Human Activity Recognition based on Wearable Sensors using Multiscale DCNN Ensemble

Jessica Sena[1] and William Robson Schwartz
Smart Sense Laboratory, Computer Science Department
Universidade Federal de Minas Gerais, Minas Gerais, Brazil

***Abstract*—Sensor-based Human Activity Recognition (HAR) provides valuable knowledge to many areas. Recently, wearable devices have gained space as a relevant source of data. However, there are two issues: large number of heterogeneous sensors available and the temporal nature of the sensor data. To handle these issues, we propose a multimodal approach that processes each sensor separately and, through an ensemble of Deep Convolution Neural Networks (DCNN), extracts information from multiple temporal scales of the sensor data. In this ensemble, we use a convolutional kernel with a different height for each DCNN. Considering that the number of rows in the sensor data reflects the data captured over time, each kernel height reflects a temporal scale from which we can extract patterns. Consequently, our approach is able to extract information from simple movement patterns such as a wrist twist when picking up a spoon, to complex movements such as the human gait. This multimodal and multi-temporal approach outperforms previous state-of-the-art works in seven important datasets using two different protocols. In addition, we demonstrate that the use of our proposed set of kernels improves sensor-based HAR in another multi-kernel approach, the widely employed inception network[1].**

I. Introduction

The use of sensors from wearable devices to recognize human activities has grown every year. As discussed by Lara et al. [1], there are many reasons for this growth: the increasing interest of several areas, such as, medical, military, and security applications; the convenience and comfort of using such devices (it does not change or hinders the action due to their use); the feeling of privacy (as opposed to monitoring with cameras where depending on the activity performed or the location, the user feels uncomfortable); and it is already naturally inserted into people's lives, facilitating the data capture. The number of sensors in such devices is increasing and the large range of sensors provide rich and complementary information regarding the activities performed by users. Therefore, an important line of research that has gained attention focuses on the investigation to combine (i.e., fuse) these multiple sensors to improve human activity recognition.

Some works perform fusion in the raw data (i.e., early fusion), concatenating the sensors into a common matrix used as input for machine learning methods. For instance, Chen and Xue [2] employed a Deep Convolutional Neural Network (DCNN) with three convolutional layers and used the size of the kernel to extract the relation between the axes and temporal information. Motivated by the architecture proposed in [2], Jordao et al. [3] suggested a DCNN able to explore the patterns among the signal axes in all the layers that compose the network. As a consequence, their proposed DCNN achieved better results than [2]. Different from [2] and [3], Jordao et al. [4] employed a DCNN and use partial least squares analysis to reduce the dimensionality of each max-pooling layer and consider the concatenation of the dimension reduction as a feature to feed a softmax classifier. To improve the data representation, Jiang and Yin [5] applied a discrete Fourier transform to pre-process the input matrix and use a DCNN composed by a stack of two convolutional layers, a fully connected and a softmax layer to recognize the activities. However, due to the multimodal nature of each sensor, merging the sensors in the raw data may not be appropriate since sensors have several dissimilarities between them, such as a different number of axes, scales, meanings, or data nature (e.g., angle, intensity or frequency).

To address the multimodality problem, some authors proposed to insert a padding between the sensors to separate the data and to be able to extract features from the sensors separately. For instance, Ha et. al. [6] preprocessed the matrix of sensors adding a zero-padding between each sensor and use a DCNN with the same layer structure as in [5]. However, this division is only effective at the first layer since, from the second layer onwards, the data from different sensors are convoluted together. In fact, in another work, Ha and Choi [7] proposed to insert zero-padding before each convolutional layer to avoid interference between sensors when 2D convolutional kernel is applied. While this approach separates in some way the data before performing fusion, it uses the same DCNN to learn features from all sensors simultaneously, which might overcharge the model since the kernel have to learn patterns from different data nature.

In a recent work, Yao et al. [8] brought a new perspective to the problem by merging multimodal data to perform sensor-based HAR. They build an architecture with three different sequential blocks: an individual deep convolutional subnet for each input sensor to learn local patterns; a common deep convolutional subnet that concatenates all sensors and learns the high-level relationship among them; and, at the end of the architecture, a stacked Gated Recurrent Unit [9] structure to learn meaningful temporal features. Since the use of convolutional and recurrent networks is already well established in the sensor community, the main advance of [8] is to go beyond just placing a boundary between the sensors in the input matrix.

[1]This work corresponds to an M.Sc. dissertation.

