

One Size Does Not Fit All: Multi-Scale, Cascaded RNNs for Radar Classification

Dhrubojoyti Roy*
Sangeeta Srivastava*
roy.174@osu.edu
srivastava.206@osu.edu
The Ohio State University

Aditya Kusupati
kusupati@cs.washington.edu
University of Washington
Microsoft Research India

Pranshu Jain
anz178419@cse.iitd.ac.in
Indian Institute of Technology Delhi

Manik Varma
manik@microsoft.com
Microsoft Research India
Indian Institute of Technology Delhi

Anish Arora
arora.9@osu.edu
The Ohio State University
The Samraksh Company

ABSTRACT

Edge sensing with micro-power pulse-Doppler radars is an emergent domain in monitoring and surveillance with several smart city applications. Existing solutions for the clutter versus multi-source radar classification task are limited in terms of either accuracy or efficiency, and in some cases, struggle with a trade-off between false alarms and recall of sources. We find that this problem can be resolved by learning the classifier across multiple time-scales. We propose a multi-scale, cascaded recurrent neural network architecture, MSC-RNN, comprised of an efficient multi-instance learning (MIL) Recurrent Neural Network (RNN) for clutter discrimination at a lower tier, and a more complex RNN classifier for source classification at the upper tier. By controlling the invocation of the upper RNN with the help of the lower tier conditionally, MSC-RNN achieves an overall accuracy of 0.972. Our approach holistically improves the accuracy and per-class recalls over machine learning models suitable for radar inferencing. Notably, we outperform cross-domain handcrafted feature engineering with purely time-domain deep feature learning, while also being up to ~3x more efficient than a competitive solution.

CCS CONCEPTS

• Computing methodologies → Neural networks; Supervised learning by classification; • Computer systems organization → Sensor networks; Real-time system architecture; System on a chip.

KEYWORDS

Radar classification, recurrent neural network, low power, edge sensing, range, joint optimization, real-time embedded systems

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BuildSys '19, November 13–14, 2019, New York, NY, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-7005-9/19/11...\$15.00

<https://doi.org/10.1145/3360322.3360860>

ACM Reference Format:

Dhrubojoyti Roy, Sangeeta Srivastava, Aditya Kusupati, Pranshu Jain, Manik Varma, and Anish Arora. 2019. One Size Does Not Fit All: Multi-Scale, Cascaded RNNs for Radar Classification. In *BuildSys '19: The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, November 13–14, 2019, New York, NY, USA*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3360322.3360860>

1 INTRODUCTION

With the rapid growth in deployments of Internet of Things (IoT) sensors in smart cities, the need and opportunity for computing increasingly sophisticated sensing inferences on the edge has also grown. This has motivated several advances in designing resource-efficient sensor inferences, particularly those based on machine learning and especially deep learning. The designs, however, encounter a basic tension between achieving efficiency while preserving predictive performance that motivates a reconsideration of state-of-the-art techniques.

In this paper, we consider a canonical inference pattern, namely discriminating clutter from several types of sources, in the context of a radar sensor. This sort of $N+1$ -class classification problem, where N is the number of source types, has a variety of smart city applications, where diverse clutter is the norm. These include triggering streetlights smartly, monitoring active transportation users (pedestrians, cyclists, and scooters), crowd counting, assistive technology for safety, and property surveillance. As an example, streetlights should be smartly triggered on for pedestrians but not for environmental clutter such as trees moving in the wind. Similarly, property owners should be notified only upon legitimate intrusions but not for passing animals.

The radar is well-suited in the smart city context as it is privacy preserving in contrast to cameras. Moreover, it consumes low power (~15mW), because of which it can be deployed at operationally relevant sites with little dependence on infrastructure, using, for instance, a small panel solar harvester or even a modest sized battery, as shown in Figure 1. Experiences with deploying sensors in visionary smart city projects such as Chicago's Array of Things [6, 38] and Sounds of New York City [5] have shown that wired deployments on poles tend to be slow and costly, given constraints of pole access rights, agency coordination, and labor unions, and can sometimes be in suboptimal locations. Using a

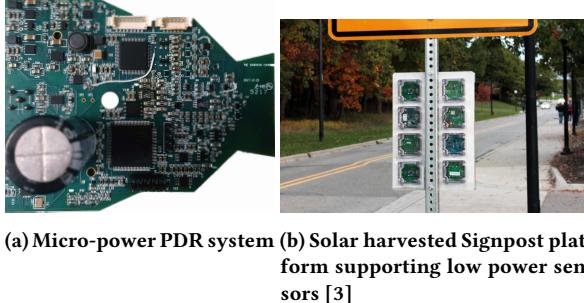


Figure 1: The micro-power pulse-Doppler radar (PDR) device can be independently deployed or interfaced with existing multi-sensor smart city platforms such as Signpost (figure adapted from [3], Copyright ©2019 ACM, Inc.)

low-power sensor that is embedded wirelessly or simply plugged in to existing platforms while imposing only a nominal power cost simplifies smart city deployment.

Table 1 illustrates an efficiency-accuracy trade-off for the canonical inference pattern with $N = 2$, wherein clutter is distinguished from human and other (i.e., non-human) sources. The more accurate deep models, the Convolutional Neural Network (CNN) [23] and the Long Short-Term Memory (LSTM) [17], that we machine-learned for this 3-class classifier from a reference dataset are significantly less efficient, in terms of speed and therefore power consumption. In contrast, the more efficient shallow solution, Support Vector Machine (SVM), is significantly less accurate. While the SVM classifier has been implemented to operate in near real-time on the Cortex-M3 single-microcontroller processor in the device depicted in Fig. 1(a), neither the CNN nor the LSTM per se yield a near real-time implementation. To implement deep models in near real-time on the M3, we therefore consider model optimization with recent state-of-art-techniques such as fast gated RNNs (FastGRNN) [26] and Early-exit Multi-Instance RNNs (EMI-LSTM and EMI-FastGRNN) [10]. However, Table 1 illustrates that the trade-off remains: the best accuracy we achieve, namely with FastGRNN, has significantly lower efficiency than the best efficiency achieved, namely with EMI-FastGRNN, but the latter’s accuracy is comparatively worse.

Problem Statement. In this work, we investigate alternative optimizations of deep models for the above classification task that achieve both high accuracy and speed. In doing so, we do not wish to sacrifice the recall performance for achieving high precision. For instance, radar sensing applications require that the clutter recall be very high so that there are minimal false alarms. However, a solution that restricts false alarms at the cost of detectability (i.e., low source recall, where a source could be either human or non-human) would be undesirable as it would have limited applicability in the smart city contexts discussed above.

