

# Learning behavioral context recognition with multi-stream temporal convolutional networks

Aaqib Saeed, Tanir Ozcelebi, \*Stojan Trajanovski, Johan Lukkien

a.saeed@tue.nl, t.ozcelebi@tue.nl, stojan.trajanovski@philips.com, j.j.lukkien@tue.nl

Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

\*Philips Research, Eindhoven, The Netherlands

## Abstract

Smart devices of everyday use (such as smartphones and wearables) are increasingly integrated with sensors that provide immense amounts of information about a person's daily life such as behavior and context. The automatic and unobtrusive sensing of behavioral context can help develop solutions for assisted living, fitness tracking, sleep monitoring, and several other fields. Towards addressing this issue, we raise the question: can a machine learn to recognize a diverse set of contexts and activities in a real-life through joint learning from raw multi-modal signals (e.g. accelerometer, gyroscope and audio etc.)? In this paper, we propose a multi-stream temporal convolutional network to address the problem of multi-label behavioral context recognition. A four-stream network architecture handles learning from each modality with a contextualization module which incorporates extracted representations to infer a user's context. Our empirical evaluation suggests that a deep convolutional network trained end-to-end achieves an optimal recognition rate. Furthermore, the presented architecture can be extended to include similar sensors for performance improvements and handles missing modalities through multi-task learning without any manual feature engineering on highly imbalanced and sparsely labeled dataset.

## Introduction

The problem of context recognition is centered on inferring person's environment, physical state, and activity performed at any particular time. Specifically, a understanding of the user's current context requires determining where and with whom the person is? and in what type of activity the person is involved in? The behavioral and activity analysis is an important and challenging task mainly because it is crucial for several applications, including smart homes (Rashidi and Cook 2009), assisted living (Lin et al. 2015; Rashidi and Mihailidis 2013), fitness tracking (Rabbi et al. 2015), sleep monitoring (Lin et al. 2012), user-adaptive services, social interaction (Lee et al. 2013) and in industry. In particular, an accurate recognition of human context can greatly benefit healthcare and wellbeing through automatic monitoring and supervision of patients with chronic diseases (Lara and Labrador 2013) such as hypertension, diabetes and dementia (Ordóñez and Roggen 2016). Furthermore, the gathered

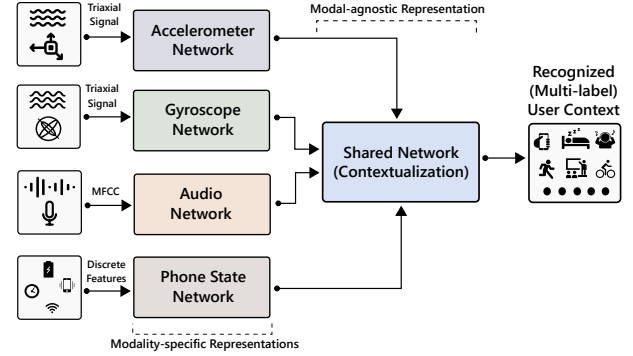


Figure 1: **Multi-modal representation learning from sensors:** Schematic of the proposed multi-stream convolutional network.

knowledge and extracted activity patterns can enable novel treatment design, adjustment of medications, better behavioral intervention and patient observation strategies (Lorincz et al. 2009).

In practice, for a context detection system to be effective in a real-life requires an unobtrusive monitoring. It is important to not distress a person in order to capture their realistic behaviors in a natural environment. The penetration of smart sensing devices (e.g. smartphones and wearables) that are integrated with sophisticated sensors in our daily lives provides a great opportunity to learn and infer about various aspects of a person's daily life. However, there is considerable variability in the human behavior in real-world situations that can cause the system to fail, if it is developed using data collected in a constrained environment. For instance, Miluzzo et al. shows that the accuracy of activity classification differs based on the interaction with the phone e.g. when in hand or carried in the bag. The various sensors embedded in the smart devices convey information about different ambient facets each with a distinct prospect. The variability issues of different patterns in phone usage, environments, and device types can be very well addressed (to improve the recognition capability of the system) through learning disentangled representations from a large-scale data source and fusing rich sensory modalities rather than separately utilizing each of them.

In the past, several studies have shown great improvement

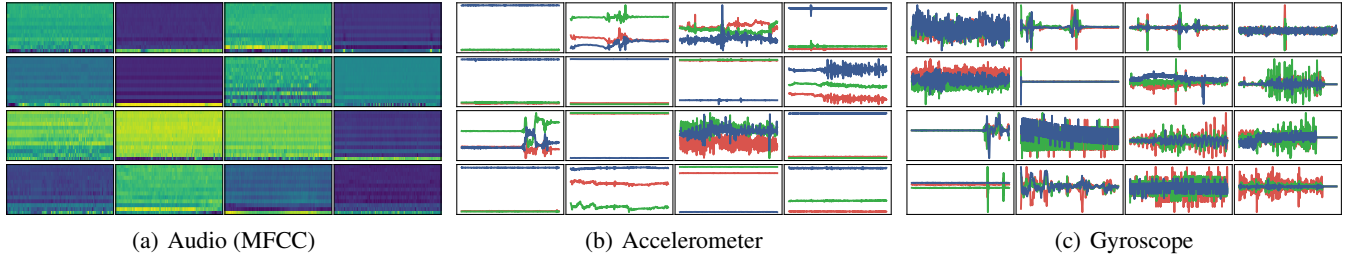


Figure 2: **Context recognition dataset:** Samples from large-scale multi-modal sensory data collected in-the-wild conditions. The individual plots within each sub-figure correspond to the same set of activities/context.

in sensor processing for basic activity recognition (Lara and Labrador 2013; Hoseini-Tabatabaei, Gluhak, and Tafazolli 2013). The majority of the earlier methods use shallow learning classifiers (such as, Random Forest and Support Vector Machine) with hand-engineered features extracted from raw sensor readings e.g. heuristically selected statistical or frequency measures (Figo et al. 2010). Likewise, many studies involve simulated controlled trials for data collection in lab environments that require users to wear extra sensors. Broadly, they also treat activity recognition as a multi-class classification problem, where a user’s activity at a specific moment can be defined by one of the  $k$  defined classes. On the contrary, people are not generally engaged in just one activity in their day-to-day living e.g. a person might surf the web while eating or talking to friends. These problems limit the applicability of these studies to detect very few rudimentary activities and make it harder for the system to generalize to real-life settings. Nevertheless, to be successful in everyday scenarios, the context recognition module should support a diverse set of activities, varying device usage, and a wide range of environments. Importantly, it must not only learn discriminative representations directly from raw signals without any ad-hoc feature engineering, but also seamlessly combine the discovered explanatory factors in the milieu of diverse sensory modalities (Bengio, Courville, and Vincent 2013).