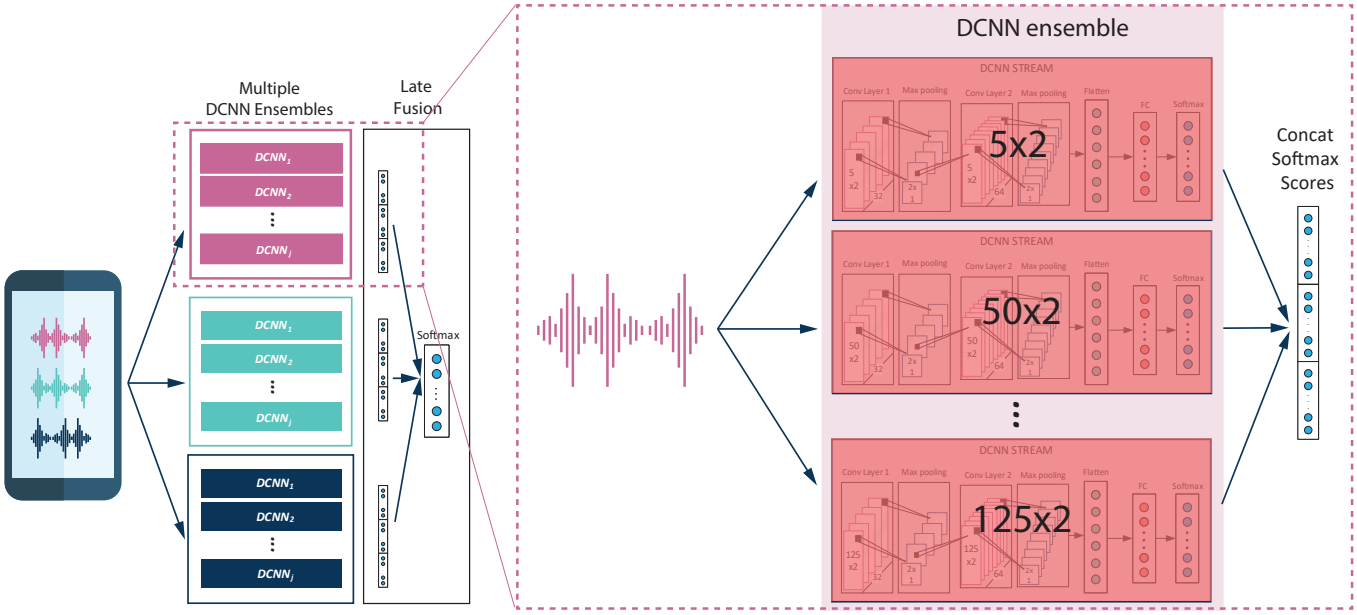

Fig. 1. Our approach, the Multimodal DCNN Ensemble (MDE) relies on two premises. The first is separately processing each sensor and the second is to extract patterns from multiple temporal scales. Thus, for each sensor, we create a DCNN ensemble that extracts multi-temporal information. This ensemble is composed of streams so that each one extracts patterns on a specific temporal scale and classifies the sample. We merge all scores into a late fusion approach which allows us to take advantage of the complementarity between both sensors and temporal scales.

Instead, before merging the data from multiple sensors (i.e., exploit the complementarity of the sensors), they separate the sensors to extract features individually to learn which patterns from each sensor better classify human activities.

Besides the sensor data heterogeneity, another issue that must be considered is the temporal nature of the data. Due to the CNN input format for sensors (where columns refer to the sensor axes and rows to data-capture over time), the height of the convolutional kernel represents the size of the temporal window used to learn patterns. Since there are several possibilities to set the kernel height, we can see each size as a temporal scale to extract potential patterns.

In traditional deep convolutional network methods [2], [3], [5], a single kernel size is set for each layer, which discards all other possible temporal scales for that particular layer. In these networks, each stacked convolutional layer learns features at a larger semantic level than the previous one and, in the sensor context, a deeper CNN network would learn features in multiple temporal scales due to its depth (each layer learns a higher temporal scale than the previous one). However, the convolutional maps that go to the next layer are the activations for the kernel in the previous layer. In this way, when one chooses a single kernel size for a specific layer, it might discard important information in this layer which would only be selected by another kernel size. Therefore, to avoid this problem, we propose the use of an ensemble of multiple kernels which is able to learn several temporal scales simultaneously. This follows the intuition that human activities are composed by different durations, i.e., while some activities can only be distinguished by small and fast movements, others need to be analyzed for longer periods of time to be classified.

