proc. IET Int. Radar Conf.*, 2021, pp. 1488–1492, doi: 10.1049/icp.2021.0609.

[33] J. Li, X. Chen, G. Yu, X. Wu, and J. Guan, "High-precision human activity classification via radar micro-doppler signatures based on deep neural network," in *Proc. IET Int. Radar Conf.*, 2020, pp. 1124–1129, doi: 10.1049/icp.2021.0651.

[34] P. Cadart et al., "Classification in C-band of doppler signatures of human activities in indoor environments," *IET Int. Radar Conf.*, 2020, pp. 412–416, doi: 10.1049/icp.2021.0597.

[35] Y. Chen, W. Wang, Q. Liu, Y. Sun, Z. Tang, and Z. Zhu, "Human activity classification with radar based on Multi-CNN information fusion," *IET Int. Radar Conf.*, 2020, pp. 538–543, doi: 10.1049/icp.2021.0676.

[36] K. Liu et al., "Micro-doppler feature and image based human activity classification with FMCW radar," in *Proc. IET Int. Radar Conf.*, 2020, pp. 1689–1694, doi: 10.1049/icp.2021.0555.

[37] Z. Chen and G. Li, "Human activity classification with neural network using radar micro-doppler and range signatures," in *Proc. IET Int. Radar Conf.*, 2020, pp. 222–227, doi: 10.1049/icp.2021.0810.

[38] J. Zhang, H. Liang, and X. Wang, "Combined spectrogram structure and intensity for human activity recognition using modified DTW," *IET Int. Radar Conf.*, 2020, pp. 1380–1385, doi: 10.1049/icp.2021.0589.

[39] B. Dawant et al., "Semi-automatic segmentation of the liver and its evaluation on the MICCAI 2007 grand challenge data set," in *Proc. MICCAI Workshop 3D Segmentation Clinic: A Grand Challenge*, 2007, pp. 215–221.

[40] D. Guffanti, A. Brunete, and M. Hernando, "Non-invasive multi-camera gait analysis system and its application to gender classification," *IEEE Access*, vol. 8, pp. 95734–95746, 2020, doi: 10.1109/ACCESS.2020.2995474.

[41] D. Tahmoush and J. Silvious, "Radar microDoppler for security applications: Modeling men versus women," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, 2009, pp. 1–4, doi: 10.1109/APS.2009.5171873.

[42] J. Zhang, "Basic gait analysis based on continuous wave radar," *Gait Posture*, vol. 36, no. 4, pp. 667–671, 2012, doi: 10.1016/j.gaitpost.2012.04.020.

[43] V. Busin et al., "Evaluation of lameness detection using radar sensing in ruminants," *Vet. Rec.*, vol. 185, pp. 572–572, 2019, doi: 10.1136/vr.105407.

[44] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop*, 2018, pp. 117–122, doi: 10.1109/IIPHDW.2018.8388338.

[45] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-aware design of multi-branch GAN for human RF micro-doppler signature synthesis," in *Proc. IEEE Radar Conf.*, 2021, pp. 1–6, doi: 10.1109/RadarConf2147009.2021.9455194.

[46] X. Zhang, Q. H. Abbasi, F. Fioranelli, O. Romain, and J. L. Kernec, "Elderly care-Human activity recognition using radar with an open dataset and hybrid maps," in *Proc. 16th EAI Int. Conf. Body Area Netw.*, 2021, pp. 39–51, doi: 10.1007/978-3-030-95593-9-4.

[47] Z. Li, J. L. Kernec, F. Fioranelli, Q. Abbasi, S. Yang, and O. Romain, "Human activity classification with adaptive thresholding using radar micro-doppler," in *Proc. CIE Int. Conf. Radar*, 2021, pp. 1511–1515.

[48] Y. Du et al., "Radar-based human activity classification with cyclostationarity," in *Proc. CIE Int. Conf. Radar*, Haikou, Hainan, China, 2021, pp. 1483–1487.

[49] M. G. Amin and R. G. Guendel, "Radar classifications of consecutive and contiguous human gross-motor activities," *IET Radar Sonar Navig.*, vol. 14, pp. 1417–1429, 2020, doi: 10.1049/iet-rsn.2019.0585.