In recent years, the fields of speech recognition, drug discovery, image segmentation and machine translation have been tremendously revolutionized thanks to the availability of massive labeled datasets and end-to-end deep representation learning (Bengio, Courville, and Vincent 2013). Similarly, the domain of human activity recognition has also started leveraging deep neural networks for automatic feature learning (Ordóñez and Roggen 2016; Radu et al. 2018; Yang et al. 2015) though commonly restricted to the detection of only elementary activities such as, walking, sitting, standing etc. There has not been the same progress in recognizing complex behavioral context in daily-life situations using devices of daily use. This can be partially attributed to the lack of a large labeled dataset, which is both expensive and time-consuming to accumulate in a real-world settings. We believe that large-scale sensory data can significantly advance context recognition. This issue is very recently addressed in (Vaizman, Ellis, and Lanckriet 2017; Vaizman, Weibel, and Lanckriet 2018) which has open-

sourced multi-modal data (see Figure 2) of activities in-the-wild. The authors provide a baseline system for sensor fusion and a unified model for multi-label classification. They trained logistic regression and fully connected neural networks on hand-crafted features that are extracted based on extensive domain-knowledge. In this paper, we utilize this heterogeneous sensors data collected over a week from sixty users to learn rich representations in an end-to-end fashion for recognizing multi-label human behavioral context.

The task of learning detailed human context is challenging, especially from imbalanced and multi-label data. Unconstrained device usage, a natural environment, different routines, and authentic behaviors are likely to result in a joint training dataset from several users with significant class skew (Vaizman, Weibel, and Lanckriet 2018) and missing labels. Another challenge with learning from multi-modal signals is developing an architecture that feasibly combines them as in diverse environments a certain sensor might perform better than others. For instance, if a person is watching a television with a phone lying on the table, the sound modality may dominate in the network as compared to an accelerometer. We address the former issue with instance weighting scheme same as (Vaizman, Weibel, and Lanckriet 2018) and later through a unified architecture that can efficiently fuse representations in multiple ways.

We present a deep temporal convolutional neural network (CNN) that learns directly from various modalities through a multi-stream architecture (accelerometer, gyroscope, sound and phone state networks). Here, a separate network facilitates learning from each modality and a contextualization module incorporates all the available information to determine the user’s context (see Figure 1). In our experiments, we show that deep multi-modal representations learned through our network without any sophisticated pre-processing or manual feature extraction achieve state-of-the-art performance.

The primary contribution of this paper is in showing how to leverage ample amount of raw sensory data to learn deep cross-modal representations for multi-label behavioral context. Although, the methods in the paper are standard, their application on a large-scale imbalanced and sparsely labeled smartphone data set is unique. The proposed network architecture achieves sensitivity and specificity score of 0.767 and 0.733, respectively averaged over 51 labels and 5-folds cross-validation. The rest of the paper describes our tech-

nique and experiments in detail. First, we review the related work on activity recognition. Then we present our multi-stream temporal convolutional network, architectural modifications for handling missing sensors, the proposed training procedure and implementation details. Next, the description of the dataset, evaluation protocol and experimental results are described, followed by the conclusions.

## Related Work

Human activity recognition has been extensively studied in simulated and controlled environments. It is concerned with classifying sensor measurements into existing activity categories. The earlier techniques are predominantly based on applying shallow learning algorithms on manually extracted features (e.g. statistical and spectral attributes) (Figo et al. 2010). Despite there are unsupervised (Bhattacharya et al. 2014; Plötz, Hammerla, and Olivier 2011) and supervised (Yang et al. 2015; Ordóñez and Roggen 2016; Ronao and Cho 2016; Zeng et al. 2014) deep learning methods applied for automatic feature extraction to detect activities, these approaches are fairly limited by the amount of labeled data (of many sensing modalities) from the real-world. Furthermore, they do not fully address the issue of multi-label context recognition. A user state is described by only one class or label, which is not true for activities humans perform in real-life. Moreover, only recently the exploration has begun into joint-learning and fusing multiple modalities for ubiquitous sensing through deep networks (Radu et al. 2018; Vaizman, Weibel, and Lanckriet 2018). The works cited here are by no means an exhaustive list, but provide a recent representative advancements made in utilizing deep neural networks for activity recognition. We recommend the interested readers to refer (Rashidi and Mihailidis 2013; Shoaib et al. 2015) for an extensive survey of former approaches.

A systematic analysis of several deep neural architectures for activity recognition is provided by Hammerla, Halloran, and Ploetz. The suitability of various models is investigated that were trained only on raw accelerometer signals for different activity classification tasks. On diverse benchmark datasets, CNN and long-short-term memory networks are found to outperform hand-crafted features by a significant margin. Likewise, Alsheikh et al. proposed an approach combining pre-training and fine-tuning of deep belief networks for sequential activity recognition. They extracted spectrograms from a triaxial accelerometer and found them to be helpful for capturing variations in the input. Similarly, Jiang and Yin used 2D activity images extracted from accelerometer signals as CNN input. The importance of unsupervised training of models in feature learning and optimization is highlighted in (Bhattacharya et al. 2014) using a combination of sparse-coding framework and semi-supervised learning. Likewise, Yang et al. developed a multi-channel CNN model to replace heuristic based hand-crafted features. Their analysis showed CNNs work well compared to traditional (shallow) learning algorithms on several datasets. Audio sensing is also employed in unconstrained acoustic environments through applying fully connected neural networks (Lane, Georgiev, and Qendro 2015).

