Evaluation Metrics for Continuous Human Activity Classification Using Distributed Radar Networks

Guendel, Ronny G.; Fioranelli, Francesco ; Yarovoy, Alexander

DOI
10.1109/RadarConf2248738.2022.9764181

Publication date
2022

Document Version
Final published version

Published in
2022 IEEE Radar Conference (RadarConf22) Proceedings

Evaluation metrics for continuous human activity classification using distributed radar networks

Ronny G. Guendel, Francesco Fioranelli, Alexander Yarovoy
Microwave Sensing, Signals and Systems (MS3), Delft University of Technology, Delft, Netherlands
{r.guendel, f.fioranelli, a.yarovoy}@tudelft.nl

***Abstract*—Continuous Human Activity Recognition (HAR) in arbitrary directions is investigated using 5 spatially distributed pulsed Ultra-Wideband (UWB) radars. Such activities performed in arbitrary and unconstrained trajectories render a more natural occurrence of Activities of Daily Living (ADL) to be recognized. An innovative signal level fusion method was applied on the Range-Time (RT) maps, and deep learning classification via Recurrent Neural Networks (RNN) with and without bidirectionality was used on the computed micro-Doppler (μD) spectrogram. To assess classification performances, novel evaluation metrics accounting for the continuous nature of the sequence of activities and for imbalances in the dataset are proposed and compared with existing metrics. It is shown that conventional accuracy evaluation is too coarse, and that the proposed metrics need to be considered for a more comprehensive evaluation.**

***Index Terms*—Micro-Doppler Classification, Distributed Radar, Deep learning, LSTM, Human Activity Recognition.**

I. Introduction

Monitoring Activities of Daily Living (ADL) via radar has gained attention for safe and independent aging-in-place of older and vulnerable subjects. This includes recording critical events such as falls, monitoring abnormalities in movements and activities, and in general providing an appraisal of well-being in terms of cognitive and physical state [1], [2].

Recently, distributed networks with multiple cooperating radars have attracted significant interest for Human Activity Recognition (HAR) to address the issue of micro-Doppler (μD) signatures recorded at unfavourable aspect angles [3], [4]. Furthermore, continuous sequences of activities are increasingly investigated, as opposed to more conventional classification of artificially separated activities [5]–[7], as they represent more realistic and natural scenarios to evaluate radar-based HAR.

However, HAR on continuous sequences of activities needs additional, alternative performance evaluation metrics beyond simple accuracy or quantities directly extracted from confusion matrices, regardless of the nature of the radar used for recording, i.e., monostatic or distributed/multistatic. Specifically, four aspects of continuous HAR data are considered:

- **Continuity:** The activities are performed in a natural way - continuous sequence, where transitions between them are not only happening at arbitrary times, but are also extended in time, i.e., it is difficult even in the ground-truth to pinpoint exactly the time instant where one activity ends and the following starts.
- **Misalignments:** As a consequence of the difficulty to estimate precisely the time instant of activity transitions, misalignments between ground-truth and predictions label can happen, i.e., time offsets between ground-truth and predictions. Depending on the overall goal of the HAR system, one needs to establish how more/less important such misalignments are in terms of the performance evaluation of a classification algorithm.
- **Interruptions:** As an activity occupies an extended number of time bins, an ideal prediction would have the corresponding correct label for all of them. However, there may be cases where the classifier returns temporary short fluctuations in the predicted label for one or a very short number of time bins. This fluctuation of the predicted label is generally overlooked when classifying human activities as artificially separated images, but needs to be considered for continuous HAR and captured by performance metrics.
- **Imbalance:** When evaluating realistic sequences of activities, imbalances in the dataset can happen. A typical example, as in this paper, can be the prevalence of the *walking* class while participants move about in the room to perform other in-place activities. It is therefore important that performance metrics for the whole sequences of continuous activities account for this.

Therefore, this paper introduces a collection of 10 possible evaluation metrics including two novel ones for continuous activity classifiers that can account for the four aforementioned aspects. Applicability of these metrics with advantages and disadvantages are discussed.