Given the aforementioned issues, we propose an approach based on multiple streams to individually process the sensor data. The core of this approach is a novel way of extracting temporal data by employing an ensemble of temporal scales implemented with multiple DCNNs. As each DCNN has a kernel size which reflects one scale of a pre-defined temporal scale range, we can extract patterns of multiple sizes, ranging from short movements, such as a gentle twist of the wrist, to large and complex motions, such as the human gait. To the best of our knowledge, this is the first work to propose extract patterns on sensor data using multiple scales to capture multiples movements.

According to experimental results, our approach outperforms previous state-of-the-art results in seven datasets using two different evaluation protocols. In addition, we adapt the Inception module [10] to compare to our DCNN Ensemble approach (without multimodal premise) and we demonstrate that our method is better than the Inception. We also show that the use our kernel set is more suitable for the sensor-based HAR than the kernels originally proposed in the Inception module.

II. Proposed Approach

In this work, we propose an approach, called *Multimodal DCNN Ensemble (MDE)*, to recognize human activities using data provided by smartphones and smartwatches. It is based on two hypotheses: (i) the use of multiples sensors might improve accuracy due to the complementarity between the sensors, (ii) activities are best described using multiple temporal scales to extract patterns. To test these hypotheses, our approach

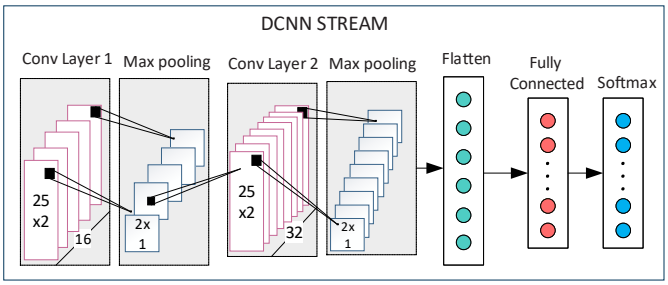

Fig. 2. The deep convolution neural network stream.

consists of three main steps. First, we divide the sensors into different inputs to process each one individually. Then, for each sensor, we build an ensemble of temporal scales extracted through DCNN streams that are subnets within our network. Finally, we use an approach based on late fusion to merge the multi-modal and multi-temporal information. The following sections detail each step of this process. Figure 1 shows an overview of our method.

A. DCNN Stream

Our approach is an end-to-end neural network composed of subnets integrated through an ensemble technique. These subnetworks, for convenience, let us call them *streams*, are Deep Convolutional Neural Networks composed of two parts, as illustrated in Figure 2.

The first block of the stream is a convolutional block with two convolutional layers intercalated by two max-pooling layers. While the convolutional layers allow us to learn patterns in the temporal scale defined for each stream, the pooling layers control overfitting, reduce the number of parameters and the computation cost. At the end of the subnet, we have a fully-connected block consisting of a flatten layer, a fully-connected layer and a softmax layer. We use scaled exponential linear units (SELUs) [11] as the activation function of the fully connected block. SELU has self-normalizing properties which make the learning highly robust and allows to train networks that have many layers. Additionally, the learning speed is faster in SELU compared to the ReLU activation function as shown in the work of [12]. While the convolutional block provides a meaningful and invariant feature space, the fully-connected block learns a non-linear function in that translates the features learned by the convolutional block to the softmax scores.

B. DCNNs Ensemble

The sensor data is commonly stored in a matrix of size $t \times a$, where a is the number of axes of the sensor (for instance, three axes (x, y, z) on motion sensors) and t is the temporal axis, where each row is a sensor sample at a given time instant. Therefore, given a 2D kernel (h, w), our premise is that the height of the kernel (h) is responsible for determining in which temporal scale we are learning the patterns. For instance, a h equal to 25 in a matrix of 500 rows (a sample of size 5 captured at a frequency of 100Hz) learns patterns of 0.25 seconds while a h equal to 250 learns patterns of 2.5 seconds. Thus, the larger the kernel height, the larger the temporal pattern it captures.