Recently, Radu et al. used deep networks for multi-modal activity recognition and compared them with traditional learning algorithms on various recognition tasks. Likewise, numerous other studies also positively utilize deep learning for detection of basic activities (Ordóñez and Roggen 2016; Ronao and Cho 2016; Zeng et al. 2014).

We differentiate ourselves from the existing approaches through utilizing a deep multi-stream CNN (with depth-wise separable convolutions) on a large and diverse context detection dataset. Specifically, we build on previous work by Vaizman, Weibel, and Lanckriet that only employed hand-engineered features for training linear and shallow neural networks. In contrast, our general-purpose approach allows us to train a deeper network that can not only automatically discover hidden latent factors, but also seamlessly combine them to achieve an end-to-end learning system without requiring domain expertise. Moreover, through taking advantage of multi-task learning (Caruana 1997) we develop an architecture that can robustly handle missing sensors.

## Learning Multi-Modal Networks

We design a deep convolutional neural network to address the problem of behavioral context recognition through learning representations from raw sensory inputs. To deal with cross-modality signals i.e. accelerometer (Acc), gyroscope (Gyro), audio (MFCC/Aud), and phone state (PS), we use a multi-stream architecture. The network comprises five main modules as demonstrated in Figure 3. This section describes each component, presents a strategy to modify the proposed architecture to handle missing sensors and provides the implementation details.

### Modality Specific Networks

We present a deep multi-modal convolutional architecture for learning context representations. We propose to use a series of depthwise-separable convolutions (DPS-Conv) (Chollet 2017) for processing different components (or channels) of raw signals. In general, CNNs are also found to be well suited for processing 1D sequences due to their ability to learn translation invariant features, scale separation, and localization of filters across time and space (Bai, Kolter, and Koltun 2018). DPS-Conv consists of two operations i.e. a depthwise convolution and a pointwise (or  $1 \times 1$ ) convolution. Specifically, the first function (depthwise convolution) performs a convolution independently over each input channel and it is followed by the second operation of  $1 \times 1$  convolution that projects the channels estimated by the earlier onto a distinct channel space to have the same number of output filters (Kaiser, Gomez, and Chollet 2017). The intuition of this formulation falls in line with the classical procedures utilized by domain experts to extract several features from each signal component independently (e.g.  $x$ ,  $y$  and  $z$  constituents of an accelerometer) but pointwise convolution goes one step further and tries to learn unified factors that may capture relationships among independent elements. Moreover, separable convolutions make efficient use of parameters as opposed to their classical counterpart and this

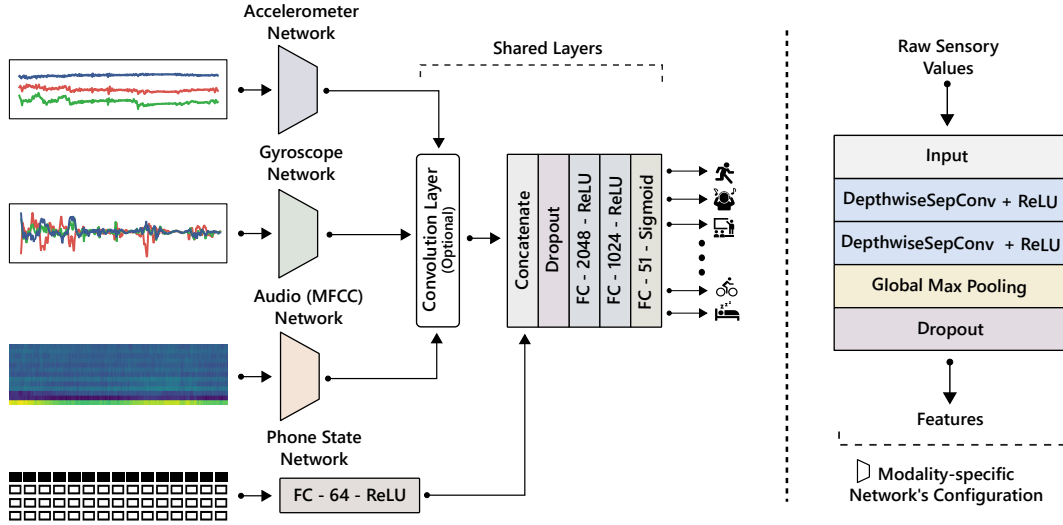


Figure 3: **End-to-end multi-modal and multi-label context recognition:** We propose a deep temporal convolutional architecture for multi-label behavioral context recognition. A separate network learns representations (features) from each modality using depthwise-separable convolutions and contextualizes this information through shared layers to infer the user context.

property has made them a very promising candidate for contemporary architectures that run on smart devices with limited computing and energy capabilities (Sandler et al. 2018; Zhang et al. 2017). Formally, in case of 1D input sequence  $\mathbf{x}$  of length  $L$  with  $M$  channels, the aforementioned operation can be formulated as follows (Kaiser, Gomez, and Chollet 2017):

$$\text{DepthwiseConv}(\mathbf{x}, \mathbf{w})_i = \sum_l (\mathbf{x}[i : i + k - 1] \odot \mathbf{w})_l$$

$$\text{PointwiseConv}(\mathbf{x}, \mathbf{w})_i = \sum_m (\mathbf{x}[i : i + k - 1] \cdot \mathbf{w})_m$$

$$\text{DepthwiseSeparableConv}(\mathbf{x}, \mathbf{w}_d, \mathbf{w}_p)_i = \text{PointwiseConv}_i(\text{DepthwiseConv}_i(\mathbf{x}[i : i + k - 1], \mathbf{w}_d), \mathbf{w}_p)$$

where  $\odot$  is elements-wise product,  $\mathbf{x}[i : j]$  represents a segment of the complete sequence with adjacent columns from  $i$  to  $j$ , and  $\mathbf{w}$  represents filter with receptive field size of  $k$ .