Solution Overview. The $N + 1$ -class radar problem, where the $+1$ -class is clutter, conflates discrimination between classes that are conceptually different. In other words, discriminating clutter from sources has a different complexity from that of disambiguating source types. This insight generalizes when the sources themselves are related by a hierarchical ontology, wherein different levels of

Table 1: Trade-offs in accuracy and runtime efficiency for the 3-class radar problem (window length 1s, feature computation overhead ignored for SVM, dataset and machine architecture details are in Section 5)

ML Model	Accuracy	FLOPS	Real-time?
SVM (15 features)	0.85	37K	Yes
LSTM	0.89	100K	No
CNN (1s FFT)	0.91	1.3M	No
EMI-LSTM	0.90	20K	Yes
FastGRNN	0.96	35K	Yes
EMI-FastGRNN	0.88	8K	Yes

source types involve concepts of correspondingly different complexity of discrimination. By way of example, in the 3-class clutter vs. human vs. non-human classification problem, discriminating clutter from sources turns out to be simpler than discriminating the more subtle differences between the source types. Using the same machine architecture for 3 classes of discrimination leads to the accuracy-efficiency trade-off, as the last two rows of Table 1 indicate. A more complex architecture suffices for discriminating among source types accurately, whereas a simpler architecture more efficiently suffices for discriminating clutter from sources, but hurts the accuracy of discriminating between source types.

We, therefore, address the problem at hand with an architecture that decomposes the classification inference into different hierarchical sub-problems. For the 3-class problem, these are: (a) Clutter vs Sources, and (b) Humans vs. Non-humans *given* Sources. For each sub-problems we choose an appropriate learning architecture; given the results of Table 1, both architectures are forms of RNN albeit *with learning at different time-scales*. The lower tier RNN for (a) uses a short time-scale RNN, the Early-exit Multi-Instance RNN (EMI-FastGRNN) [10, 26], whereas the higher tier for (b) uses a longer time-scale RNN, a FastGRNN [26], which operates at the level of windows (contiguous, fixed-length snippets extracted from the time-series) as opposed to short instances within the window. The upper tier uses the features created by the lower tier as its input; for loss minimization, both tiers are jointly trained. To further improve the efficiency, we observe that source type discrimination needs to occur only when a source is detected and clutter may be the norm in several application contexts. Hence, the less efficient classifier for (b) is invoked only when (a) discriminates a source: we refer to this as cascading between tiers. The joint training loss function is refined to emulate this cascading. We call this architecture *Multi-Scale, Cascaded RNN (MSC-RNN)*.

Contributions. Our proposed architecture exploits conditional inferencing at multiple time-scales to jointly achieve superior sensing and runtime efficiency over state-of-the-art alternatives. To the best of our knowledge, this approach is novel to deep radar systems. For the particular case of the 3-class problem, MSC-RNN performs as follows on the Cortex-M3:

Accuracy	Clutter Recall	Human Recall	Non-human Recall	FLOPS
0.972	1	0.92	0.967	9K

Its accuracy and per-class recalls are mostly better than, and in remaining cases competitive with, the models in Table 1. Likewise, its efficiency is competitive with that of EMI-FastGRNN, the most efficient of all models, while substantially outperforming it in terms of sensing quality. We also validate that this MSC-RNN solution is superior to its shallow counterparts not only comprehensively, but at each individual tier as well. The data and training code for this project are open-sourced at [33, 35].

Other salient findings from our work are summarized as follows:

- (1) Even with deep feature learning purely in the time-domain, MSC-RNN surprisingly outperforms handcrafted feature engineering in the amplitude, time, and spectral domains for the source separation sub-problem. Further, this is achieved with 1.75-3× improvement in the featurization overhead.
- (2) The Tier 1 component of MSC-RNN, which classifies legitimate sources from clutter, improves detectability by up to 2.6× compared to popular background rejection mechanisms in radar literature, even when the false alarm rate is controlled to be ultra-low.
- (3) MSC-RNN seems to tolerate the data imbalance among its source types better than compared EMI-RNN models. In particular, it enhances the non-dominant human recall by up to 20%, while simultaneously maintaining or improving the dominant non-human recall and overall accuracy.

Organization. In Section 2, we present related research and outline the basics of micro-power radar sensing in Section 3. In Section 4, we detail the various components in our solution and discuss the training and inference pipelines. We provide evaluation and prototype implementation details in Sections 5 and 6 respectively. We conclude and motivate future research in Section 7.

2 RELATED WORK

Shallow and Deep Radar Sensing. Micro-Doppler features have been used in myriad radar sensing applications ranging from classification [15, 20, 27] to regression [14]. Most of these applications employ the short-time Fourier transform (STFT) input representation for analyzing micro-Doppler signatures. Although shallow classifiers can be computationally cheaper than deep solutions, the spectrogram generation over a sliding window for the STFT incurs significant computational overhead for real-time applications on single microcontroller devices. In order to decrease the computational overhead, different feature extraction methods [18, 30] have been investigated in the past. Notably, feature engineering not only requires sophisticated domain knowledge, but also may not transfer well to solutions for other research problems. Moreover, selection of relevant and non-redundant features requires care for sensing to be robust [34]. In recent years, there has been significant use of deep learning in radar applications. Most works with architectures like CNNs and autoencoders use spectrogram-based input [19, 21]. The authors of [28] generate a unique spectral correlation function for the Deep Belief Network to learn signatures from. The pre-processing needed in these applications and the resulting model sizes make them unsuitable for single microcontroller devices. Therefore, we use raw time-series data in conjunction with RNN variants to make our deep learning solution faster and

more efficient in addition to avoiding spectrogram computation altogether.

Efficient RNN. The ability of RNNs in learning temporal features makes them ubiquitous in various sequence modeling tasks. They, however, often suffer from training instabilities due to the exploding and vanishing gradient problem (EVGP) [31]. Gated RNNs like LSTMs [17] and GRUs [8] circumvent EVGP and achieve the desired accuracy for a given task, but compromise on model sizes and compute overheads, making them unattractive for real-time, single microcontroller implementations. Recently, FastGRNN [26] has been proposed to achieve prediction accuracies comparable to gated RNNs while ensuring that the learned models are significantly smaller for diverse tasks. The hierarchical classifier solution we present in this paper is based on this architecture.

Multi-Instance Learning and Early Classification. MIL is a weakly supervised learning technique that is used to label sub-instances of a window. MIL has found use in applications from vision [40] to natural language processing (NLP) [22]. It enables a reduction in the computational overhead of sequential models like RNNs by localizing the appropriate activity signature in a given noisy and coarsely-labeled time-series data along with early detection or rejection of a signal [10]. We use it as our lower-tier classifier for clutter versus source discrimination.