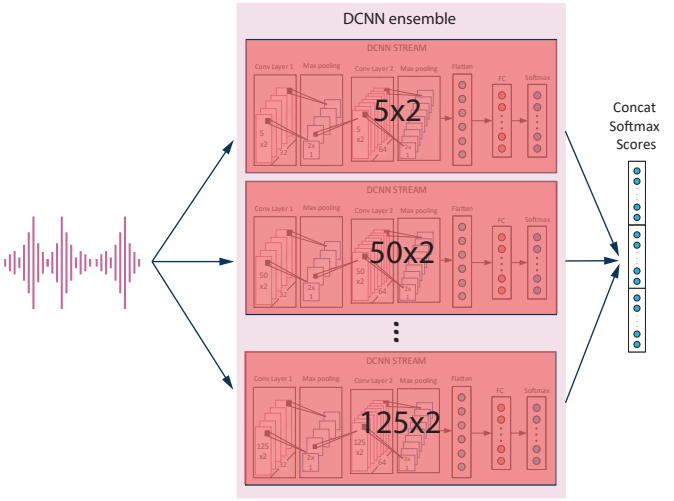

Fig. 3. The DCNN ensemble is composed of streams so that each stream extracts patterns from a specific temporal scale and classifies the sample for that scale. We merge all scores into a late fusion approach which allows us to take advantage of the complementarity between sensors and between temporal scales.

Based on the aforementioned premise, we employ an ensemble of deep convolutional streams with different kernel sizes to extract information from multiple temporal scales. The architecture of our multiscale ensemble is illustrated in Figure 3. The number of DCNNs in each ensemble is predefined as a parameter of our architecture, called *pool*. The pool is a set of kernels $K = \{K_1, K_2, ..., K_j\}$ containing a variety of kernel sizes ranging from a small to a large kernel. For each kernel in our pool, we add a DCNN in the ensemble and set its two convolutional layers with the specific kernel. For instance, in Figure 3, we have a pool of j kernels where three of them have their streams explicitly drawn in the figure composing a kernel pool $K = \{5 \times 2, 25 \times 2, ..., 250 \times 2\}$.

The multiscale ensemble is the most important contribution of this thesis since, to the best of our knowledge, we are the first to extract patterns on sensor data using multiple scales. As shown in the architecture of our main approach (Figure 1), an ensemble is built for each sensor, so we have several ensembles according to the number of sensors processed (in the example illustrated in Figure 1, we have three sensors and consequently, three ensembles of DCNNs).

C. Decision Level Fusion

At the end of the DCNN ensemble stage, we have an ensemble for each sensor, and each ensemble is composed of j streams. Thus, it is necessary to merge this information to take advantage of the complementarity provided by both the multiple sensors and the multiple temporal scales. According to our experimental results, the best way to merge these streams is by using meta-learning of the scores. Therefore, we concatenate all the scores of the streams ($j \times$ number of sensors) in a single feature vector and pass it to a classification layer (softmax).

The training of our network is performed in an end-to-end way, which optimizes the weights of the entire network since

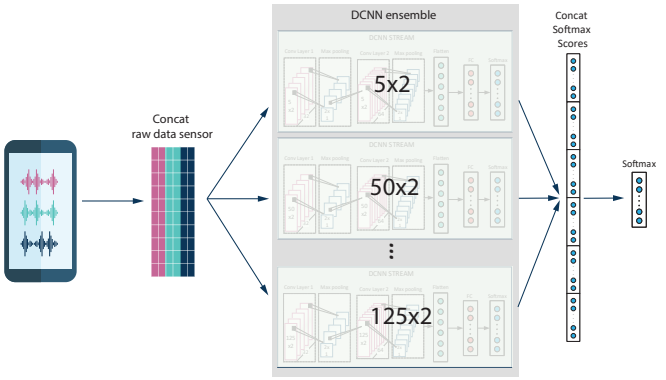

Fig. 4. DCNN Ensemble approach on previous concatenated sensors data.

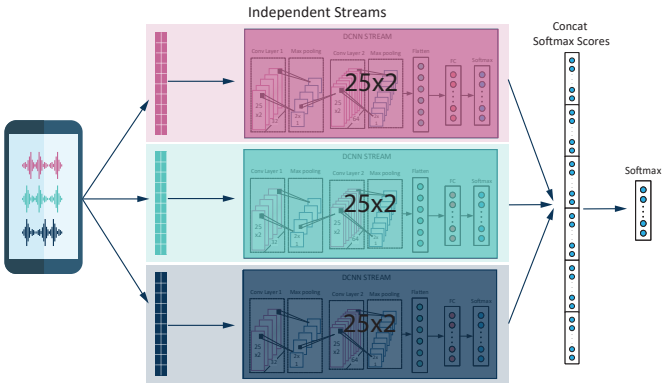

Fig. 5. Multimodal stream approach. A simplified version to evaluate one of the main steps of our MDE approach. This version is based on the premise of processing the sensors separately to extract meaningful features before fusion. We do not use the DCNN ensemble in this version, instead, we use only one stream to process each sensor and we set the kernel size as 25×2.

it maps the input of all the modalities to a single output. Consequently, the network dynamically learns which scales and sensors are most appropriate to classify each activity.