The proposed network takes four different signals as input, each with its independent disjoint pathway in the earlier layers of the network. Towards the end, they are merged into shared layers that are common across all modalities that are described in the next subsection. This network configuration has the benefit of not just extracting modality-specific (and channel-specific) features but it can also feasibly extract mutual representations through shared layers. Each of the presented Acc and Gyro networks consist of 2 temporal convolution layers which act as feature extractors over raw signals of dimensions  $800 \times 3$ . The convolution layers have kernel sizes of 64 and 32 with a stride of 2 and each layer has 32 and 64 filters, respectively. We use rectified linear activation in all the layers and apply depth-wise L2-regularization with a

rate of 0.0001. The audio network takes mel frequency cepstral coefficients (see Section Dataset and Modalities) of size  $420 \times 13$  as input and it has a similar architecture except the kernel size, which is set to 8 and 6 in the first and second layers, respectively. Likewise, the discrete attributes indicating PS are fed into a single layer fully-connected (FC) network with 64 units and L1-penalty is used on the weights with a rate of 0.0001. Furthermore, we explore different mechanisms to get a fixed dimension vector from each modality that can be fed into a shared network. Specifically, we use: a) global max pooling (GMP), b) global average pooling (GAP), c) a FC layer, and d) exactly pass the representations without any transformation to the shared network.

### Shared Network (Contextualization)

Given the concepts extracted from each modality, the shared network generates a modal-agnostic representation. To achieve this, we fuse the output of earlier networks either through concatenation or apply standard convolution (only for Acc, Gyro and Aud). We then feed the output into 2 FC layers having 2048, 1024 hidden units, respectively. Same as earlier, we use rectified linear non-linearity and L1-regularization with a weight decay coefficient of 0.0001. The final output layer contains 51 units (one for each label) with sigmoid activation. Figure 3 visualizes the sharing of the network layers, where, earlier layers are modality specific but downstream layers become more general.

### Missing Sensors

In a real-life setting, a context recognition system may encounter missing modalities which can limit its inference capability. To make the model robust against such a situation, we develop a multi-task network (Caruana 1997), where learning from each sensor is posed as a task. The initial configuration of the model is the same as before but an addi-

tional layer (of 128 units for Acc, Gyro, MFCC/Aud and 64 units for PS) with a separate loss function is added after only a single shared layer of 1024 hidden units. Figure 4 provides a high-level overview of the architecture. We employ joint-training (with a learning rate of 0.0003) on all the modalities through aggregating cost functions of each model in order to get a total loss. This architectural configuration allows not only to learn independent and shared factors but enables inference even when any of the sensors is missing. It does so through averaging (which can be weighted) over probabilities produced by the individual networks.

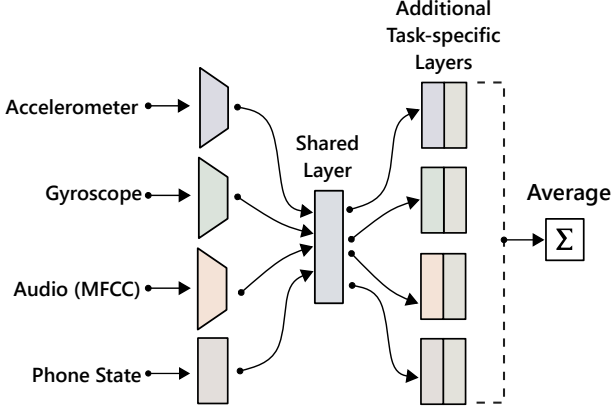


Figure 4: **Handling Missing Sensors with a Multi-task Network:** A variant of the earlier defined architecture with additional task (modality-specific) layers and a separate loss function for each modality. It is able to recognize user context even if only one sensor is producing data and the others are unavailable.

## Implementation and Training Details

The networks are implemented in Tensorflow (Abadi et al. 2016) and the models are learned from scratch; initializing the weights with Xavier technique (Glorot and Bengio 2010). Dropout (Srivastava et al. 2014) is applied on the hidden layers with a probability of 0.2. We use the Adam optimizer with a learning rate of 0.0001 (unless mentioned otherwise) and use a batch size of 100. We optimize the model weights for a fixed number of iterations (i.e. 15000) with mini-batch stochastic gradient descent and backpropagation using instance-weighted cross-entropy objective function:

$$\mathcal{J}_C = \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \Psi_{i,c} \cdot \mathcal{L}_{CE}(\hat{y}_{i,c}, y_{i,c})$$

$$\mathcal{L}_{ce}(\hat{y}, y) = -[(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))]$$

where  $\mathcal{L}_{ce}$  is the binary cross-entropy loss, and  $\Psi$  is an instance-weighting matrix of size  $N \times C$  (i.e. number of training examples and total labels, respectively). The instance weights in  $\Psi$  are assigned by inverse class frequency. Likewise, the entries for the missing labels are set to zero, to impose no contribution in the overall cost from such examples.

## Experimental Results

We conduct several experiments to analyze the capability of the proposed method. First, we provide a brief description of the utilized dataset and signals. Second, we describe the evaluation approach and metrics used to determine the model’s performance on a multi-label and imbalanced dataset. Finally, we discuss our empirical observations, effect of different modalities’ representation, comparison of various procedures to learn shared factors and visualization of the internal representation.

### Dataset and Modalities

We choose to learn discriminative representations directly from raw Acc, Gyro, Aud/MFCC and PS attributes from a smartphone because of their wide adoptability and ubiquity. For this purpose, we chose to leverage *ExtraSensory Dataset* (Vaizman, Ellis, and Lanckriet 2017) since it is collected in a natural environment from users’ personal devices. The experimental setup was not scripted but data collection was performed when participants were busy with their daily routines to capture varied activities and context combinations, in-the-wild conditions. This data source contains over 300,000 multi-labeled instances (with classes such as ‘outside’, ‘at a restaurant’, ‘with friends’ from a total of 51 labels) from sixty users. The complete data collection protocol is described in (Vaizman, Ellis, and Lanckriet 2017). Here, we provide a high-level overview of the signals that we used in this study. The samples are collected for 20 seconds duration every minute from tri-axis Acc and Gyro at a sampling frequency of 40Hz, mel frequency cepstral coefficients (MFCCs) for 46msec frame are extracted from Aud recorded at 22,050Hz. Likewise, several phone state binary features are also collected such as those specifying, time of day, battery level, ringer mode and Wi-Fi connection etc. A few randomly selected samples of these signals are illustrated in Figure 2.