Specifically, the proposed evaluation metrics are applied to the classifiers of ADL simultaneously recorded with 5 pulsed Ultra-Wideband (UWB) radars circularly spaced and covering a surveillance area of 4.39 m as in Figure 1. This comprehensive dataset includes 30 sequences of 2 minutes duration for each of the 15 participants, with activities performed in both predefined (sequence A) and random locations (sequence B) within the surveillance area. It should be noted that the participants were free to move in unconstrained directions between performing each activity, and to face random directions in terms of aspect angles [8].

The data from 5 radars are combined in the Range-Time (RT) domain with incoherent fusion, followed by generation of μD spectrograms used as input to several types of Recurrent Neural Networks (RNN). Their predictions are used to calculate and compare the proposed evaluation metrics.

The rest of the paper is organized as follows. Section II

2022 IEEE Radar Conference (RadarConf22) | 978-1-7281-5368-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/RADARCONF2248738.2022.9764181

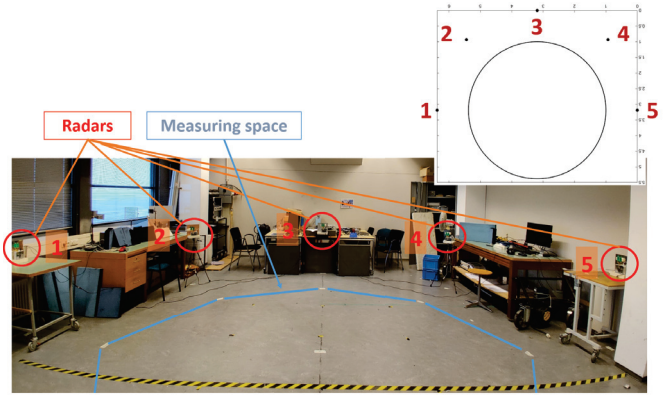


Fig. 1: Distributed radar network covering the surveillance area of about 4.39 m diameter at the Microwave Sensing, Signals and Systems (MS3) laboratory at TU Delft.

describes the experimental setup and the data format and pre-processing. Section III introduces the 10 evaluation metrics addressing concerns to evaluate continuous ADL sequences. In Section IV, evaluation metrics and their use cases are discussed, with final remarks given in Section V.

II. Experimental Setup and Signal Model

This section introduces the comprehensive dataset collected with a radar network of 5 distributed monostatic nodes for 15 participants, and discusses the radar and signal fusion model.

A. Dataset description

The collected set consists of data acquired for 5 classes, namely: (I) *translation activities* (walking), (II) *stationary activities*, (III) *in-place activities* (sitting down, standing up from sitting, bending while sitting and standing), (IV) *falling while standing or walking*, and (V) *standing up from falling*.

The sequences of 15 participants were split into test and training data by excluding one participant from the training data for testing. The procedure is well known as *leave one person out* (L1PO). For all participants, each of the collected recordings has a total duration of 2 min with (sequence A) all activities performed in predefined locations and (sequence B) freely chosen locations within the surveillance area. It should be noted that the participants were free to move in unconstrained directions between performing each activity, and to face random directions in terms of aspect angles [8].

B. Radar model

Five coherent pulsed radar nodes are employed with coded waveform capabilities minimizing interference between nodes. The experimental pulse repetition frequency (PRF) of the Humatics P410 (former PulsON) radar nodes is f_{PRF} of 122 Hz (PRI: 8.2 ms). The unambiguous Doppler frequency results in ± 61 Hz (± 2.2 m/s), and the radar filterbanks have a time-of-flight sampling rate of $\tau = 61$ ps. The unambiguous range by using a bandwidth of $B = 2.2$ GHz is approximately 68 mm according to $R = \frac{c}{2 \cdot B}$.