III. Experimental Results

To evaluate the contribution of each part of the proposed approach (i.e., the ensemble of convolutional neural networks and the individual processing of the sensors), we implemented two simplified versions of our proposed approach. The first, called *DCNN Ensemble*, is illustrated in Figure 4. In this version, we do not separate the sensors, instead, we concatenate all sensors into a single array, in the same way of the majority of works. Thus, we employ only a single ensemble of kernels since we have only one input. The goal is to measure the contribution of the multi-temporal scale approach implemented with the DCNN ensemble in a scenario without multimodal processing of sensors.

Figure 5 shows the second simplified version of our approach, called *Multimodal Stream*, that aims at measuring the contribution of individual processing of the sensors. In this version, we create a network following the multimodal hypothesis but without using the DCNN ensemble approach. Instead, we employ only a single DCNN stream (see Figure 2) for each sensor. In this DCNN stream, we empirically choose the value of 25×2 to set the kernel size.

We compare our approach and its simplified versions with all methods evaluated by Jordão et al. [13]. Thereby, in addition to the methods mentioned in Section I, we also show results from three other handcrafted methods [14]–[16] surveyed by Jordão et al. [13]. Usually, this family of methods extracts statistical features and applies a classifier to recognize activities. We include them in our evaluation mainly because they present better results in some datasets than approaches based on deep learning. Finally, due to the multimodal nature, we evaluate the MDE and Multimodal Stream only on datasets that contain more than one sensor.

A. Experimental Setup

One of the most latent problems in wearable sensor-based human activity recognition is the lack of standardization of metrics, evaluation protocols, and datasets, which makes it difficult to compare the methods in the literature. While some works record their own datasets to perform experiments, others use datasets from the literature but do not clarify the evaluation protocol employed, which prevents the reproducibility of the results. Recently, a work has endeavored to solve this issue by bringing the first standardization to the domain. Jordão et al. [13] performed a thorough study and standardized seven datasets of the wearable sensor literature in four different protocols.

As pointed by Jordão et al. [13], most of the works based on convolution networks omit some important parameters, hindering comparison between methods. To handle this issue, the protocol created by Jordão et al. [13] sets some parameters. The maximum number of epochs was set to 200 and the method stops its training when the loss function reaches a value less or equal than 0.2. These values were set empirically by observing the trade-off between execution time and accuracy. Regarding deep learning implementation decisions, we use *cross-entropy* as the loss function of our network. Cross-entropy measures the performance of a classification model whose output is a probability value between 0 and 1. The loss increases as the predicted probability diverges from the actual probability. We employ the RMSprop [17] as optimizer since it provides an efficient execution time. As dropout layer, we use alpha dropout [11] since it fits well to scaled exponential linear units by randomly setting activations to the negative saturation value. Alpha dropout keeps the mean and variance of inputs to their original values, to ensure the self-normalizing property even after dropout. We set the dropout rate to 0.1.

Regarding the ensemble implementation, it is important to mention that in the DCNN stream (see Figure 2) we use 16 filters in the first convolutional layer and 32 filters in the second. In addition, the results shown in our experiments were performed using $K = \{2 \times 2, 3 \times 3, 5 \times 2\, 12 \times 2\, 25 \times 2\}$ as our pool of kernels.

Jordão et al. [13] conducted a survey in the literature and gathered seven important datasets: WHARF [18],

USCHAD [19], UTD-MHAD (set 1 and 2) [20], WISDM [21], PAMAP2P [22] and MHEALTH [23]. This set of datasets composes an interesting diversity of number of samples, types of activities performed and number of available sensors, making it possible to evaluate the robustness of the methods in different scenarios. The datasets were processed and standardized with a sampling rate of five seconds, except for the UTD-MHAD dataset that had to be sampled at 1-second rate. We evaluate our approach in these seven datasets following strictly the procedure defined by Jordão et al. [13][2].