We seek to process raw sensory values without manual feature engineering. Thus, the only pre-processing we applied is to transform variable length inputs to an identical temporal length. For this purpose, the MFCCs of environmental audio are repeated (along time dimension) to get equal size input, this is reasonable for ambient soundscapes as we are not particularly interested in inferring a specific sound event. Similarly, the Acc and Gyro samples of varying sizes are zero-padded and instances, where MFCC length is shorter than twenty are discarded. Furthermore, we treat Acc, Gyro and Aud as  $m$ -channels inputs (3, 3, and 13 channels, respectively) as it allows us to efficiently learn independent factors from every sensor axis, thus maximally utilizing the large-scale dataset.

### Evaluation and Metrics

Our models are evaluated with five-folds cross-validation with the same divisions of sixty users as of (Vaizman, Weibel, and Lanckriet 2018), where training and test folds contain 48 and 12 users, respectively. For hyper-parameter optimization, we use nested cross-validation (Cawley and Talbot 2010) by randomly dividing the training fold data



into training and validation sets with ratio of 80-20. After hyper-parameters selection, we train our models on the complete dataset of training folds (individually, each time from scratch) and calculate metrics on the testing folds. Furthermore, it is mentioned earlier that the considered dataset is highly imbalanced with sparse labels. In this case, simply calculating naive accuracy will be misleading due to not taking underrepresented classes into account. Similarly, precision and f1-score are also very likely to be affected by the class-skew due to involvement of true positives in the denominator. Hence, we adopt a metric named balanced accuracy (BA) (Brodersen et al. 2010) as used in (Vaizman, Weibel, and Lanckriet 2018), which incorporates both recall (or true positive rate) and true negative rate:  $BA = \frac{Sensitivity + Specificity}{2}$ . BA can be interpreted as average accuracy achieved on either class (positive or negative regarding binary classification). It stays identical to traditional accuracy, if a model performs equally well on each class but drops to a random chance (i.e. 50%) if a classifier performs poorly on a class with few instances (Brodersen et al. 2010). We calculate BA for each label independently and average them afterwards to get a trustworthy score of the model’s overall performance.

## Results and Analysis

**Analysis of Fusing Multi-Modal Representations:** We quantify the effect of different procedures for getting a fixed dimension feature vector from each modality-specific network and examine their fusion through different configurations of the shared network. It is important to note that, we keep an entire network’s configuration same but only the layers under consideration are changed. Table 1 provides the averaged (metrics) scores over 51 contextual labels and 5-folds as a result of applying global (max and average) pooling, using FC layer or simply feeding the extracted representations to the shared network for further processing. For the latter, we explore learning mutual representation from Acc, Gyr, and Aud/MFCC through an additional standard convolution layer and compare its performance with directly using flattened representations. Our experiments suggest that global max pooling (GMP) over each modality’s features outperforms other utilized techniques; achieving BA of 0.750 with a sensitivity rate of 0.767. We believe the reason is that, GMP is capable of picking-up high-level shift-invariant features, which are most discriminative among others. Figure 5 presents per label metrics for this network on all the 51 labels in the dataset. Specifically, we notice majority of the labels have BA score in range of 70%-80%.

**Comparison of Convolution Variants:** We evaluate the complete multi-stream model through replacing only DPS-Conv layers with standard convolution (Std-Conv) in modality-specific networks. We did not observe major performance differences between the two models as shown in Table 2. Nevertheless, a model with DPS-Conv should be preferred because of having lower computational cost than Std-Conv (Sandler et al. 2018).

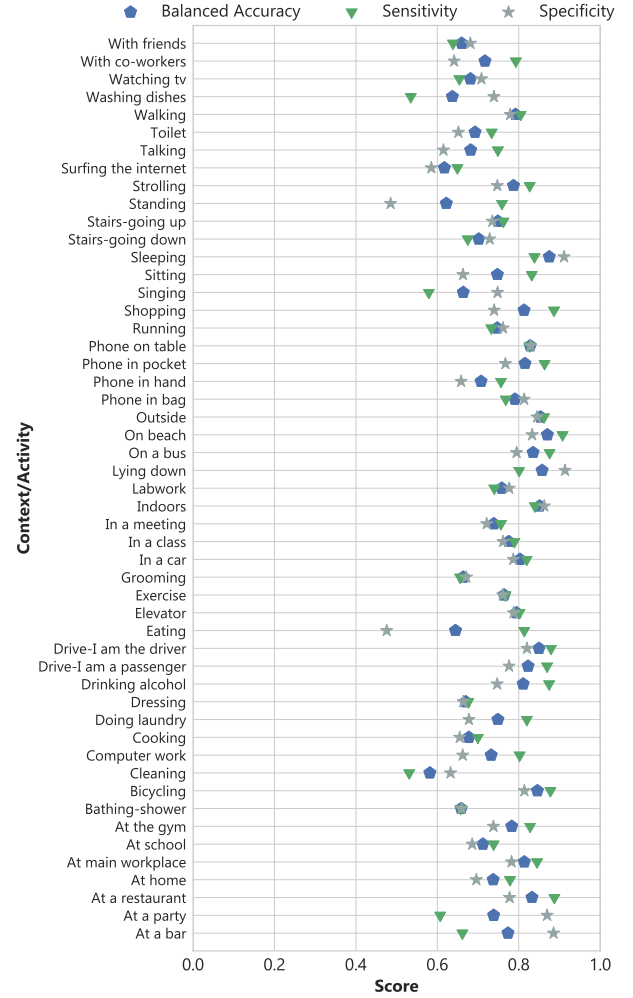


Figure 5: **Performance metrics per label of the best performing model (with GMP):** The scores are averaged over 5-folds cross-validation.

Table 1: **Multi-modal context recognition:** The metrics are reported for 5-folds cross-validation averaged over 51 class labels. BA stands for balanced accuracy.

