IMPROVED GESTURE RECOGNITION BASED ON sEMG SIGNALS AND TCN

Panagiotis Tsinganos[1,2,], Bruno Cornelis[2], Jan Cornelis[2], Bart Jansen[2,3], Athanassios Skodras[1]*

[1]University of Patras, Department of Electrical and Computer Engineering, 26504 Patras, Greece
[2]Vrije Universiteit Brussel, Department of Electronics and Informatics, 1050 Brussels, Belgium
[3]imec, Kapeldreef 75, 3001 Leuven, Belgium
{panagiotis.tsinganos, skodras}@ece.upatras.gr, {bcorneli, jpcornel, bjansen}@etrovub.be

ABSTRACT

In recent years, the successful application of Deep Learning methods to classification problems has had a huge impact in many domains. In biomedical engineering, the problem of gesture recognition based on electromyography is often addressed as an image classification problem using Convolutional Neural Networks. In this paper, we approach electromyography-based hand gesture recognition as a sequence classification problem using Temporal Convolutional Networks. The proposed network yields an improvement in gesture recognition of almost 5% to the state of the art reported in the literature, whereas the analysis helps in understanding the limitations of the model and exploring new ways to improve its performance.

Index Terms— sEMG, Gesture Recognition, Deep Learning, CNN, TCN.

1. INTRODUCTION

Accurate gesture recognition is important for a number of applications including human computer interaction [1], prosthesis control [2] and rehabilitation gaming [3, 4]. Surface electromyography (sEMG) signals measured from the forearm contain useful information for decoding muscle activity and hand motion.

Machine Learning (ML) classifiers have been used extensively for determining the type of hand motion from sEMG data. A complete pattern recognition system based on ML includes acquiring data, extracting features, specifying a model and reasoning about new data. In the case of gesture recognition based on sEMG, electrodes attached to the arm and/or forearm acquire the EMG signals, and the typical extracted features are RMS, variance, zero crossings and frequency coefficients that are applied as inputs to classifiers like k-NN, SVM, MLP and Random Forests [5].

Recently, *Deep Learning (DL)* models have been successfully applied to sEMG-based gesture recognition. In these approaches, EMG data are represented as images and a Convolutional Neural Network (CNN) is used to determine the type of gesture. Although EMG signals are time-series data, no appropriate DL model (e.g. Recurrent Neural Network – RNN) has been used so far, to our knowledge.

In this work, taking into account the outcomes of [6] we investigate the application of temporal convolutional networks (TCN) [7] to the problem of sEMG-based gesture recognition. In contrast to the image classification approach, EMG signals are considered as a multi-dimensional time-series and only 1D convolutions are applied. Additionally, the outputs of the convolutions are computed using only past and current data (causal convolutions).

The main contributions presented in this paper are:

- the approach of the problem of sEMG-based gesture recognition as a time sequence classification problem using TCN,
- the improvement of the state-of-the-art accuracy by approximately 5%.

The paper is organized as follows. Section 2 provides an overview of the related gesture recognition approaches. In Section 3, we give a detailed description of the proposed TCN architecture. The experiments performed for the evaluation of the model are presented in Section 4, while the results and a discussion are given in Section 5. Finally, Section 6 describes the conclusions and outlines future work.

2. RELATED WORK

The problem of sEMG-based hand gesture recognition has been studied thoroughly using either conventional ML techniques or DL methods. In the case of ML-based methods, the first significant study is presented in [8] for the classification of four hand gestures using time-domain features extracted from sEMG measured with two electrodes. The authors of [9] achieve a 97% accuracy in the classification of three grasp motions using the RMS value from seven electrodes as the input to an SVM classifier. The works of [10, 11, 12] evaluate a wide range of EMG features with various classifiers for the recognition of 52 gestures (Ninapro dataset [11, 13]). The best performance is observed

The work is supported by the Andreas Mentzelopoulos Scholarships for the University of Patras and the VUB-UPatras International Joint Research Group (IJRG) on ICT.

with a combination of features and a Random Forest classifier resulting in 75% accuracy.