Multi-Scale RNN. One of the early attempts to learn structure in temporally-extended sequences involved using reduced temporal sequences [16] to make detectability over long temporal intervals feasible in recurrent networks [29, 36]. With the resurgence of RNNs, multi-scale recurrent models are being used to discover the latent hierarchical multi-scale structure of sequences [9]. While they have been traditionally used to capture long-term dependencies, we use them to design a computationally efficient system by incorporating different scales of temporal windows for the lower- and upper-tier RNNs. By conditioning the Tier 2 classifier, which works on longer windows and is hence bulkier, we make sure that the former is invoked only when necessary, i.e., when Tier 1 predicts a source.

Compression Techniques. Sparsity, low-rank, and quantization have been proven to be effective ways of compressing deep architectures like RNNs [39, 41] and CNNs [13, 25]. We incorporate low-rank representation, Q15 quantization, and piecewise-linear approximation [26] to make MSC-RNN realizable on Cortex-M3 microcontrollers.

3 RADAR AND CLASSIFIER MODELS

3.1 Micro-power Radar Model

The monostatic PDR sensor depicted in Figure 1 has a bandwidth of nearly 100 MHz and a center frequency at about 5.8 GHz. It is a short-range radar with an anisotropic radiation pattern yielding a maximum detection range of ~13 m. Sensing itself consumes 15mW of power, not counting the inference computation on the associated microcontroller. The radar response is low pass filtered to 100 Hz; hence the output is typically sampled at rates over 200Hz.

The output signal from the radar is a complex time-series with In-phase (I) and Quadrature (Q) components. When a source moves within the detection range, in addition to the change in received

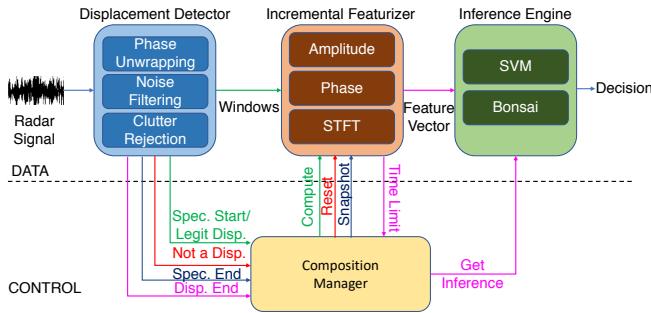


Figure 2: SVM classifier data and control planes; control signal-response pairs are color coded

power, the phase of this signal changes according to the direction of motion. Consequently, its relative displacement can be estimated with high accuracy (typically, sub-cm scale), and a rich set of features can be derived by tracking its phase evolution over time.

3.2 Classifier Architectures

3.2.1 Input and Feature Representation. The radar classifier system uses the aforementioned complex time-series as input. Extant end-to-end architectures for micro-power radar sensing mostly eschew deep feature learning for cheap handcrafted feature engineering in the amplitude, time, and spectral domains [14, 34]. However, these solutions incur significant featurization overhead; this is exemplified in Table 2 on 1-second snippets extracted from the complex time-series. Even ignoring the SVM computation latency, it can be seen that the main computation bottleneck is this incremental overhead which results in >30% duty cycle on the Cortex-M3, of which ~10% constitutes the FFT overhead alone.

Table 2: Computation overheads in a shallow (SVM) radar solution on Cortex-M3 (10 features, 1s windows)

Component	Latency (ms)
FFT	80
Incremental feature computation	212
SVM inference (700 SVs)	55

3.2.2 Shallow Classifier Architecture. As shown in Figure 2, a prototypical shallow radar classifier system consists of three subsystems: (i) a *displacement detector* for discriminating clutter vs. sources, (ii) an *incremental featurizer*, (iii) an *end inference engine* that discriminates source types, and (iv) a *composition manager* that handles their interactions. The displacement detector thresholds unwrapped phase over incoming windows of radar data ($\frac{1}{2}$ s or 1s) to detect legitimate source displacements in the scene, filtering in-situ clutter that tends to yield self-canceling phase unwraps. When a source displacement is speculatively detected, the featurizer is invoked till the current displacement ends or a pre-specified time limit is reached. The final feature vector is fed to an end classifier such as SVM [34]. Note that incremental feature computation overhead is the primary impediment in realizing efficiency in these systems,

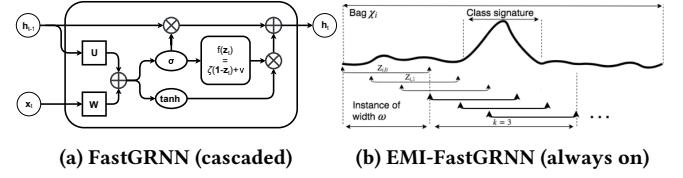


Figure 3: FastGRNN & EMI-FastGRNN (images from [10, 26])

hence techniques like replacing the heavy SVM classifier with the much lighter Bonsai [24], or observing longer displacements to run inference infrequently do not alleviate this problem.

3.2.3 Deep Classifier Architecture. In the interest of designing resource efficient solutions, in this work, we use purely time-domain learning instead of all-domain feature engineering. While we preserve the aforementioned classifier hierarchy in our solution, we replace the simple “ensemble” with a principled 2-tier RNN approach. In the next sections, we present our proposed architecture and discuss how our approaches to deep feature learning can be used to successfully resolve the above issues.

4 2-TIER DEEP CLASSIFIER ARCHITECTURE

MSC-RNN is a multi-scale, cascaded architecture that uses EMI-FastGRNN as the lower-tier clutter discriminator and FastGRNN as the upper-tier source classifier. While EMI-FastGRNN efficiently localizes the source signature in a clutter prone time-series ensuring smaller sequential inputs along with early classification, FastGRNN reduces the per-step computational overhead over much heavier alternatives such as LSTM. We begin with the relevant background for each of these components.

4.1 Candidate Classifiers

FastGRNN. FastRNN [26] provably stabilizes RNN training by helping to avoid EVGP by using only two additional scalars over the traditional RNN. FastGRNN is built over FastRNN and it extends the scalars of FastRNNs to vector gates while maximizing the computation reuse. FastGRNN also ensures its parameter matrices are low-rank, sparse and byte quantized to ensure very small models and very fast computation. FastGRNN is shown to match the accuracies of state-of-the-art RNNs (LSTM and GRU) across various tasks like keyword spotting, activity recognition, sentiment analysis, and language modeling while being up to 45x faster.

Let $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$ be the input time-series, where $x_t \in \mathbb{R}^D$. The traditional RNN’s hidden vector $\mathbf{h}_t \in \mathbb{R}^{\hat{D}}$ captures long-term dependencies of the input sequence: $\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b})$. Typically, learning \mathbf{U} and \mathbf{W} difficult due to the gradient instability. FastGRNN (Figure 3(a)) uses a scalar controlled peephole connection for every coordinate of \mathbf{h}_t :

$$\begin{aligned} \mathbf{h}_t &= (\zeta(1 - \mathbf{z}_t) + v) \odot \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}_h) + \mathbf{z}_t \odot \mathbf{h}_{t-1}, \\ \mathbf{z}_t &= \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}_z) \end{aligned}$$

Here, $0 \leq \zeta, v \leq 1$ are trainable parameters, and \odot represents the vector Hadamard product.