	BA	Sensitivity	Specificity
GMP	0.750 ( $\pm 0.012$ )	0.767 ( $\pm 0.015$ )	0.733 ( $\pm 0.016$ )
GAP	0.748 ( $\pm 0.009$ )	0.753 ( $\pm 0.012$ )	0.742 ( $\pm 0.015$ )
FC	0.744 ( $\pm 0.009$ )	0.735 ( $\pm 0.014$ )	0.753 ( $\pm 0.008$ )
Flattened	0.742 ( $\pm 0.014$ )	0.734 ( $\pm 0.029$ )	0.749 ( $\pm 0.007$ )
Conv	0.738 ( $\pm 0.011$ )	0.725 ( $\pm 0.022$ )	0.752 ( $\pm 0.022$ )

Table 2: **Performance evaluation with different convolution layers.**

	BA	Sensitivity	Specificity
Std-Conv	0.751 ( $\pm 0.011$ )	0.750 ( $\pm 0.017$ )	0.751 ( $\pm 0.007$ )
DPS-Conv	0.750 ( $\pm 0.012$ )	0.767 ( $\pm 0.015$ )	0.733 ( $\pm 0.016$ )

**Quantifying Modality Influence:** To examine the effect of different combinations of sensors (or features learned

from them) on the recognition capability of the model, we experimented with training several networks with modified architectures. Specifically, in this case the model only consisted of layers that are relevant to the signals under consideration e.g. for evaluating models with only Acc, Aud, and PS, we removed the Gyro network entirely and then trained it end-to-end from scratch. Table 3 shows the evaluation results that highlights the importance of joint-learning and fusion of multiple modalities to improve detection rate.

Table 3: **Effect of different modalities on recognition performance.**

	BA	Sensitivity	Specificity
Acc	0.633 ( $\pm 0.011$ )	0.668 ( $\pm 0.027$ )	0.599 ( $\pm 0.017$ )
Gyro	0.639 ( $\pm 0.011$ )	0.638 ( $\pm 0.017$ )	0.640 ( $\pm 0.020$ )
Aud	0.669 ( $\pm 0.024$ )	0.731 ( $\pm 0.028$ )	0.608 ( $\pm 0.025$ )
PS	0.712 ( $\pm 0.005$ )	0.723 ( $\pm 0.011$ )	0.700 ( $\pm 0.013$ )
Acc, Gyro, PS	0.733 ( $\pm 0.010$ )	0.744 ( $\pm 0.021$ )	0.722 ( $\pm 0.014$ )
Acc, Gyro, Aud	0.708 ( $\pm 0.010$ )	0.722 ( $\pm 0.027$ )	0.693 ( $\pm 0.012$ )
Acc, Aud, PS	0.745 ( $\pm 0.013$ )	0.757 ( $\pm 0.025$ )	0.733 ( $\pm 0.015$ )
Gyro, Aud, PS	0.748 ( $\pm 0.012$ )	0.768 ( $\pm 0.014$ )	0.728 ( $\pm 0.014$ )
All	0.750 ( $\pm 0.012$ )	0.767 ( $\pm 0.015$ )	0.733 ( $\pm 0.016$ )

**Fusion and Effect of Missing Sensors:** We now evaluate the modified architecture’s predictive performance (presented in Section Missing Sensors), confronting various combinations of missing signals. Table 4 provides experimental results showing that the proposed multi-task network can handle lost modalities, achieving similar BA score as when separate models for each modality are developed (see Table 3). However, this flexibility comes at the price of slightly lower BA but makes a model capable of operation in the face of unavailable sensors.

Table 4: **Assessment of multi-task network for handling missing modalities.** Each row provide averaged metrics score as earlier but only mentioned modalities that are used for determining user’s context.

	BA	SN	SP
Acc	0.634 ( $\pm 0.008$ )	0.652 ( $\pm 0.027$ )	0.616 ( $\pm 0.013$ )
Gyro	0.619 ( $\pm 0.016$ )	0.632 ( $\pm 0.040$ )	0.606 ( $\pm 0.023$ )
Aud	0.656 ( $\pm 0.026$ )	0.670 ( $\pm 0.046$ )	0.641 ( $\pm 0.015$ )
PS	0.688 ( $\pm 0.009$ )	0.709 ( $\pm 0.015$ )	0.667 ( $\pm 0.012$ )
Acc, Gyro	0.646 ( $\pm 0.009$ )	0.670 ( $\pm 0.028$ )	0.622 ( $\pm 0.018$ )
Acc, Aud	0.687 ( $\pm 0.015$ )	0.695 ( $\pm 0.035$ )	0.679 ( $\pm 0.008$ )
Acc, PS	0.708 ( $\pm 0.007$ )	0.713 ( $\pm 0.015$ )	0.702 ( $\pm 0.012$ )
Gyro, Aud	0.687 ( $\pm 0.020$ )	0.699 ( $\pm 0.045$ )	0.676 ( $\pm 0.015$ )
Gyro, PS	0.708 ( $\pm 0.007$ )	0.719 ( $\pm 0.023$ )	0.696 ( $\pm 0.019$ )
Aud, PS	0.708 ( $\pm 0.013$ )	0.717 ( $\pm 0.027$ )	0.698 ( $\pm 0.010$ )
Acc, Gyro, Aud	0.690 ( $\pm 0.012$ )	0.703 ( $\pm 0.031$ )	0.677 ( $\pm 0.011$ )
Acc, Gyro, PS	0.705 ( $\pm 0.007$ )	0.714 ( $\pm 0.023$ )	0.696 ( $\pm 0.019$ )
Acc, Aud, PS	0.721 ( $\pm 0.007$ )	0.729 ( $\pm 0.019$ )	0.712 ( $\pm 0.011$ )
Gyro, Aud, PS	0.721 ( $\pm 0.011$ )	0.730 ( $\pm 0.030$ )	0.711 ( $\pm 0.017$ )
All	0.720 ( $\pm 0.008$ )	0.728 ( $\pm 0.025$ )	0.712 ( $\pm 0.015$ )

**Reliance on Instance Weighting and Regularization:** Our results thus far have been obtained through training a model with cross-entropy loss. This incorporated instance-