On the other hand, the first DL-based architecture, was proposed in [14]. The authors built a CNN-based model for the classification of six common hand movements resulting in a better classification accuracy compared to SVM. In [15], a simple model consisting of five blocks of convolutional and average pooling layers resulted in accuracy figures comparable, though not higher, to what was obtained with classical ML methods. In our previous work [16], we have investigated methods to improve the performance of this basic model. The results have shown that opting for max pooling and inserting dropout [17] layers after the convolutions boosts the accuracy by 3% (from 67% to 70%). The works of [18] and [19] incorporate dropout and batch normalization [20] techniques in their methodology. Apart from differences in model architectures, they measure EMG signals using a high-density electrode array, which has been proven effective to myoelectric control problems [21, 22, 23]. Using instantaneous EMG images, [18] achieves 89% accuracy on a set of eight movements, whereas in [19] a multi-stream CNN architecture is employed resulting in 85% accuracy on the Ninapro dataset.

Other important works based on DL architectures deal with the problem of model adaptation. In [24], the technique of adaptive batch normalization (AdaBN) [25] updates only the normalization parameters of a pretrained model, whereas in [26] the authors use weighted connections between a source network and the model instantiated for a new subject. Additionally, in [26] they propose data augmentation methods for sEMG signals.

3. PROPOSED MODEL

Unlike existing works in sEMG-based gesture recognition that address the problem as an image classification task, in this paper we develop a time sequence recognition model. It is based on the architecture of TCN presented in [6] for sequence prediction problems. The main characteristics of TCNs are the use of causal convolutions and the mapping of an input sequence to an output sequence of the same length. In addition, accounting for sequences with long history, this model uses dilated convolutions that enable a large receptive field (RF) [27] as well as residual connections [28] that allow training deeper networks. Considering that our task is to classify sEMG signals, the output layer of TCN is further processed by either an average over time (AoT) calculation or an attention (Att) mechanism [29] so that a single class label characterizes a complete sequence (Figure 1).

Given an input sequence of length N, $\{x\} = \{x_0, \dots, x_{N-1}\}$, the output of a causal convolutional layer is a sequence $\{y\} = \{y_0, \dots, y_{\mathrm{N}-1}\}$ such that the calculation of $y_n, n < N$ depends only on $\{x_0, \dots, x_n\}$. The dilated convolutions are calculated as:

$$y_n = (x *_d h)_n = \sum_m x_{n-dm} h_m$$

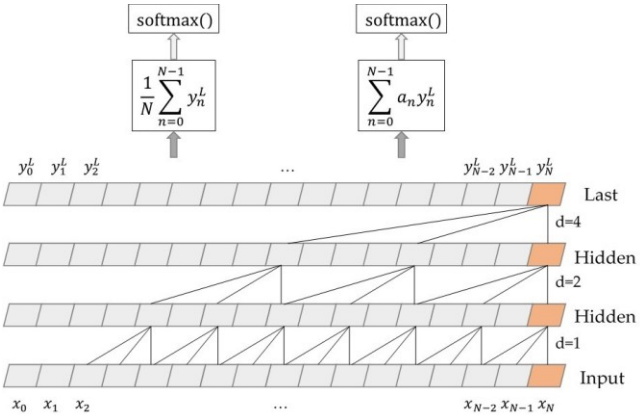

Figure 1. Graphical representation of the proposed model (Figure adjusted from [6])

where $*_d$ is the operator for dilated convolutions, d is the dilation factor and h is the filter's impulse response. For a TCN with L layers, the output of the last layer, y^L, is used for the sequence classification. When using the AoT, the class label $\hat{o}$ attributed to the sequence is found through a fully connected layer with softmax activation function:

$$s = \frac{1}{N} \sum_{n=0}^{N-1} y_n^L$$

$$\hat{o} = \mathrm{softmax}(W_o \cdot s + b_o)$$

where W_o, b_o are trainable parameters.