EMI-RNN. Time-series signals when annotated are rarely precise and often coarsely labeled due to various factors like human errors and smaller time frames of activities themselves. EMI-RNN [10] tackles the problem of signal localization using MIL, by splitting the i^{th} data window into instances $\{Z_{i,\tau}\}_{\tau=1,\dots,T-\omega+1}$ of a fixed width ω (Figure 3(b)). The algorithm alternates between training the classifier and re-labeling the data based on the learned classifier until convergence. A simple thresholding scheme is applied to refine the instances: in each iteration, k consecutive instances are found with maximum prediction sum for the class label. Only these instances are included in the training set for the next iteration. Here, k is a hyperparameter that intuitively represents the number of instances expected to cover the source signature. In the end, EMI-RNN produces precise signal signatures which are much smaller than the raw input, thus reducing the computation and memory overhead over the traditional sequential techniques. EMI-RNN also ensures early detection of noise or keywords thereby removing the need of going through the entire signal before making a decision. When combined, EMI-FastGRNN provides very small models along with very fast inference for time-series classification tasks. Codes for FastGRNN [26] & EMI-RNN [10] are part of Microsoft Research India's EdgeML repository [11].

4.2 MSC-RNN Design

While EMI-RNN is by itself equipped to handle multi-class classification efficiently, we find its accuracy and non-dominant source recall to be sub-optimal for the radar time-series, especially at smaller hidden dimensions and shorter window lengths. FastGRNN, on the other hand, is a relatively heavier solution to be used as a continuously running 3-class discriminator. To redress this trade-off, we make the following observations:

- (i) clutter, which yields self-canceling phase, can be rejected at a relatively shorter time-scale,
- (ii) disambiguating source types from their complex returns is a harder problem requiring a potentially longer window of observation, and
- (iii) the common case in a realistic deployment constitutes clutter; legitimate displacements are relatively few.

MSC-RNN, therefore, handles the two sub-problems at different time-scales of featurization (see Figure 4): the lower tier, an EMI-FastGRNN, discriminates sources from clutter at the level of short instances, while the upper one, a windowed FastGRNN, discriminates source types at the level of longer windows. Further, the upper tier is invoked *only* when a source is discriminated by the lower tier and operates on the instance-level embeddings generated by the latter.

4.2.1 Joint Training and Inference. The training of the lower tier inherits from that of EMI-training. We recap its training algorithm [10], which occurs in two phases, the MI phase and the EMI phase. In the MI phase, where the source boundaries are refined in a clutter-prone window, the following objective function is optimized:

$$\min_{f_l, s_l} \frac{1}{n} \sum_{i,\tau} \mathbb{1}_{\tau \in [s_i, s_i+k]} \ell(f_l(Z_{i,\tau}), y_i)$$

Here, ℓ represents the loss function of FastGRNN, and the classifier f_l is based on the final time-step in an instance. In the EMI

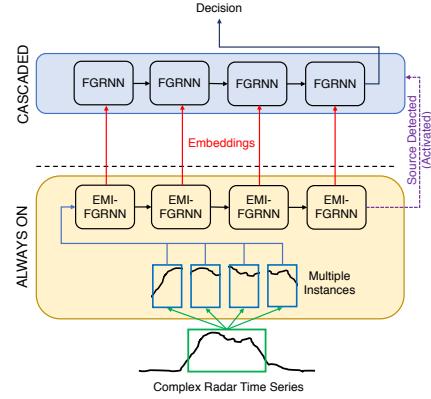


Figure 4: MSC-RNN architecture – the lower EMI-FastGRNN runs continuously, while the higher FastGRNN is invoked only for legitimate displacements

phase, which incorporates the early stopping, the loss \mathcal{L}_{EMI} is obtained by replacing the previous loss function with the sum of the classifier loss at every step: $\min \sum_{t=1}^T \ell(w^T o_{i,t})$, where w is the fully connected layer and $o_{i,t}$ the output at step t . The overall training proceeds in several rounds, where the switch to the EMI loss function is typically made halfway in.

For training the upper tier, in keeping with the divide-and-conquer paradigm of MSC-RNN, the Tier 2 FastGRNN cell should only learn to separate the source types, while ignoring instances of training data that are clutter. Therefore, we devise a conditional training strategy that captures the cascading behavior. To achieve this, the standard cross-entropy loss function of the upper tier is modified as:

$$\min_{f_u} \frac{1}{n} \sum_i \mathbb{1}_{y_i \neq -1} \ell(f_u(\mathcal{E}(\{Z_{i,\tau}^{tr}\})), y_i)$$

where f_u represents the upper classifier, and $\mathcal{E} : \mathbb{R}^{(T-\omega+1) \times \omega \times F} \rightarrow \mathbb{R}^{(T-\omega+1) \times H_l}$ represents the instance-level embedding vector from EMI-RNN with a hidden dimension of H_l (here, F represents the feature dimension for the radar time-series). Intuitively, this means that the upper loss is unaffected by clutter points, and thus the tiers can be kept separate.

The training algorithm for MSC-RNN is outlined in Algorithm 1. The two tiers are first separately initialized using their respective loss functions, and in the final phase, both are jointly trained to minimize the sum of their losses. Inference is simple: the instance-level EMI-RNN stops early with a decision of “Source” when a probability threshold \hat{p} is crossed; $\geq k$ consecutive positives constitute a discrimination for which the cascade is activated.

5 COMPARATIVE & TIER-WISE EVALUATION

5.1 Datasets

Table 3(a) lists the radar source and clutter datasets collected in various indoor and outdoor environments, which are used in this work. Some of these locations are documented in Figure 5; small or crammed indoor spaces such as office cubicles have been avoided



Figure 5: Some locations where source and clutter data was collected for experiments

Algorithm 1: MSC-RNN training algorithm

Input: Multi-instance training data
 $\{\{Z_{i,\tau}^{tr}\}_{\tau=1,\dots,T-\omega+1}, y_i^{tr}\}_{i=1,\dots,n}$, the number of rounds n_r , k

Training:

 Freeze FastGRNN, unfreeze EMI-FastGRNN

repeat

 Train **EMI-FastGRNN**($\{\{Z_{i,\tau}^{tr}\}, \mathbb{1}_{y_i^{tr} \neq -1}\}$)

until convergence

 Freeze EMI-FastGRNN, unfreeze FastGRNN

repeat

 Train **FastGRNN**($\{\mathcal{E}(\{Z_{i,\tau}^{tr}\}, y_i^{tr})\}$), minimizing loss
 $\frac{1}{n} \sum_i \mathbb{1}_{y_i \neq -1} \ell(f_u(\mathcal{E}(\{Z_{i,\tau}^{tr}\})), y_i)$

until convergence

 Unfreeze both EMI-FastGRNN and FastGRNN

for $r \in n_r$ **do**

if $r < \frac{n_r}{2}$ **then**
 $\mathcal{L}_{lower} \leftarrow$ MI-loss

else
 $\mathcal{L}_{lower} \leftarrow$ EMI-loss

end if

repeat

 Train **MSC-RNN**($\{\{Z_{i,\tau}^{tr}\}, \mathbb{1}_{y_i^{tr} \neq -1}\}$) minimizing loss
 $\mathcal{L}_{lower} + \frac{1}{n} \sum_i \mathbb{1}_{y_i \neq -1} \ell(f_u(\mathcal{E}(\{Z_{i,\tau}^{tr}\})), y_i)$

until convergence

end for

to prevent the radar returns from being adversely affected by multipath effects and because they are not central to the smart city scenarios. A partial distribution of displacement durations is provided in Figure 6(a). Each data collect has associated with it the corresponding ground truth, recorded with motion-activated trail cameras or cellphone video cameras, with which the radar data was correlated offline to “cut” and label the source displacement snippets appropriately¹. The datasets have been balanced in the number of human and non-human displacement points where possible, and windowed into snippets of 1, 1.5, and 2 seconds which correspond to 256, 384, and 512 I-Q sample pairs respectively. We note that due to the duration of collections and differences in average displacement lengths, etc., humans are underrepresented in these datasets compared to the other labels. Table 3(b) shows the

¹The radar dataset, which we have open-sourced, does not include individually identifiable information of living individuals and is thus not considered research with human subjects per 45 CFR §46.102(e)(1)(ii).

number of training, validation, and test points for each of these window lengths on a roughly 3:1:1 split. Currently, only the cattle set has multiple concurrent targets; efforts to expand our datasets with target as well as radar type variations are ongoing.

Table 3: Radar evaluation datasets

(a) Source displacement counts and clutter durations

Env.	Data Type	
	Type	Count
Building foyer	Human, Gym ball	52, 51
Indoor amphitheater	Human, Gym ball	49, 41
Parking garage bldg.	Human	268
Parking lot	Human, Car	50, 41
Indoor soccer field	Human, Gym ball	90, 82
Large classroom	Human, Gym ball	48, 50
Cornfield	Human, Dog	117, 85
Cattle shed	Cow	319
Playground	Clutter	45 mins
Parking garage bldg.	Clutter	45 mins
Public park	Clutter	45 mins
Garden	Clutter	45 mins
Lawn	Clutter	20 mins

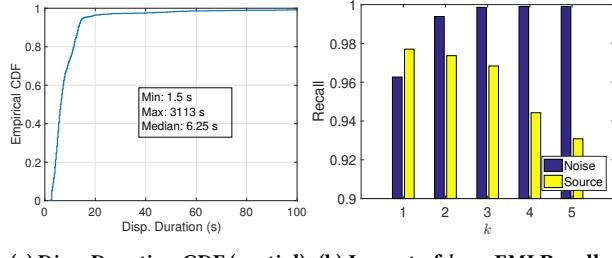
(b) Windowed data from (a) showing number of training, validation, and test points

Window Len. (s)	#Windows		
	Training	Validation	Testing
1	17055	5685	5685
1.5	11217	3739	3739
2	8318	2773	2773

5.2 Evaluation Methodology

Our proposed architecture is compared with existing shallow radar solutions that use feature handcrafting in the amplitude, phase and spectral domains, as well as with other MIL RNNs. In all cases involving RNNs, the radar data is represented purely in the time-domain. The models chosen for this evaluation are:

- (a) **2-tier SVM with phase unwrapped displacement detection.** Phase unwrapping [12] is a widely used technique in radar displacement detection due to its computational efficiency. The idea is to construct the relative trajectory of a source by accumulating differences in successive phase measurements, whereby clutter can be filtered out. We contrast

(a) Disp. Duration CDF (partial) (b) Impact of k on EMI Recalls**Figure 6: Source detected duration CDF for the data in Table 3(a) and how the hyperparameter k in 2-class EMI affects their detection (1-second windows)**

MSC-RNN with a two-tier solution proposed in [34], which uses a robust variant of phase unwrapping with adaptive filtering of clutter samples.

- (b) **3-class SVM.** A clutter vs human vs non-human SVM solution that uses feature handcrafting.
- (c) **EMI-FastGRNN.** An EMI version of FastGRNN (Section 4).
- (d) **EMI-LSTM.** An EMI version of the LSTM. Note that this is a much heavier architecture than the former, and should not be regarded as suitable for a microcontroller device.

Since shallow featurization incurs high incremental overhead, real-time micro-power radar solutions typically only compute the highest information features. For the SVM solutions, training is performed with the best 15 features selected by the *minimal-redundancy maximal-relevance* (mRMR) criterion [32].

For the MIL experiments, the windowed data from Table 3(b) is further reshaped into instances of length 48×2 samples with a fixed stride of 16×2 , where 2 refers to the number of features (I and Q components of radar data). For example, for 1-second windows, the shape of the training data for MIL experiments is $(17055, 14, 48, 2)$, and the shape of the corresponding instance-level one-hot labels is $(17055, 14, 3)$. In the interest of fairness and also to avoid a combinatorial exploration of architectural parameters, we present results at fixed hidden sizes of 16, 32, and 64. For MSC-RNN, the lower tier’s output (embedding) dimension and upper tier’s hidden dimension are kept equal; however, in practice, it is easy to parameterize them differently since the former only affects the latter’s input dimension.

5.2.1 Hyperparameters. Table 4 lists the hyperparameter combinations used in our experiments. For the upper-tier source discrimination comparison in Section 5.3.3, FastGRNN is also allowed to select its optimum input length from 16, 32, and 64 samples.

The selection of the EMI hyperparameter k merits some discussion, in that it controls the extent of “strictness” we assign to the definition of *displacement*. A higher k makes it more difficult for a current window to be classified as a source unless the feature of interest is genuinely compelling. Expectedly, this gives a trade-off between clutter and source recall as is illustrated in Figure 6(b). As explained in Section 1, controlling for false positives is extremely important in radar sensing contexts such as intrusion detection. Hence, we empirically set k to 10, the smallest value that gives a clutter recall of 0.999 or higher in our windowed datasets.