[50] A. Bordat, P. Dobias, J. L. Kernec, D. Guyard, and O. Romain, "Towards real-time implementation for the pre-processing of radar-based human activity recognition," in *Proc. IEEE 31st Int. Symp. Ind. Electron.*, Anchorage, AK, USA, 2022, pp. 635–638.

[51] A. Pesin, A. Louzir, and A. Haskou, "A novel approach for radar-based human activity detection and classification," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2021, pp. 1–4, doi: 10.1109/ICCE50685.2021.9427670.

[52] Z. Yang, H. Wang, P. Ni, P. Wang, Q. Cao, and L. Fang, "Real-time human activity classification from radar with CNN-LSTM network," in *Proc. IEEE 16th Conf. Ind. Electron. Appl.*, 2021, pp. 50–55, doi: 10.1109/ICIEA51954.2021.9516401.

[53] D. Gusland, J. M. Christiansen, B. Torvik, F. Fioranelli, S. Z. Gurbuz, and M. Ritchie, "Open radar initiative: Large scale dataset for benchmarking of micro-Doppler recognition algorithms," in *Proc. IEEE Radar Conf.*, 2021, pp. 1–6, doi: 10.1109/RadarConf2147009.2021.9455239.

[54] R. Guendel, M. Unterhorst, F. Fioranelli, and A. Yarovoy, "Dataset of continuous human activities performed in arbitrary directions collected with a distributed radar network of five nodes," 4TU.ResearchData. Dataset, 2021, doi: 10.4121/16691500.v2.

[55] J. C. Ayena, L. Chioukh, M. J. Otis, and D. Deslandes, "Risk of falling in a timed Up and Go test using an UWB radar and an instrumented insole," *Sensors*, vol. 21, no. 3, 2021, Art. no. 722, doi: 10.3390/s21030722.

[56] 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, doi: 10.1109/LGRS.2015.2491329.

[57] N. Andrew, "Machine learning yearning," 2017. [Online]. Available: http://www.mlyearning.org/(96)

[58] B. Erol, S. Z. Gurbuz, and M. G. Amin, "GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition," in *Proc. IEEE Radar Conf.*, 2019, pp. 1–5.

[59] F. Fioranelli, S. Zhu, and I. Roldan, "Benchmarking classification algorithms for radar-based human activity recognition.," in *Proc. IEEE Aerosp. Electron. Syst. Mag.*, 2022, pp. 1–4, doi: 10.1109/MAES.2022.3216262.

[60] M. Ritchie, R. Capraru, and F. Fioranelli, "Dop-NET: A micro-Doppler radar data challenge," *Electron. Lett.*, vol. 56, pp. 568–570, 2020, doi: 10.1049/el.2019.4153.

[61] G. Bhavanasi, L. Werthen-Brabants, T. Dhaene, and I. Couckuyt, "Patient activity recognition using radar sensors and machine learning," *Neural Comput. Appl.*, vol. 34, pp. 16033–16048, 2022, doi: 10.1007/s00521-022-07229-x.

[62] Z. S. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, "Cross-frequency training with adversarial learning for radar micro-Doppler signature classification (Rising Researcher)," *Proc. SPIE*, vol. 11408, pp. 58–68, 2020, doi: 10.1117/12.2559155.

[63] M. J. Bocus et al., "OPERAnet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors," *Sci. Data*, vol. 9, 2022, Art. no. 474, doi: 10.1038/s41597-022-01573-2.

[64] Z. Meng, "Gait recognition for co-existing multiple people using millimeter wave sensing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 849–856.

[65] S. Schellenberger et al., "A dataset of clinically recorded radar vital signs with synchronised reference sensor signals," *Sci. Data*, vol. 7, no. 1, Sep. 2020, Art. no. 291, doi: 10.1038/s41597-020-00629-5.

[66] G. Beltrão et al., "Contactless radar-based breathing monitoring of premature infants in the neonatal intensive care unit," *Sci. Rep.*, vol. 12, 2022, Art. no. 5150, doi: 10.1038/s41598-022-08836-3.

[67] S. Yoo et al., "Radar recorded child vital sign public dataset and deep learning-based age group classification framework for vehicular application," *Sensors*, vol. 21, no. 7, Mar. 2021, Art. no. 2412, doi: 10.3390/s21072412.