Otherwise, when using the Att mechanism, the class label is calculated as follows [29]:

$$v_n = \tanh(W_a \cdot y_n^L + b_a)$$

$$a_n = \mathrm{softmax}(v_n^T u_a)$$

$$s = \sum_{n=0}^{N-1} a_n y_n^L$$

$$\hat{o} = \mathrm{softmax}(W_o \cdot s + b_o)$$

where W_a, b_a are trainable parameters that transform the TCN output into a hidden representation v_n, u_a is a learnable context vector, a_n is the normalized importance for each time-step and s is the weighted sum of y^L based on the importance weights.

According to [6], the advantages of TCN include the ability to process sequences of arbitrary lengths, while they require less memory in training compared to RNNs due to shared filter parameters of the convolutions. On the other hand, during inference RNNs are more memory-friendly since their calculations at time n are based only on the current input and the hidden state, whereas a TCN needs the input sequence until the RF classification step is reached.

For the case of sEMG signals classification, the use of a time sequence model, which seems to be a natural choice given the nature of the EMG input data, is investigated in this paper, and compared to the image classification methods used so far. The hyper-parameters that have to be determined are the number of residual blocks and the number of layers per block, both of which affect the size of the RF.

Table 1. Details of the evaluated models. The number of layers refers only to convolutions.

Model	Classifier	RF	Size	Layers
AoT(300)	AoT	300ms	60K	4
AoT(2500)	AoT	2500ms	70K	7
Att(300)	Att	300ms	75K	4
Att(2500)	Att	2500ms	85K	7

Table 2. Performance of the evaluated models.

Model	Top-1	Top-3
AoT(300)	89.51% (3.43%)	97.42% (1.51%)
AoT(2500)	89.29% (3.80%)	97.37% (1.50%)
Att(300)	89.67% (3.50%)	97.35% (1.77%)
Att(2500)	89.76% (3.49%)	97.11% (1.68%)

4. EXPERIMENTS

The proposed TCN architecture was evaluated on data from the first dataset of the Ninapro database. It includes data acquisitions of 27 healthy subjects that perform each of the 52 gestures 10 times (repetition sequences). The types of gestures can be divided into three groups: i) basic finger movements, ii) isometric, isotonic hand configurations and basic wrist movements, and iii) grasping and functional movements. The data are acquired with 10 electrodes, of which eight are placed around the forearm and the other two are placed on the main activity spots of the large flexor and extensor muscles of the forearm [11].

The evaluation in this paper is based on existing works that have used this dataset [15, 16, 19]. Specifically, for each subject a new model is trained on data from seven repetitions and tested on the remaining three. As performance metrics we use the top-1 and top-3 accuracies (i.e. the accuracy when the highest and any of the 3 highest output probabilities match the expected gesture) averaged over all the subjects.

The performed experiments evaluated TCNs with RFs that correspond to 300ms (short) and 2500ms (long) of input sequences. In addition, an exponential dilation factor $d = 2^l$ for the l^{th} layer in the network was used. The classification was either based on the AoT or the Att mechanism. Therefore, four models were evaluated using complete repetition sequences as input. The details of each model are shown in Table 1.

All networks were trained using the adam optimizer [30] for 30 epochs with constant learning rate of 0.01 and a batch size of 128. To avoid overfitting the networks due to the small training set (size of $53 \times 7 = 371$), the training data of each subject were augmented by a factor of 10 using the time-warping, magnitude-warping and jittering methods described in [31]. Finally, dropout layers were appended after each convolution with a forget rate of 0.05. These values were selected after performing a grid search on a validation set of five randomly selected subjects.

The models were trained on a workstation with an Intel Xeon, 2.40 GHz (E5-2630v3) processor, 16 GB RAM and a Nvidia GTX1080, 8GB GPU. Each epoch was completed in approximately 20s. The results are summarized in Table 2.