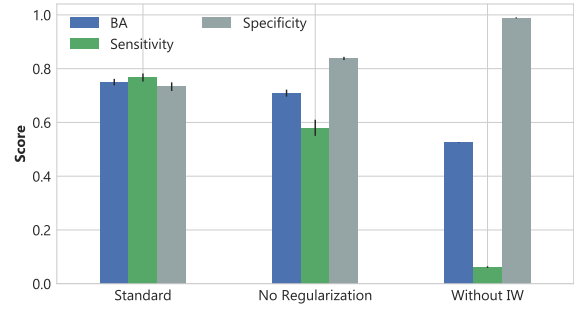


Figure 6: **Assessment of instance-weighting and regularization:** We determine the impact of cost sensitive loss function and regularization (i.e. weight decay and dropout) on the network’s predictive power. The results labeled under standard are with both IW and regularization.

weights to handle class-imbalance. To test network’s dependence on the cost sensitive loss function ( $\mathcal{J}_c$ ), we examined a model’s performance that is trained without it. As expected, the overall BA score drastically drops to a random chance (see Figure 6) with worse performance on positive samples in comparison with the negative ones. Likewise, we also trained a model without any sort of regularization i.e. removing dropout, L1 and L2 penalties from the network. The average recall rate on the held-out testing folds dropped to 0.58 which can be an indication of overfitting the training set. Hence, incorporating both instance-weighting (IW) and regularization improved performance significantly in learning from this imbalanced dataset. However, further work will be necessary to investigate other techniques for managing (sparse) rare labels such as oversampling and data augmentation in case of multi-labeled instances.

**Visualization:** In order to illustrate the semantic relevance of the learned features, we applied t-SNE (van der Maaten and Hinton 2008) to project high-dimensional data to 2D embedding. We take the output of the last FC layer (see Figure 3) from the shared network by feeding a limited (but randomly selected) subset of the dataset to extract the embeddings. Further, as the data under consideration is multi-labeled, we identified sets of mutually-exclusive labels (e.g. Indoors vs. Outside) that can be used to color code the data points to visually identify meaningful clusters. Figure 7 provides a visualization for various sets of labels suggesting the network is able to disentangle possible factors of variation that may distinguish a class from the rest in large-scale sensory data. Furthermore, to get better insights in the diversity of the extracted features from each modality, in Figure 8, we visualize the feature maps produced by the first layer of the DPS-Conv layer of modal-specific networks.

## Conclusions

In this work, we tackled the problem of multi-label behavioral context recognition with deep multi-modal convolutional neural networks. We propose to train an end-to-end model for jointly-learning from low-level sensory data (accelerometer, gyroscope, audio and phone state) of

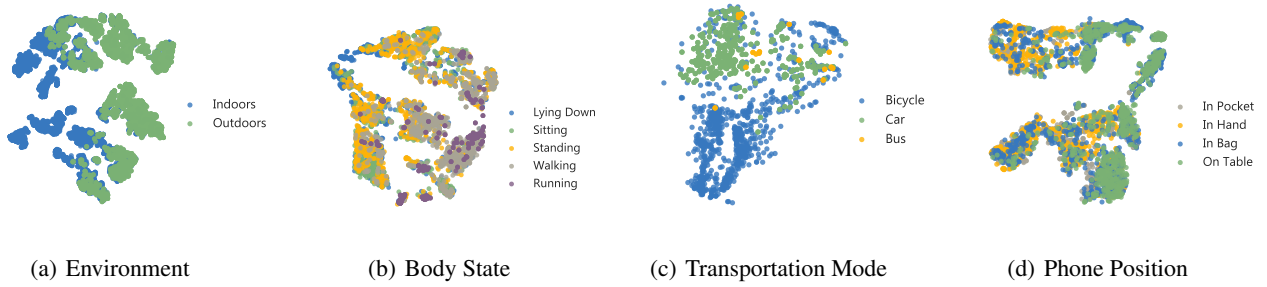


Figure 7: **t-SNE embeddings:** We visualize the mutual features learned through fusion of multiple modalities (from the last layer) in the shared network. Four sets of mutually-exclusive labels are identified from multi-labeled data to use during final visualization of semantically related clusters extracted through t-SNE.

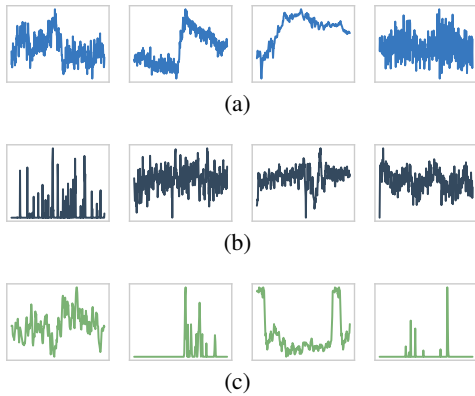


Figure 8: **Feature Maps from Modality-Specific Networks:** Illustration of randomly selected (learned) features from first layer of convolutional networks. (a), (b) and (c) represent outputs from Acc, Gyro and Aud models, respectively.

smart devices collected in-the-wild. Our empirical results demonstrated various strategies for feasibly fusing representations learned from different modalities and quantifying their contribution on the predictive performance. We also showed that instance-weighted cross-entropy loss (as also leveraged in (Vaizman, Weibel, and Lanckriet 2018)) and regularization schemes enable the model to generalize well on highly imbalanced (sparsely labeled) dataset. Furthermore, we present a slight modification in the proposed network’s architecture to handle missing sensors; potentially taking advantage of multi-task learning. We believe, the proposed methodology is generic enough and can be applied to other related problems of learning from multivariate time series. Additionally, potential directions for future work would involve developing techniques to handle imbalanced multi-label data, optimal sensor selection to reduce computation and battery consumption, and incorporating other analogous sensors to further improve the detection rate.

**Acknowledgment** SCOTT (www.scott-project.eu) has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737422. This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and Aus-

tria, Spain, Finland, Ireland, Sweden, Germany, Poland, Portugal, Netherlands, Belgium, Norway.

Various icons used in the figures are created by Anuar Zhu-maev, Tim Madle, Korokoro, Gregor Cresnar, Shmidt Sergey, Hea Poh Lin, Natalia Jacquier, Trevor Dsouza, Adrien Coquet, Alina Oleynik, Llisole, Alena, AdbA Icons, Jeevan Kumar, Artd-abana@Design, lipi, Alex Auda Samora, and Michelle Colonna from the Noun Project.