Table 4: Training hyperparameters used

Model	Hyperparameter	Values
EMI/FastGRNN	Batch size	64, 128
	Hidden size	16, 32, 64
	Gate non-linearity	sigmoid, tanh
	Update non-linearity	sigmoid, tanh
	k	10
	Keep prob. (EMI-LSTM)	0.5, 0.75, 1.0
SVM	Optimizer	Adam
	c	1e-3, 1e-2, 0.1, 1, 10, 100, 1e3, 1e5, 1e6
	γ	1e-3, 1e-2, 5e-2, 0.1, 0.5, 1, 5, 10

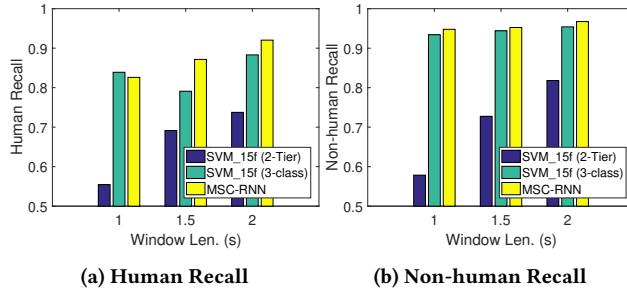
5.3 Results

5.3.1 Comparative Classifier Performance. We compare the inference accuracy and recalls of MSC-RNN, with the RNN and shallow solutions outlined in Section 5.2.

Recall that we have purposefully devised a purely time-domain solution for source discrimination for efficiency reasons, since one of the main components of featurization overhead is that of FFT computations. Figure 7 compares MSC-RNN with engineered features in the amplitude, time, and spectral domains that are optimized for micro-power radar classification. For the two-tier SVM, the source recalls for increasing window sizes are inferred from Figure 8 (discussed in Section 5.3.3). We find that MSC-RNN significantly outperforms the 2-tier SVM solution in terms of human and non-human recalls, even with features learned from the raw time-series. Similarly, for the 3-class case, our solution provides much more stable noise robustness and is generally superior even to the much heavier SVM solution.

Figure 9 contrasts our model with 3-class EMI-FastGRNN and EMI-LSTM, for fixed hidden sizes of 16, 32, and 64 respectively. It can be seen that MSC-RNN outperforms the monolithic EMI algorithms on all three metrics of accuracy, non-human and human recalls (with one exception for EMI-LSTM). Notably, cascading significantly enhances the non-dominant class recall over the other methods, especially for larger hidden sizes, and therefore offers better resilience to the source type imbalance in radar datasets.

5.3.2 Runtime Efficiency Comparison - MSC-RNN vs. Feature Hand-crafting. Table 5 lists the runtime duty cycle estimates of MSC-RNN versus shallow SVM alternatives in two deployment contexts with realistic clutter conditions, supported by usage statistics of a popular biking trail in Columbus, OH [2]. While the 2-tier SVM understandably has the lowest duty cycle due to a cheap lower tier, it is not a competitive solution as established in Section 5.3.1. The 3-class SVM, on the other hand, is dominated by the feature computation overhead. While the 48×2 MSC-RNN formulation is about $1.75 \times$ as efficient as using handcrafted features, it is possible to reduce instance-level computations even further by using longer input vectors and reducing the number of iterations. As an example, an alternate formulation of MSC-RNN with a 16-dimensional input vector at the instance level is $3 \times$ more efficient than feature engineering.



(a) Human Recall

(b) Non-human Recall

Win. Len. (s)	Accuracy		Clutter Recall	
	SVM_15f (3-class)	MSC-RNN	SVM_15f (3-class)	MSC-RNN
1	0.851	0.944	0.758	0.999
1.5	0.934	0.954	0.996	0.999
2	0.959	0.972	0.999	1.000

(c) Accuracy and Clutter Recall (3-class SVM and MSC-RNN)

Figure 7: Classification comparison of purely time-domain FastGRNN with two SVM solutions: (a) a 2-tier system using a phase unwrapped clutter rejector as the lower tier, and (b) a 3-class SVM. Both use 15 high information features hand-crafted in the amplitude, time, and spectral domains

Table 5: Estimated featurization duty cycle comparison on ARM Cortex-M3

Architecture	Est. Duty Cycle (Cortex-M3)	
	97% Clutter	98% Clutter
MSC-RNN (Inp. dim.=2)	21.00%	20.00%
MSC-RNN (Inp. dim.=16)	10.87%	10.70%
2-Tier SVM	2.05%	1.70%
3-Class SVM	35.00%	35.00%

5.3.3 Tier-wise Evaluation. We next compare the lower-tier and upper-tier classifiers individually to their shallow counterparts in the 2-tier SVM solution.

Tier 1 Classifier. Figure 8 compares the probabilities of missed detections versus displacement durations for the 3-outof-4 displacement detector and the EMI component of our solution (for a principled approach to choosing parameters for the former, refer to Appendix A) at hidden sizes of 16, 32, and 64. It can be seen that, for the shortest cut length of 1.5 s in the dataset, the detection probability is improved by up to 1.5× (1.6×) over the 3-outof-4 detector with false alarm rates of 1/week and 1/month respectively even when the false alarm rate (1-test clutter recall) of EMI is 0, which translates to a false alarm rate of <1 per year. Further, the EMI detector converges to 0 false detects with displacements ≥ 2.5 s, and is therefore able to reliably detect walks 2.6× shorter than the previous solution. Therefore, it is possible to restrict false positives much below 1/month while significantly improving detectability over the M-outof-N solution. Since the clutter and source datasets span various backgrounds (Figure 5), MSC-RNN offers superior cross-environmental robustness.

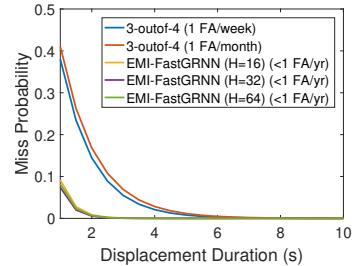


Figure 8: Comparison of miss probabilities versus displacement durations of Tier 1 classifier vs. 3-outof-4 phase unwrapped displacement detector

Tier 2 Classifier. We now show that the gains of MSC-RNN over the 2-tier SVM solution are not, in fact, contingent on the quality of the underlying displacement detector for the latter. For this experiment, we train a 2-class FastGRNN on embeddings derived from the Tier 1 EMI-FastGRNN. Table 6 compares its performance with the upper-tier SVM from the latter when trained with the best 15 cross-domain features obtained from the raw radar samples. It can be seen that the purely time-domain FastGRNN still generally outperforms the 2-class SVM on all three metrics of accuracy, human recall, and non-human recall. Thus, it is possible to replace feature engineering with deep feature learning and enjoy the dual benefits of improved sensing and runtime efficiency for this class of radar applications.