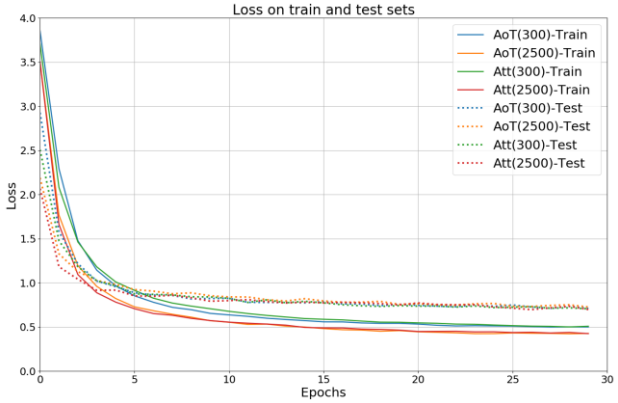

Figure 2. Average loss graphs during training and testing show convergence of the models.

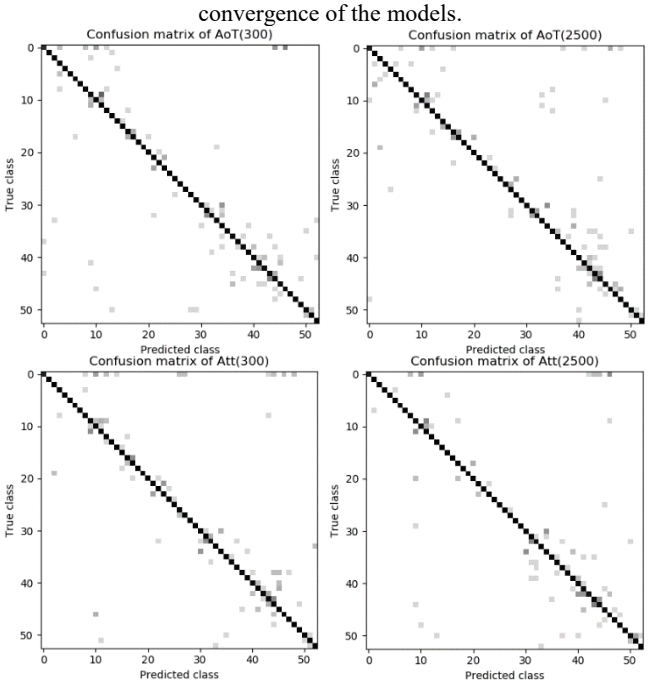

Figure 3. Average confusion matrix for each TCN model.

5. RESULTS AND DISCUSSION

The problem of hand gesture recognition based on sEMG is addressed as a sequence classification task using TCN models; a type of CNN that performs only temporal causal convolutions. Two hyperparameters of the model are explored: the receptive field of the convolutions and the type of classification. A comparison of the loss graphs during training and testing (Figure 2) shows that all models have been trained until convergence. In addition, the degree of overfitting is very small, therefore adjusting the forget rate of the dropout layers could only slightly improve the test accuracy.

Error analysis based on the average confusion matrix shows that some gestures are difficult to classify correctly.

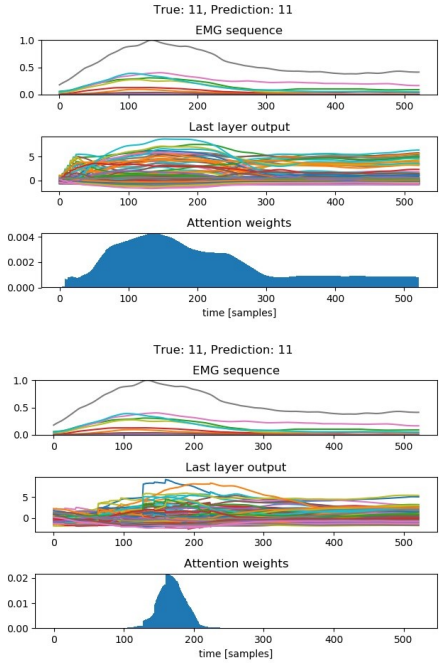

Figure 4. Correct classification of the thumb flexion gesture of subject-11 with Att(300) (up) and Att(2500) (down).