C. Incoherent signal level fusion

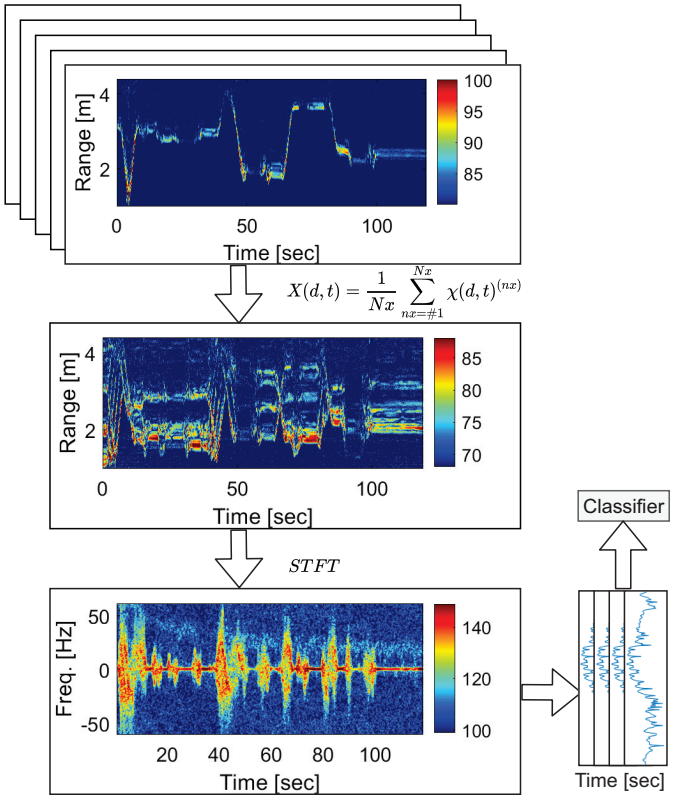


Fig. 2: Pipeline of incoherent signal fusion: from individual Range-Time (RT) plots to spectrograms fed directly into the Recurrent Neural Network.

The received radar echoes in fast-time provide the target's range, with the mainlobe typically associated with the target's position, and the sidelobes defining the noise floor, assuming sufficient SNR conditions. The summation of the complex RT matrices combining all radars as:

$$X(d,t) = \frac{1}{Nx} \sum_{nx=\#1}^{Nx} \chi(d,t)^{(nx)} \tag{1}$$

This generates $X(d,t)$ as a resulting RT matrix with information from all radar nodes indicated as $nx=\{\#1,\ldots,\#5\}$, as shown in Figure 2. This is then used to calculate a μD spectrogram to be used as input to the classifier.

A variety of Short-time Fourier transform (STFT) window sizes and step-widths were tested for the best performance between clutter suppression and clarity of limb motions. Clutter cancellation is performed by subtracting the average Doppler frequencies from the μD spectrogram, with satisfying classification achieved with STFT step-width of 10 samples (82 ms $\rightarrow$ t') and a Hanning window of 150 samples (1.23 s) [9]–[11].

The STFT is applied on the RT, X(d,t), as computed in [12], obtaining the μD spectrogram, Ψ(m,t'), containing the Doppler/velocity information of the target from all nodes,

where m refers to the spectrogram Doppler bins and t' indicates the slow-time bins, respectively. The proposed method uses directly the slow-time bins of the μD spectrogram as feature vectors. This approach is used in literature when spectrograms are fed directly into classifiers based on recurrent neural networks [13], and is an alternative to the extraction of features from sliding windows across the spectrograms.

III. Evaluation metrics

This section defines and discusses the proposed evaluation metrics to address the priorly mentioned challenges to classify a dataset with continuous, sequential activities.

TABLE I: Notation for metrics' definitions.

y, A / $\hat{y}$, $\hat{A}$	ground truth / predicted label / area
$\overline{y}$ / $\overline{\hat{y}}$	mean ground truth / mean predicted samples
s / $\hat{s}$	ground truth / predicted block (Figure 3)
$(\cdot)_p$, P	sample, set of samples
$(\cdot)^{(c)}$	class index (later neglected for readability)
tp	true positive rate
tn	true negative rate
fp	false positive rate
fn	false negative rate

A. Accuracy

To compare predictions and ground truth labels the identity function for each class c $I^{(c)}(\hat{y}_p, y_p)$ is introduced to measure false predictions:

$$I^{(c)}(\hat{y}_p, y_p) = \begin{cases} 0 \leftarrow \hat{y}_p = y_p \\ 1 \leftarrow \hat{y}_p \neq y_p \end{cases} \tag{2}$$

The number of misclassifications is provided by:

$$M^{(c)} = \frac{1}{P} \sum_{p=1}^{P} I^{(c)}(\hat{y_p}, y_p) \tag{3}$$

with the resulting accuracy being equal to:

$$A^{(c)} = 1 - M^{(c)} \tag{4}$$

Classical accuracy for evaluating classification performances does not capture inequalities of false negative (fn) and false positive (fp) and does not account for imbalanced datasets. This may lead to overlook drops in performance [14].