Table 6: Independent of the Tier 1 classifier, the Tier 2 source-type classifier outperforms the SVM

Win. Len. (s)	Accuracy		Human Recall		Non-human Recall	
	SVM	Fast- _15f	SVM	Fast- _15f	SVM	Fast- _15f
1	0.93	0.93	0.90	0.90	0.93	0.94
1.5	0.93	0.93	0.90	0.93	0.95	0.95
2	0.93	0.96	0.86	0.96	0.96	0.97

6 LOW-POWER IMPLEMENTATION

The radar sensor described in Figure 1(a) uses an ARM Cortex-M3 microcontroller with 96 KB of RAM and 4 MB of flash storage. It runs eMote [37], a low-jitter near real-time operating system with a small footprint. We emphasize that energy efficient compute, not working memory or storage, is the bigger concern for efficient real-time operation. Hence, we take several measures to efficiently implement the multi-scale RNN to run at a low duty cycle on the device. These include low-rank representation of hidden states, Q15 quantization, and *piecewise-linear* approximations of non-linear functions. The latter in particular ensures that all the computations can be performed with integer arithmetic when the weights and inputs are quantized. For example, $\tanh(x)$ can be approximated as: $\text{quantTanh}(x) = \max(\min(x, 1), -1)$, and $\text{sigmoid}(x)$ can be approximated as: $\text{quantSigmoid}(x) = \max(\min(\frac{x+1}{2}, 1), 0)$. The underlying linear algebraic operations are implemented using the CMSIS-DSP

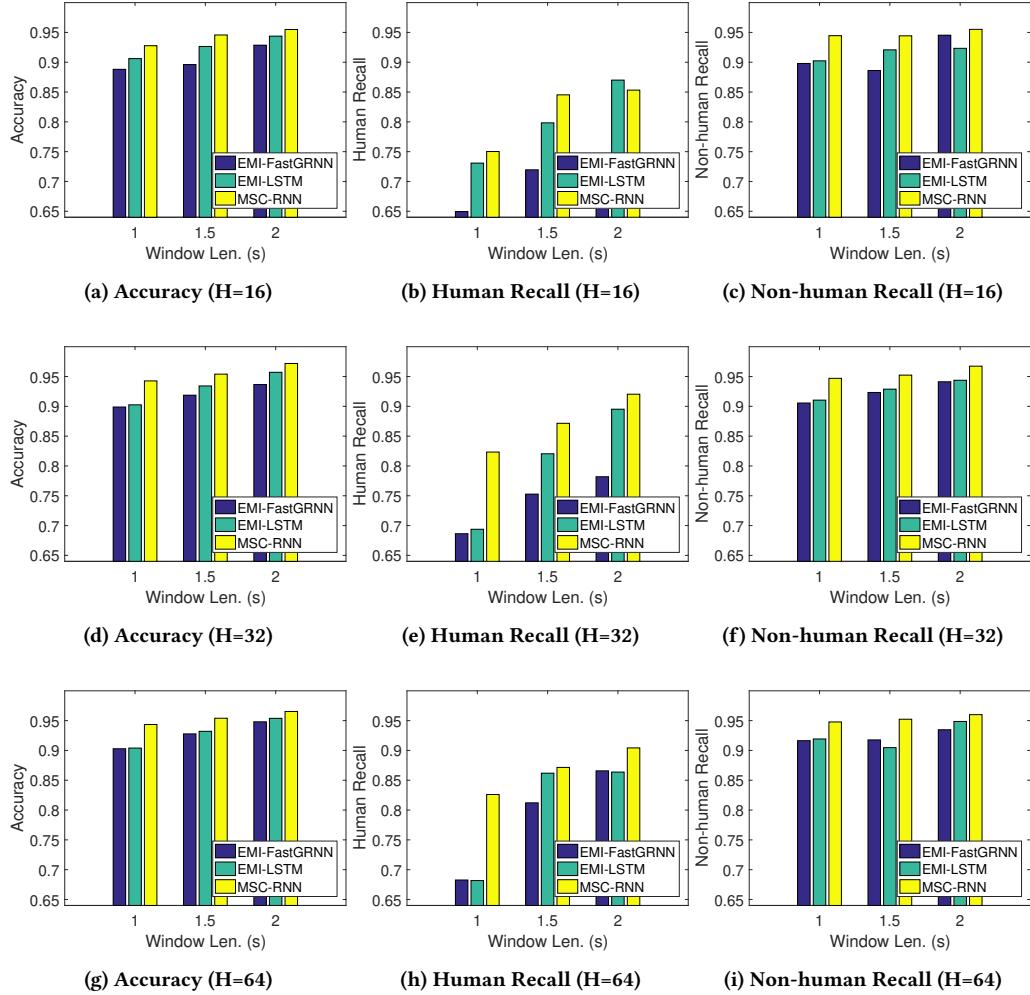


Figure 9: Sensing performance comparison of MSC-RNN with EMI-FastGRNN and EMI-LSTM

library [1]. While advanced ARM processors such as Cortex-M4 offer floating point support, it should be noted that, for efficiency reasons, using sparse, low rank matrices and quantization techniques are beneficial in general.

7 CONCLUSION AND FUTURE WORK

In this work, we introduce multi-scale, cascaded RNNs for radar sensing, and show how leveraging the ontological decomposition of a canonical classification problem into clutter vs. source classification, followed by source type discrimination on an on-demand basis can improve both sensing quality as well as runtime efficiency over alternative systems. Learning discriminators at the time-scales relevant to their respective tasks, and jointly training the discriminators while being cognizant of the cascading behavior between them yields the desired improvement.

The extension of MSC-RNNs to more complicated sensing contexts is a topic of future work. Of interest are regression-based radar “counting” problems such as occupancy estimation or active transportation monitoring, where the competitiveness of MSC-RNN to

architectures such as TCNs [4] could be insightful. We also believe that MSC-RNN could also apply to alternative sensing for smart cities and built environments where the sources have intrinsic ontological hierarchies, such as in urban sound classification [5].

ACKNOWLEDGEMENTS

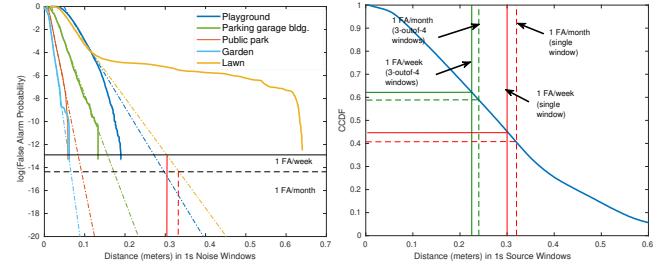
We thank our shepherd, Zheng Yang, and the anonymous reviewers for their comments. We are indebted to Don Dennis, Prateek Jain, and Harsha Vardhan Simhadri at Microsoft Research India for their suggestions and feedback. The computation for this work was supported by the Ohio Supercomputer Center [7] project PAS1090, the IIT Delhi HPC facility, and Azure services provided by Microsoft Research Summer Workshop 2018: Machine Learning on Constrained Devices ².