## References

- [Abadi et al. 2016] Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, 265–283.
- [Alsheikh et al. 2016] Alsheikh, M. A.; Selim, A.; Niyato, D.; Doyle, L.; Lin, S.; and Tan, H.-P. 2016. Deep activity recognition models with triaxial accelerometers. In *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*.
- [Bai, Kolter, and Koltun 2018] Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- [Bengio, Courville, and Vincent 2013] Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.
- [Bhattacharya et al. 2014] Bhattacharya, S.; Nurmi, P.; Hammerla, N.; and Plötz, T. 2014. Using unlabeled data in a sparse-coding framework for human activity recognition. *Pervasive and Mobile Computing* 15:242–262.
- [Brodersen et al. 2010] Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; and Buhmann, J. M. 2010. The balanced accuracy and its posterior distribution. In *Pattern recognition (ICPR), 2010 20th international conference on*, 3121–3124. IEEE.
- [Caruana 1997] Caruana, R. 1997. Multitask learning. *Machine learning* 28(1):41–75.
- [Cawley and Talbot 2010] Cawley, G. C., and Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11(Jul):2079–2107.
- [Chollet 2017] Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.



- [Figo et al. 2010] Figo, D.; Diniz, P. C.; Ferreira, D. R.; and Cardoso, J. M. 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* 14(7):645–662.
- [Glorot and Bengio 2010] Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.
- [Hammerla, Halloran, and Ploetz 2016] Hammerla, N. Y.; Halloran, S.; and Ploetz, T. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*.
- [Hoseini-Tabatabaei, Gluhak, and Tafazolli 2013] Hoseini-Tabatabaei, S. A.; Gluhak, A.; and Tafazolli, R. 2013. A survey on smartphone-based systems for opportunistic user context recognition. *ACM Computing Surveys (CSUR)* 45(3):27.
- [Jiang and Yin 2015] Jiang, W., and Yin, Z. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1307–1310. ACM.
- [Kaiser, Gomez, and Chollet 2017] Kaiser, L.; Gomez, A. N.; and Chollet, F. 2017. Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*.
- [Lane, Georgiev, and Qendro 2015] Lane, N. D.; Georgiev, P.; and Qendro, L. 2015. Deeppear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 283–294. ACM.
- [Lara and Labrador 2013] Lara, O. D., and Labrador, M. A. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials* 15(3):1192–1209.
- [Lee et al. 2013] Lee, Y.; Min, C.; Hwang, C.; Lee, J.; Hwang, I.; Ju, Y.; Yoo, C.; Moon, M.; Lee, U.; and Song, J. 2013. Socio-phone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 375–388. ACM.
- [Lin et al. 2012] Lin, M.; Lane, N. D.; Mohammad, M.; Yang, X.; Lu, H.; Cardone, G.; Ali, S.; Doryab, A.; Berke, E.; Campbell, A. T.; et al. 2012. Bewell+: multi-dimensional wellbeing monitoring with community-guided user feedback and energy optimization. In *Proceedings of the conference on Wireless Health*, 10. ACM.
- [Lin et al. 2015] Lin, Q.; Zhang, D.; Connelly, K.; Ni, H.; Yu, Z.; and Zhou, X. 2015. Disorientation detection by mining gps trajectories for cognitively-impaired elders. *Pervasive and Mobile Computing* 19:71–85.
- [Lorincz et al. 2009] Lorincz, K.; Chen, B.-r.; Challen, G. W.; Chowdhury, A. R.; Patel, S.; Bonato, P.; Welsh, M.; et al. 2009. Mercury: a wearable sensor network platform for high-fidelity motion analysis. In *SenSys*, volume 9, 183–196.
- [Miluzzo et al. 2008] Miluzzo, E.; Lane, N. D.; Fodor, K.; Peterson, R.; Lu, H.; Musolesi, M.; Eisenman, S. B.; Zheng, X.; and Campbell, A. T. 2008. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, 337–350. ACM.
- [Ordóñez and Roggen 2016] Ordóñez, F. J., and Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multi-modal wearable activity recognition. *Sensors* 16(1):115.
- [Plötz, Hammerla, and Olivier 2011] Plötz, T.; Hammerla, N. Y.; and Olivier, P. 2011. Feature learning for activity recognition in ubiquitous computing. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, 1729.
- [Rabbi et al. 2015] Rabbi, M.; Aung, M. H.; Zhang, M.; and Choudhury, T. 2015. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 707–718. ACM.
- [Radu et al. 2018] Radu, V.; Tong, C.; Bhattacharya, S.; Lane, N. D.; Mascolo, C.; Marina, M. K.; and Kawsar, F. 2018. Multi-modal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1(4):157.
- [Rashidi and Cook 2009] Rashidi, P., and Cook, D. J. 2009. Keeping the resident in the loop: Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans* 39(5):949–959.
- [Rashidi and Mihailidis 2013] Rashidi, P., and Mihailidis, A. 2013. A survey on ambient-assisted living tools for older adults. *IEEE journal of biomedical and health informatics* 17(3):579–590.
- [Ronao and Cho 2016] Ronao, C. A., and Cho, S.-B. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications* 59:235–244.
- [Sandler et al. 2018] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.
- [Shoaib et al. 2015] Shoaib, M.; Bosch, S.; Incel, O. D.; Scholten, H.; and Havinga, P. J. 2015. A survey of online activity recognition using mobile phones. *Sensors* 15(1):2059–2085.
- [Srivastava et al. 2014] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.
- [Vaizman, Ellis, and Lanckriet 2017] Vaizman, Y.; Ellis, K.; and Lanckriet, G. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing* 16(4):62–74.
- [Vaizman, Weibel, and Lanckriet 2018] Vaizman, Y.; Weibel, N.; and Lanckriet, G. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1(4):168.
- [van der Maaten and Hinton 2008] van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.
- [Yang et al. 2015] Yang, J.; Nguyen, M. N.; San, P. P.; Li, X.; and Krishnaswamy, S. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*, 3995–4001.
- [Zeng et al. 2014] Zeng, M.; Nguyen, L. T.; Yu, B.; Mengshoel, O. J.; Zhu, J.; Wu, P.; and Zhang, J. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on*, 197–205. IEEE.
- [Zhang et al. 2017] Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2017. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*.