B. F_β score with precision, recall and specificity

The F_β score provides a more concise metric accounting for fp and fn imbalances, and consists of precision and recall. Together with precision and recall, the specificity is also computed as:

$$\begin{aligned} precision &= \frac{tp}{tp+fp}, \\ TPR = sensitivity = recall &= \frac{tp}{tp+fn}, \\ TNR = \text{specificity} &= \frac{tn}{tn+fp} \end{aligned} \tag{5}$$

Precision and recall are needed to compute the F_β score as:

$$F_\beta = (1+\beta^2) \times \frac{precision \times recall}{\beta^2 \cdot precision + recall} \tag{6}$$

with precision and recall evenly treated if $\beta = 1$, known as F_1 score. Otherwise, the formula favors precision if $\beta > 1$ [15].

C. Dice index

The Dice similarity index (also named as Sørensen-Dice coefficient) normalizes the length of the vector labels $\hat{y}$ and ground truth y and divides them by the total number of non-zero entries. The factor 2 multiplier scales the measurement range between $[0, 1]$ with 1 meaning label vectors identical to the ground truth [16]. It is expressed as:

$$Dice^{(c)} = 2 \times \frac{\left|\hat{A} \cap A\right|}{\left|\hat{A}\right| + |A|} = \frac{2tp}{2tp + fp + fn} \tag{7}$$

D. Jaccard index

The Jaccard index or Tanimoto coefficient defines the intersection divided by the union of two label vectors.

$$Jac^{(c)} = \frac{\left|\hat{A} \cap A\right|}{\left|\hat{A}\right| + |A| - \left|\hat{A} \cap A\right|} = \frac{tp}{tp + fp + fn} \tag{8}$$

It will be noted that the denominator determines the union as, $\left|\hat{A}\right| + |A| - \left|\hat{A} \cap A\right| = \left|\hat{A} \cup A\right|$. Furthermore, the Jaccard index is always smaller than the Dice index except at their extrema $[0, 1]$, with the relation between them as:

$$Jac^{(c)} = \frac{Dice^{(c)}}{2 - Dice^{(c)}} \tag{9}$$

and is described in [17].

E. Consecutive block detection (CBD)

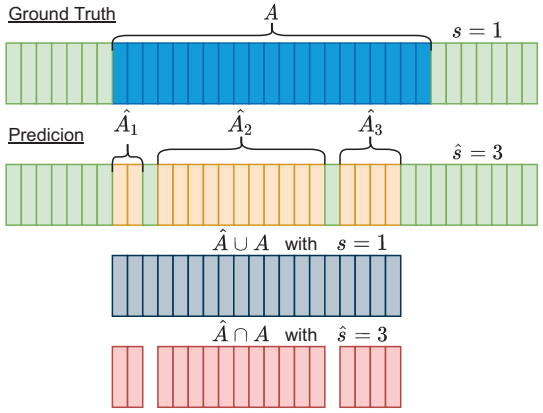


Fig. 3: The IoU with intersection and union sequences are demonstrated, as well as, the penalization term $H(\hat{s}, s) = 3/4$ used in Equation (16).

This proposed metric penalizes interruptions and fluctuations in the sequence of predicted samples, $\hat{y}_p$, with respect

to the corresponding ground truth labels, y_p. To the best of our knowledge, this aspect is not always well considered in the literature when evaluating radar-based HAR for continuous activities.