²<https://www.microsoft.com/en-us/research/event/microsoft-research-summer-workshop-2018-machine-learning-on-constrained-devices/>

REFERENCES

- [1] [n. d.]. CMSIS-DSP Software Library. <http://www.keil.com/pack/doc/CMSIS/DSP/html/index.html>.
- [2] [n. d.]. Olentangy Trail usage, Columbus, OH. [https://www.columbus.gov/recreationandparks/trails/Future-Trails-\(Updated\)./](https://www.columbus.gov/recreationandparks/trails/Future-Trails-(Updated)./)
- [3] Joshua Adkins, Branden Ghena, et al. 2018. The signpost platform for city-scale sensing. In *IPSN*. IEEE, 188–199.
- [4] Shaojie Bai, J Zico Kolter, et al. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).
- [5] Juan P. Bello, Claudio Silva, et al. 2019. SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution. *CACM* 62, 2 (Jan. 2019), 68–77.
- [6] Charles E Catlett, Peter H Beckman, et al. 2017. Array of things: A scientific research instrument in the public way: platform design and early lessons learned. In *SCOPE*. ACM, 26–33.
- [7] Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. <http://osc.edu/ark:/19495/f5s1ph73>
- [8] Kyunghyun Cho, Bart Van Merriënboer, et al. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
- [9] Junyoung Chung, Sungjin Ahn, et al. 2016. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704* (2016).
- [10] Don Dennis, Chirag Pabbaraju, et al. 2018. Multiple instance learning for efficient sequential data classification on resource-constrained devices. In *NIPS*. Curran Associates, Inc, 10975–10986.
- [11] Don Kurian Dennis, Sridhar Gopinath, et al. [n. d.]. EdgeML: Machine learning for resource-constrained edge devices. <https://github.com/Microsoft/EdgeML>
- [12] Richard M Goldstein, Howard A Zebker, et al. 1988. Satellite radar interferometry: Two-dimensional phase unwrapping. *Radio science* 23, 4 (1988), 713–720.
- [13] Song Han, Huizi Mao, et al. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *ICLR* (2016).
- [14] Jin He and Anish Arora. 2014. A regression-based radar-mote system for people counting. In *PerCom*. IEEE, 95–102.
- [15] Jin He, Dhrubojoyti Roy, et al. 2014. Mote-scale human-animal classification via micropower radar. In *Sensys*. ACM, 328–329.
- [16] Geoffrey E Hinton. 1990. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence* 46, 1–2 (1990), 47–75.
- [17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [18] Rios Jesus Javier and Youngwook Kim. 2014. Application of linear predictive coding for human activity classification based on micro-Doppler signatures. *GRSL, IEEE* 11 (2014), 1831–1834.
- [19] Branka Jokanovic, Moeness Amin, et al. 2016. Radar fall motion detection using deep learning. In *RADAR*. IEEE, 1–6.
- [20] Youngwook Kim and Hao Ling. 2008. Human activity classification based on micro-Doppler signatures using an artificial neural network. In *AP-S*. IEEE, 1–4.
- [21] Youngwook Kim and Brian Toomajian. 2016. Hand gesture recognition using micro-Doppler signatures with convolutional neural network. *IEEE Access* 4 (01 2016), 1–1.
- [22] Dimitrios Kotzias, Misha Denil, et al. 2014. Deep multi-instance transfer learning. *arXiv preprint arXiv:1411.3128* (2014).
- [23] Alex Krizhevsky, Ilya Sutskever, et al. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
- [24] Ashish Kumar, Saurabh Goyal, et al. 2017. Resource-efficient machine learning in 2 KB RAM for the Internet of Things. In *ICML*. ACM, 1935–1944.
- [25] Sangeeta Kumari, Dhrubojoyti Roy, et al. 2019. EdgeL³: Compressing L³-Net for mote scale urban noise monitoring. In *IPDPSW*. IEEE, 877–884.
- [26] Aditya Kusupati, Manish Singh, et al. 2018. FastGRNN: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In *NIPS*. Curran Associates, Inc., 9030–9041.
- [27] Liang Liu, Mihail Popescu, et al. 2011. Automatic fall detection based on Doppler radar motion signature. In *PervasiveHealth*, Vol. 222. 222–225.
- [28] Gihan Mendis, Tharindu Randeny, et al. 2016. Deep learning based Doppler radar for micro UAS detection and classification. In *MILCOM*. IEEE, 924–929.
- [29] Michael C Mozer. 1992. Induction of multiscale temporal structure. In *NIPS*. Morgan-Kaufmann, 275–282.
- [30] Dustin P. Fairchild and Ram Narayanan. 2014. Classification of human motions using empirical mode decomposition of human micro-Doppler signatures. *Radar, Sonar & Navigation, IET* 8 (06 2014), 425–434.
- [31] Razvan Pascanu, Tomas Mikolov, et al. 2013. On the difficulty of training recurrent neural networks. In *ICML*. ACM, 1310–1318.
- [32] Hanchuan Peng, Fuhui Long, et al. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.
- [33] Dhrubojoyti Roy, Sangeeta Kumari, et al. [n. d.]. MSC-RNN: Multi-Scale, Cascaded RNNs for Radar Classification. <https://github.com/dhruboroy29/MSCRNN>

A PARAMETER SELECTION FOR M-OUTOF-N DISPLACEMENT DETECTOR



(a) Clutter threshold selection for (b) Relaxation of per-window 1 FA/week and 1 FA/month threshold through aggregation

Figure 10: Shallow displacement detector parameter selection using the datasets from Table 3(a); here, M=3 and N=4

We discuss the parameter selection process for the unwrapped-phase displacement detector [34] referenced in Figures 7 and 8 in a principled manner. Figure 10(a) shows the cumulative distribution of unwrapped phase changes of environmental clutter, translated into real distance units, in various environments for 1-second integration windows from the clutter datasets in Table 3(a). The data is extrapolated using linear fitting on a logarithmic scale to estimate the required phase thresholds to satisfy false alarm rates of 1 per week and 1 per month respectively (derived using Bernoulli probabilities). We see that the unwrapped thresholds for 1 false alarm per week and month correspond to 0.3 and 0.32 m respectively. In this analysis, we fix the IQ rejection parameter at 0.9, which gives us the most lenient thresholds.

Figure 10(b) illustrates the CCDFs of phase displacements for all source types (humans, gym balls, dogs, cattle, and slow-moving vehicles) in our dataset combined, calculated over 1-second windows. Setting thresholds based on the previous analysis, the probability of false negatives per window is still significant. In practice, the algorithm improves detection by basing its decision on 3-outof-4 sliding windows, where detectability improves since the threshold per window is now $\frac{3}{4} \times$ the original threshold. For 1 false alarm per week (month), the displacement threshold for the 3-outof-4 detector reduces to 0.22 m (0.24 m) per window, with an improved detection probability of 0.59 (0.62).