Regarding the evaluation protocols, according to Jordão et al. [13], Leave-One-Subject-Out (LOSO) and Leave-One-Trial-Out (LOTO) are the most appropriate for reporting results in sensor-based HAR. In the LOSO protocol, the data are separated in training and test so that the test has only one subject at a time and the training has the other subjects. In the LOTO, the trial consists of a transition from one activity to another, so the data is separated into trials where each trial contains only a continuous capture of an activity. Therefore, the training is performed with all the trials except one that is set as test. LOSO represents the real scenario of applications for wearables devices, where a method is trained in known subjects and applied to new subjects afterwards. This protocol also analyzes the generalization quality of the method since the training and test data do not have the same distribution. On the other hand, LOTO protocol has the benefits of generating a large number of samples and certifying that the contents of a trial do not appear in training and testing at the same time, different from the cross-validation protocols inappropriately used in the literature, which ensures a correct evaluation of the performance.

B. Comparison with a Kernel Ensemble Baseline

To analyze the contribution of the pool of kernels and to evaluate the contribution of our DCNN ensemble, we use the Inception network module proposed by Szegedy et al. [10] as a baseline. Although the inception was originally designed for object detection in images, it is analogous to our approach since it also applies multiple kernels to the same input to extract different pattern sizes. We were not able to compare our DCNN Ensemble with inception's full architecture because the available datasets to sensor-based HAR do not have enough data to train a network the size of inception (in the object detection domain the inception was trained using 1.2 million of images provided by ImageNet dataset [24], in our context the dataset with the largest number of samples used in our evaluation has 20,000 samples). One option would be to use the network pre-trained on the ImageNet and perform transfer learning. However, the pre-trained model is restricted to the use of three channels and requires a minimum array of 139x139 pixels as input. The sensor data is composed of one channel and our largest dataset has a matrix of 500x10. Therefore, it is not possible to use the pre-trained inception network without deforming our data.

[2]Refer to [13] for more details regarding the evaluation procedure.

Due to the aforementioned restrictions, we performed a study of the appropriate number of inception modules that should be used for the context of wearable sensors. Our experiments showed that the addition of more than one module deteriorates the results. Therefore, all inception-based experiments in this work were executed by using only one inception module. Another important point is that we add the fully connected block used in our stream to the inception module. This considerably increased the inception performance since the fully connected block is capable of fusing different patterns extracted by different kernels sizes and also regularize the network since we use SELU activation function. We employed two modules proposed by Szegedy et al. [10]: the naïve and the dimensionality reduction module as baselines. In addition, to evaluate our kernel pool, we adapt each type of inception module to the wearable sensors domain by using the same pool of kernels used by our approach instead of the kernels proposed in Szegedy et al. [10].

Table I shows the results obtained with the described approaches on LOTO and LOSO evaluation protocols. According to the results, the use of our pool of kernels improves the result of the inception original modules for all datasets. This support our hypothesis that extracting multiple temporal scales is appropriate for the sensor domain. Besides, our DCNN ensemble approach outperforms all four inception-based methods using LOSO and LOTO on the seven datasets, which points out that our ensemble is more suitable to employ multiple kernels to extract temporal information in the context of wearables sensors.

C. Comparison with a Multimodal Baseline

Yao et al [8] brought advances to sensor fusion by employing multiple streams to process each sensor separately. To the best of our knowledge, that is the only multimodal method using multiple streams that have been proposed so far in the context of wearables sensors. Our multimodal stream and MDE explore this intuition. It is important to note that due to the multimodal premise of the approaches, we do not evaluate the work of [8] and our multimodal approaches (MDE and Multimodal Stream) on WHARF and WISDM datasets since they consider only the accelerometer sensor.

The approach proposed by Yao et al [8] shows poor results both on LOTO and LOSO (see Table I) protocols reporting accuracy lower than very simple approaches such as handcrafted methods in all datasets evaluated. Particularly, their approach performs poorly in UTD-MHAD family and MHEALTH datasets. We believe this is because the network proposed by Yao et al [8] has a very complex structure which can cause overfitting since these datasets do not have a large number of samples. In addition, in the datasets of the UTD-MHAD family, the sample size does not allow it to be divided into time-steps to fed the network, which is essential to the approach of Yao et al [8] since it uses recurrent neural network (RNN).