1) Unweighted consecutive block detection: Firstly, the individual ground truth blocks and the prediction blocks are counted as shown for the ground truth in Equation (10) and the predictions in Equation (11), respectively, as:

$$s\left(y_p\right) = \frac{1}{2}\sum_{p=2}^{P-1}\sqrt{\left(y_p - y_{p-1}\right)^2} \tag{10}$$

and

$$\hat{s}\left(\hat{y}_p\right) = \frac{1}{2}\sum_{p=2}^{P-1}\sqrt{\left(\hat{y}(\hat{y}|y_p=1)_p - \hat{y}(\hat{y}|y_p=1)_{p-1}\right)^2} \tag{11}$$

with the counter index, $(\cdot)_\mathrm{p}$, in the sequence of a total length, P. The ratio of blocks, as shown in Figure 3, is computed as:

$$Ed = \frac{s\left(y\right)}{\hat{s}\left(\hat{y}\right)} \tag{12}$$

with the range between $[0, 1]$, where 1 indicates the same number of blocks found within the ground truth sequence of a class and the prediction. It should be noted that block length differences are not considered in Equations (10) to (12), and this can in fact affect the result.

2) Weighted consecutive block detection: Due to the aforementioned effect of the block length differences on the metric, a corresponding penalty factor is computed as:

$$w = \sqrt{\frac{\left|\hat{A}\cap A\right|}{|A|}} \tag{13}$$

with the numerator indicating the intersection between the ground truth and the prediction over the ground truth, $|A|$. The non-linearity impact of the weight, $w = \sqrt{(\cdot)}$, is introduced to minimize penalization on small misalignments. The weighted consecutive block detection is finally computed by combining the Equations (12) and (13) as:

$$Ed_w = Ed \cdot w = \frac{s\left(y\right)}{\hat{s}\left(\hat{y}\right)} \cdot \sqrt{\frac{\left|\hat{A}\cap A\right|}{|A|}} \tag{14}$$

F. Intersection-Over-Union (IoU)

IoU is another metric that penalizes interruptions and fluctuations in the sequences of predictions. It is a known technique for evaluating camera-based object detection algorithms and is equivalent to the Jaccard index (under certain conditions). This method defines the similarity on the bounding boxes [18], which are generally uninterrupted entities in vision-based detection methods.

A modified algorithm can account for interruptions in labels such as:

$$H(\hat{s}, s) = 1 - \left(\frac{2\cdot\hat{s}}{\hat{s}+s} - 1\right)^2 \tag{15}$$

with s and $\hat{s}$, computed by using Equations (10) and (11), the concatenated sequence blocks for ground truth and predictions, respectively, such as:

$$\begin{aligned}\mathrm{IoU}^{(\mathrm{c})} &= Jac \cdot H(\hat{s}, s)\\ &= \left(\frac{\hat{A}\cap A}{\hat{A}\cup A}\right)\cdot\left(1 - \left(\frac{2\cdot\hat{s}}{\hat{s}+s} - 1\right)^2\right)\end{aligned} \tag{16}$$

Equation (16) penalizes interrupted sequences, even if the predictions are broadly corresponding and aligned with the ground truth [19].

G. Correlation index or Matthews Correlation Coefficient (MCC)

The correlation index or Matthews Correlation Coefficient (MCC) provides a rather uncommon evaluation of sequences. The method calculates the Pearson's Linear Correlation Coefficient, typically used to find linear similarities between vectors. This can also be used for sequence classification as:

$$R(\hat{y}, y) = \frac{\sum_{p=1}^{P}\left(y_p - \bar{y}\right)\left(\hat{y}_p - \overline{\hat{y}}\right)}{\sqrt{\sum_{p=1}^{P}\left(y_p - \bar{y}\right)^2 \sum_{p=1}^{P}\left(\hat{y}_p - \overline{\hat{y}}\right)^2}},\ R\in\mathbb{R};[-1,1] \tag{17}$$

with $\overline{\hat{y}}$ and $\overline{y}$ the means of the ground truth and prediction vector, respectively. Alternatively, the equation can be expressed as:

$$R = \frac{tp\cdot tn + fp\cdot fn}{\sqrt{(tp+fp)\cdot(tp+fn)\cdot(tn+fp)\cdot(tn+fn)}} \tag{18}$$

and is known as Matthews Correlation Coefficient (MCC) [20]. It should be noted that $R(\hat{y}, y)=-1$ is equivalent to perfectly misclassified sequences, and $R(\hat{y}, y)=0$ is the expected value from an unbiased "coin tossing classifier" for a balanced dataset.