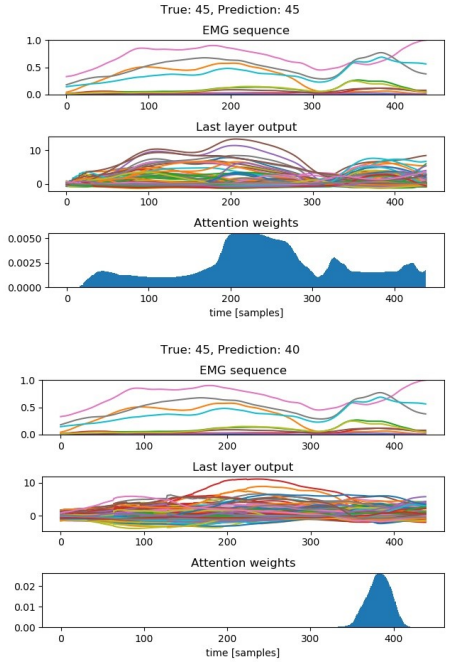

Figure 5. Classification of the quadpod grasp gesture of subject-11 with Att(300) (up) and Att(2500) (down). The latter misclassifies the gesture as a three-finger sphere grasp.

All confusion matrices display a few clusters of misclassifications between the following gestures: i) thumb adduction-abduction and flexion-extension (labels 9-12), ii) thumb opposing little finger and abduction of all fingers (labels 16-17), iii) power grip and hook grasp (labels 31-32), iv) types of three finger grasps (labels 40-42), and v) types of pinch grasps (labels 43-44). The first three cases are gestures were all the fingers except the thumb are fully or partially extended, while in the last two the gestures are very similar and hence some of the subjects might have been confused. In our previous work [16], we had shown that these groups of gestures were mostly misclassified by an image-based approach as well.

Table 3. Comparison with existing works.

	Top-1	**Model parameters***
[15]	66.59%	85K
[16]	70.48%	85K
[18]	76.10%	500K
[19]	85%	2.5M
This work	89.76%	85K

* calculated for 53 classes based on model description

A comparison between AoT and Att reveals that the specific attention (Att) mechanism did not yield much better results than the time average (AoT). The performance gain from using Att is only 0.2%. However, the values of the attention weights show that the model can identify discriminative features for the classification, since in general the weights values are not equally spread (Figures 4-5).

In Figure 4, the attention weights of the short RF and the long RF model are shown for the same gesture. It is clear that with a longer RF the network can isolate useful features. However, it might occur that an important region of the input is completely ignored (Figure 5) resulting in a classification error. Therefore, more experiments should be conducted to gain a better understanding of the relation between the RF and attention methods.

Compared to other experimental results on the Ninapro benchmark dataset, the proposed model outperforms the state-of-the-art on sEMG-based gesture recognition (Table 3). Previous approaches classified sEMG images generated from sliding windows of 150ms [15, 16] and 200ms [19], whereas in this work only complete sequences are considered. In addition to achieving the highest performance in the Ninapro dataset, the TCN-based model uses only a few parameters, which allows training with less data.

6. CONCLUSIONS

This paper investigated the application of a TCN model to the problem of sEMG-based recognition. In contrast to existing works that address the problem as an image classification task, the proposed model can categorize complete sEMG sequences. The architecture consists of a stack of layers that perform temporal causal convolutions, while the class label is computed either with an AoT or an Att method. The results showed that it outperforms the state-of-the-art by about 5% on a benchmark dataset. Finally, future research will investigate ways to better understand the effect of the receptive field and model depth on recognition accuracy.

8. REFERENCES

[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," Artif. Intell. Rev., vol. 43, no. 1, pp. 1–54, Jan. 2015.