In contrast to the approach proposed by Yao et al [8], our method showed superior results even using only the

TABLE I

COMPARISON OF OUR MULTIMODAL DCNN ENSEMBLE (MDE) AND ITS SIMPLIFICATIONS (DCNN ENSEMBLE AND MULTIMODAL STREAM) TO THE STATE-OF-THE-ART ARCHITECTURES SURVEYED BY [13] USING LOTO AND LOSO PROTOCOLS ON SEVEN DATASETS. ALSO, WE SHOW THE RESULTS OF TWO INCEPTION MODULES [10] USING THE ORIGINAL PROPOSED KERNELS AND OUR PROPOSED POOL OF KERNELS. CELLS WITH THE SYMBOL "-" DENOTES THAT IT IS NOT POSSIBLE TO EXECUTE THE METHOD ON THE RESPECTIVE DATASET DUE TO ITS ARCHITECTURE.

	WHARF	**UTD-1**	**UTD-2**	**WISDM**	**USCHAD**	**MHEALTH**	**PAMA**	**WHARF**	**UTD-1**	**UTD-2**	**WISDM**	**USCHAD**	**MHEALTH**	**PAMA**
METHODS	LOTO (ACCURACY (%))							LOSO (ACCURACY (%))						
Kwapisz et al. [14]	44.51	15.99	69.61	79.08	76.52	89.75	70.58	42.19	13.04	66.67	75.31	70.15	90.41	71.27
Catal et al. [15]	64.84	47.80	81.37	80.52	87.77	91.84	81.03	46.84	32.45	74.67	74.96	75.89	94.66	85.25
Kim et al. [16]	61.12	50.98	75.27	56.26	85.70	91.51	78.08	51.48	38.05	64.60	50.22	64.20	93.90	78.08
Chen and Xue [2]	72.55	-	-	86.55	84.66	89.95	82.32	61.94	-	-	83.89	75.58	88.67	83.06
Jiang and Yin [5]	70.79	-	-	83.82	80.73	52.78	-	65.35	-	-	79.97	74.88	51.46	-
Ha et al. [6]	-	-	-	-	-	85.31	80.13	-	-	-	-	-	88.34	73.79
Ha and Choi [7]	-	-	-	-	-	82.75	71.19	-	-	-	-	-	84.23	74.21
Yao et al. [8]	×	12.70	22.41	×	81.34	31.35	70.59	×	11.45	22.40	×	71.52	31.88	72.61
Inception naïve mod [10]	43.98	50.87	76.27	83.02	-	-	-	36.64	40.71	72.55	78.64	-	-	-
Inception naïve + pool	49.86	53.06	76.71	84.89	-	-	-	41.14	41.44	72.46	81.99	-	-	-
Inception mod [10]	51.76	52.36	74.62	79.18	-	-	-	42.07	39.62	68.34	73.86	-	-	-
Inception + pool	60.74	56.66	78.62	86.83	-	-	-	49.97	42.23	72.96	80.99	-	-	-
DCNN Ensemble	**75.50**	62.03	81.63	**89.01**	88.49	**93.09**	**83.99**	**69.79**	46.75	79.38	**86.22**	82.66	**96.27**	**87.59**
Multimodal Stream	×	48.90	79.82	×	85.95	83.17	79.62	×	36.99	74.59	×	79.68	90.20	80.58
MDE	×	**69.61**	**83.78**	×	**90.08**	84.61	76.35	×	**57.13**	**81.99**	×	**83.40**	88.97	77.70

multimodal hypothesis through our multimodal stream approach (without DCNN ensemble as explained in the beginning of this section). Furthermore, using the multimodal DCNN ensemble, we solve the temporal issue in an apparently more efficient way since it does not use RNNs and still is able to surpass more sophisticated approaches such as Yao et al [8].

D. State-of-the-art Comparison

Table I shows the results of our main approach, the multimodal DCNN ensemble (MDE), and its two simplifications, the DCNN ensemble and the multimodal stream (both explained at the beginning of this section). Our approaches overcome the results of our two baselines (inception module [10] and Yao et al [8]), as discussed before, and all methods of the literature surveyed by Jordão et al. [13]. Our method achieves, to the best of our knowledge, the state-of-the-art in the seven datasets evaluated. We reiterate that many efforts have been done to achieve modest improvements in HAR based on wearable sensor data, which reinforces that the Multimodal DCNN Ensemble and the DCNN Ensemble provide notable improvements.