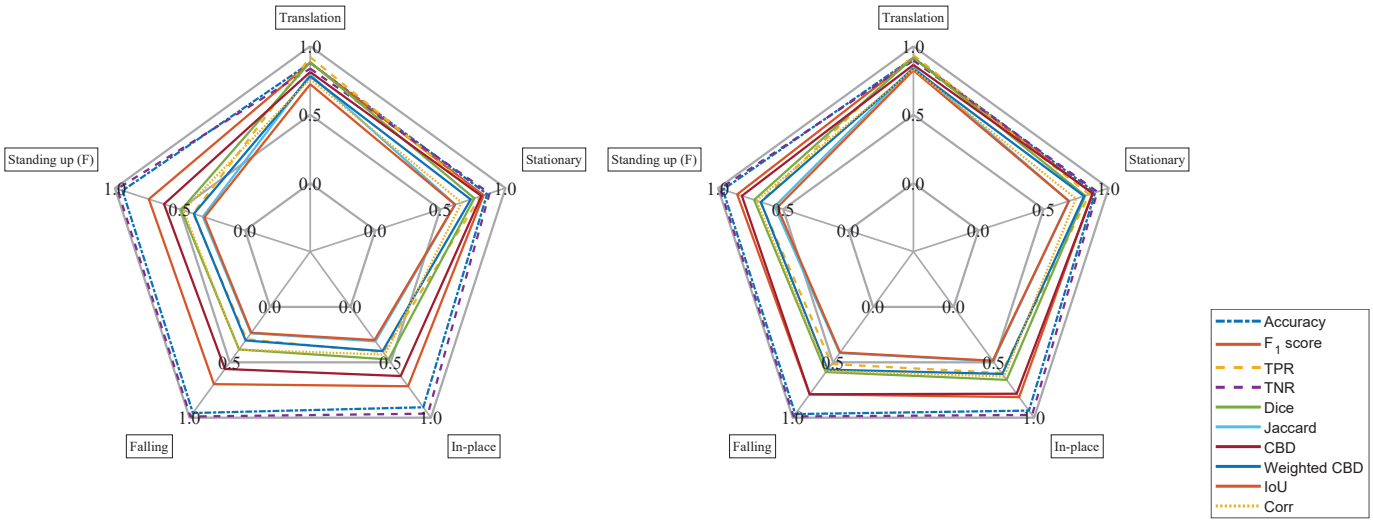
IV. Case Study with Experimental Data

To evaluate the proposed metrics, 3 different RNN are used as classifiers: the Gated Recurrent Units (GRU), the Long Short-Term Memory (LSTM), and the Bidirectional LSTM (Bi-LSTM), with results reported in Table II. The classification performance of a single radar is compared to the proposed incoherent signal fusion method, shown in Figure 4.

A. Discussion

The objective is not to identify the most suitable metric for continuous sequences of activities, but to discuss the applicability and strengths or weaknesses for each metric. Specific results for individual classes are only mentioned for the case of incoherent signal fusion using Bi-LSTM, as the best performing RNN from the results presented in this paper.

1) Accuracy: Accuracy appears to be very high for all considered classifier (macro accuracy of $91\,\%$). However, it should be noted that accuracy alone does not capture the true performances due to dataset imbalances, especially for classes with few samples such as *falling*.



(a) Average single radar Bi-LSTM (b) Signal fusion Bi-LSTM (c) Legend

Fig. 4: Spider diagrams with the performance of the 10 proposed evaluation metrics for the classifier Bi-LSTM using signal fusion in Figure 4b and with single radar classification in Figure 4a.

TABLE II: The proposed metrics for signal fusion used RNNs vs. single radar classification with Bi-LSTM are compared, with the mean and standard deviation across all networks.

	Accuracy	F1 score	TPR	TNR	Dice	Jaccard	CBD	Weighted CBD	IoU	Corr.
Signal fusion GRU	0.914	0.784	0.589	0.934	0.608	0.477	0.721	0.557	0.400	0.560
Signal fusion LSTM	0.889	0.741	0.555	0.919	0.560	0.421	0.688	0.516	0.360	0.481
Signal fusion Bi-LSTM	0.936	0.849	0.715	0.952	0.741	0.603	0.827	0.698	0.592	0.698
Single radar average Bi-LSTM	0.909	0.773	0.566	0.930	0.599	0.456	0.687	0.521	0.436	0.543
Mean across all RNN	0.912	0.787	0.606	0.934	0.627	0.489	0.731	0.573	0.447	0.570
Standard deviation across all RNN	0.019	0.046	0.074	0.014	0.079	0.079	0.066	0.085	0.102	0.092

2) F_β score, TPR, TNR: Evaluating TPR (sensitivity or recall), precision, and TNR (specificity) on their own is less effective than using the F_β score, as this can provide a better global view on performances for each specific class. An average of the F_β score across all classes, macro F_β score, is also possible. For this case study, the performance differences between individual classes increase to approximately $12\,\%$ for signal fusion using Bi-LSTM, specifically referring to the *translation* ($91.6\,\%$) and *standing up from falling* ($78.9\,\%$) activity, as shown in Figure 4b.

3) Dice index: The Dice index is a harder metric than the prior shown metrics of accuracy or F_β score. Here, for example *standing up from falling* degrades to $58.8\,\%$ (F1 score: $78.9\,\%$) and *translation* to $91.7\,\%$ (F1 score: $91.6\,\%$).

4) Jaccard index: The Jaccard index has a linear relationship with the Dice index as shown in Equation (9), with performance always lower than the Dice index except at their extrema. In fact, with this metric even lower performances are reported for certain classes, e.g., *standing up from falling* degrades to $41.6\,\%$ (Dice index: $58.8\,\%$).

5) Consecutive block detection (CBD): The CBD operates differently than the previously shown metrics. Here, interruptions of prediction label blocks have an impact. For example, *stationary* and *in-place* activities provide the best and worst classes for incoherent fusion with Bi-LSTM classification (see Figure 4b) with $88.0\,\%$ and $78.3\,\%$, respectively.

However, CBD accounts for the number of detected blocks only. The weighted CBD considers also the detection length of the predictions versus the ground truth labels. Specifically (see Figure 4b), the best and worst class become *translation* and *falling* activity with $83.5\,\%$ and $56.6\,\%$, respectively.

6) Intersection-Over-Union (IoU): The IoU metric is the second metric accounting for detected blocks within the prediction vector. The IoU is the most extreme evaluation metric for our dataset since it is a product of the Jaccard index (hard metric on its own) multiplied with a block detection term [19]. The activity *standing up from falling* degrades to $41.1\,\%$ (Jaccard index: $41.6\,\%$), and the *translation* activity to $82.4\,\%$ (Jaccard index: $84.7\,\%$) as the best class.

7) Correlation index or Matthews Correlation Coefficient (MCC): This metric is rather challenging to compare with the previously introduced metrics. In contrast, the strength of this metric is a distinct indication if classifiers provide outputs resulting in $R(\hat{y}, y) < 0$. Such results are immediately an indication of a mismatch between the ground truth and prediction samples. The activities *falling* and *translation* activities provide the worst and best results with $57.7\,\%$ and $83.1\,\%$, respectively.

B. Discussion overview

As previously mentioned, the scope of this paper is not to find the most suitable evaluation metric, but to identify the pros and cons of each metric concerning the characteristics of the dataset and the overall classification objective of the HAR system. Some considerations from this initial analysis follow:

1) For equally-distributed (balanced) data evaluation: The conventional accuracy metric can provide satisfactory results, even if it does not describe where mistakes (e.g., missed detections or false alarms) occur for a class. For that, F_β score are more suitable by accounting for precision and recall.

2) For skewed (imbalanced) data evaluation: The F_β score becomes a more suitable metric than plain accuracy, and is widely used. This is very important as accuracy can significantly overestimate performances, as seen in our case study. The same applies to Dice and Jaccard coefficients/indexes. Also, the correlation index or Matthews Correlation Coefficient (MCC) accounts for imbalances in the dataset as it is widely used in the medical domain, as shown in [19].