According to the results, in the MHEALTH and PAMAP2P datasets, the DCNN ensemble shows superior results when compared to the multimodal DCNN ensemble in both protocols tested. We believe this is occurring because we had to reduce the number of parameters in the MDE network for these two datasets due to the limited computational resources available to run our experiments. Thus, we use a smaller pool of kernels and a fully connected with fewer neurons in the stream fusion block in these datasets.

IV. CONCLUSIONS

In this work, we proposed a multiscale ensemble-based approach of deep convolutional neural networks to address sensor-based human activity recognition (HAR). Our approach is able to learn individually the features of each sensor before performing the fusion and to model multiple temporal scales of an activity sequence. We demonstrate its suitability for HAR on wearable sensor data by performing an evaluation on seven important datasets. Our approach outperforms previous state-of-the-art results and an Inception module network adaptation used as a baseline to our convolutional kernel ensemble premise. We demonstrate that our approach works directly on the raw sensor data, with no pre-processing, which makes it general and minimizes engineering bias. As future work, we intend to study a dynamically way to choose the kernels employed in the ensemble.

During the development of this work, a technical paper entitled *´´Multiscale DCNN Ensemble Applied to Human Activity Recognition Based on Wearable Sensors"* containing the contributions of this thesis was published in the proceedings of the 26th European Signal Processing Conference (EUSIPCO) [25]. Additionally, we contribute as co-author in the journal paper *´´Human Activity Recognition based on Wearable Sensor Data - A Benchmark"*, which created a significant standardization of metrics and protocols on seven important datasets and made an extensive evaluation of several methods for human activity recognition based on wearable sensor domain. Currently, this work is under major review in the IEEE Sensors Journal and pre-printed in the arxiv.org database [13].

ACKNOWLEDGMENTS

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 311053/2016-5 and 438629/2018-3), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17), the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). Part of the results presented in this paper were obtained through research on a project titled ´´HAR-HEALTH: Reconhecimento de Atividades Humanas associadas a Doenças Crônicas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91. This study was financed in part by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors." *IEEE Communications Surveys and Tutorials*, 2013.

[2] Y. Chen and Y. Xue, "A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer," in *SMC*, 2015.

[3] A. Jordao, L. A. B. Torres, and W. R. Schwartz, "Novel approaches to human activity recognition based on accelerometer data," *Signal, Image and Video Processing*, 2018.

[4] A. Jordao, R. B. Kloss, and W. R. Schwartz, "Latent hypernet: Exploring all layers from convolutional neural networks," in *IJCNN*, 2018.

[5] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *ACM Multimedia Conference*, 2015.

[6] S. Ha, J.-M. Yun, and S. Choi, "Multi-modal convolutional neural networks for activity recognition," in *SMC*, 2015.

[7] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *IJCNN*, 2016.

[8] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," 2017.

[9] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation."

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, 2014.

[11] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," 2017.

[12] D. Pedamonti, "Comparison of non-linear activation functions for deep neural networks on mnist classification task," *arXiv preprint arXiv:1804.02763*, 2018.

[13] A. Jordao, A. C. Nazare Jr, J. Sena, and W. R. Schwartz, "Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art," *arXiv preprint arXiv:1806.05226*, 2018.

[14] J. R. Kwapisz, G. M. Weiss, and S. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explorations*, 2010.

[15] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *Applied Soft Computing*, 2015.

[16] H.-J. Kim and Y. S. Choi, "Eating activity recognition for health and wellness: A case study on asian eating style," in *ICCE*, 2013.

[17] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[18] B. Bruno, F. Mastrogiovanni, and A. Sgorbissa, "Wearable Inertial Sensors: Applications, Challenges, and Public Test Benches," *in IEEE Robot. Automat. Mag.*, 2015.

[19] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *UbiComp*, 2012.

[20] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015.

[21] J. W. Lockhart, G. M. Weiss, J. C. Xue, S. T. Gallagher, A. B. Grosner, and T. T. Pulickal, "Design considerations for the wisdm smart phone-based sensor mining architecture," in *SensorKDD*, 2011.

[22] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *ISWC*, 2012.

[23] O. Baños, R. García, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. R. Ruiz, A. Saez, and C. Villalonga, "mhealthdroid: A novel framework for agile development of mobile health applications," in *IWAAL*, 2014.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[25] J. Sena, J. B. Santos, and W. R. Schwartz, "Multiscale dcnn ensemble applied to human activity recognition based on wearable sensors," in *Signal Processing Conference (EUSIPCO), 2018 Proceedings of the 26th European*. IEEE, 2018.