3) For evaluating continuous sequences of activities: The prior metrics suffer by evaluating continuous sequences of activities with random and seamless transitions between them. With their modification as weighted CBD and the IoU, the proposed CBD is preferable for such cases. These metrics can account for outliers (i.e., fluctuations and interruptions) in the prediction label vector and are well suited for HAR of continuous sequences. These metrics are particularly suited for RNN classifiers for HAR as they can directly process their sequential output predictions and penalize fluctuations/interruptions that could propagate errors within the networks' memory cells.

V. Conclusion

This paper presents and compares state-of-the-art and proposed evaluation metrics for radar-based HAR of continuous sequences of human activities. The metrics' pros and cons are discussed, referring to an experimental dataset collected with a network of 5 distributed UWB radars and including 15 participants. Notably, the collected sequences were performed in random locations and with arbitrary and unconstrained trajectories and aspect angles to the radar sensors. Data from 5 radars were combined with incoherent signal level fusion to generate a combined μD spectrogram fed to RNN classifiers, namely, GRU, LSTM, and Bi-LSTM.

The paper demonstrates the need for metrics other than plain accuracy or precision/recall when evaluating continuous HAR, especially by using recurrent classification to estimate classifier performance at a fine scale. Specifically, evaluation metrics that account for outliers in the prediction vector (i.e., misalignments, interruptions, and fluctuations), the *Weighted CBD* and *IoU* are more sensitive than conventional Accuracy evaluations: while IoU shows almost $20\,\%$ difference between poor and good performing classifiers, conventional Accuracy evaluation gives only $2\,\%$ difference, i.e., very coarse assessment. Regarding the used dataset, classifiers with bidirectionality provide superior classification, principally for a imbalanced dataset as a use case.

References

[1] S. A. Shah and F. Fioranelli, "RF sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 11 2019.

[2] M. G. Amin and R. G. Guendel, "Radar classifications of consecutive and contiguous human gross-motor activities," *IET Radar, Sonar and Navigation*, vol. 14, no. 9, pp. 1417–1429, 09 2020.

[3] U. M. Khan, Z. Kabir, S. A. Hassan, and S. H. Ahmed, "A deep learning framework using passive WiFi sensing for respiration monitoring," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[4] 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 and Navigation*, vol. 9, no. 9, pp. 1231–1239, 12 2015.

[5] M. Wang, Y. D. Zhang, and G. Cui, "Human motion recognition exploiting radar with stacked recurrent neural network," *Digital Signal Processing: A Review Journal*, vol. 87, pp. 125–131, 2019.

[6] C. Ding, H. Hong, Y. Zou, H. Chu, X. Zhu, F. Fioranelli, J. Le Kernec, and C. Li, "Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6821–6831, 2019.

[7] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7590–7603, 2021.

[8] R. G. 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," Nov 2021. [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/16691500/2

[9] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Continuous human activity recognition for arbitrary directions with distributed radars," in *2021 IEEE Radar Conference (RadarConf21)*, 5 2021, p. 6.

[10] A. Petroff, "A practical, high performance ultra-wideband radar platform," in *2012 IEEE Radar Conference*, 2012, pp. 0880–0884.

[11] Y. He, P. Molchanov, T. Sakamoto, P. Aubry, F. Le Chevalier, and A. Yarovoy, "Range-Doppler surface: a tool to analyse human target in ultra-wideband radar," *IET Radar, Sonar Navigation*, vol. 9, no. 9, pp. 1240–1250, 2015.

[12] R. G. Guendel, F. Fioranelli, and A. Yarovoy, "Phase-based classification for arm gesture and gross-motor activities using histogram of oriented gradients," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7918–7927, 2021.

[13] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2020.

[14] J. Watt, R. Borhani, and A. K. Katsaggelos, *Machine Learning Refined: Foundations, Algorithms, and Applications*, 2nd ed. Cambridge University Press, 2020.

[15] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*, A. Sattar and B.-h. Kang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1015–1021.

[16] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999.

[17] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.

[18] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2 2019.

[19] P. Chen, X. Wang, M. Wang, X. Yang, S. Guo, C. Jiang, G. Cui, and L. Kong, "Multi-view real-time human motion recognition based on ensemble learning," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20 335–20 347, 2021.

[20] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.