[2] X. Chen et al., "Hand gesture recognition research based on surface EMG sensors and 2D-accelerometers," in 2007 11th IEEE Intern. Symp. on Wearable Computers, pp. 1–4, 2007.

[3] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," Res. Dev. Disabil., vol. 32, no. 6, pp. 2566–2570, Nov. 2011.

[4] L. Omelina et al., "Serious games for physical rehabilitation: designing highly configurable and adaptable games," in Proceedings of the 9th Intern. Conf. on Disability, Virtual Reality & Associated Technologies, pp. 195–201, 2012.

[5] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," J. Rehabil. Res. Dev., vol. 48, no. 6, p. 643, 2011.

[6] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic Convolutional and Recurrent Networks for sequence modeling," ArXiv e-prints, Apr. 2018.

[7] C. Lea et al., "Temporal Convolutional Networks for action segmentation and detection," in 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1003–1012, 2017.

[8] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," IEEE Trans. Biomed. Eng., vol. 40, no. 1, pp. 82–94, 1993.

[9] C. Castellini, A. E. Fiorilla, and G. Sandini, "Multi-subject/daily-life activity EMG-based control of mechanical hands.," J. Neuroeng. Rehabil., vol. 6, p. 41, Nov. 2009.

[10] I. Kuzborskij, A. Gijsberts, and B. Caputo, "On the challenge of classifying 52 hand movements from surface electromyography," in 2012 Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology Society, pp. 4931–4937, 2012.

[11] M. Atzori et al., "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," Sci. Data, vol. 1, p. 140053, 2014.

[12] A. Gijsberts et al., "Movement error rate for evaluation of Machine Learning methods for sEMG-based hand movement classification," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 4, pp. 735–744, Jul. 2014.

[13] M. Atzori et al., "Building the Ninapro database: A resource for the biorobotics community," Proc. IEEE RAS EMBS Int. Conf. Biomed. Robot. Biomech., pp. 1258–1265, 2012.

[14] K.-H. Park and S.-W. Lee, "Movement intention decoding based on Deep Learning for multiuser myoelectric interfaces," in 2016 4th Intern. Winter Conf. on Brain-Computer Interface (BCI), pp. 1–2, 2016.

[15] M. Atzori, M. Cognolato, and H. Müller, "Deep Learning with Convolutional Neural Networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," Front. Neurorobot., vol. 10, Sep. 2016.

[16] P. Tsinganos et al., "Deep Learning in EMG-based Gesture Recognition," in Proceedings of the 5th Intern. Conf. on Physiological Computing Systems, pp. 107–114, 2018.

[17] N. Srivastava et al., "Dropout: A simple way to prevent Neural Networks from overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

[18] W. Geng et al., "Gesture recognition by instantaneous surface EMG images," Sci. Rep., vol. 6, no. 36571, 2016.

[19] W. Wei et al., "A multi-stream Convolutional Neural Network for sEMG-based gesture recognition in muscle-computer interface," Pattern Recognit. Lett., Dec. 2017.

[20] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," ArXiv e-prints, Mar. 2015.

[21] N. Jiang et al., "Is accurate mapping of EMG signals on kinematics needed for precise online myoelectric control?," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 3, pp. 549–558, May 2014.

[22] S. Muceli, N. Jiang, and D. Farina, "Extracting signals robust to electrode number and shift for online simultaneous and proportional myoelectric control by factorization algorithms," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 3, pp. 623–633, May 2014.

[23] A. Stango, F. Negro, and D. Farina, "Spatial correlation of high density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 23, no. 2, pp. 189–198, Mar. 2015.

[24] Y. Du et al., "Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation," Sensors, vol. 17, no. 3, p. 458, Feb. 2017.

[25] Y. Li et al., "Revisiting Batch Normalization for practical domain adaptation," ArXiv e-prints, Nov. 2016.

[26] U. Côté-Allard et al., "Deep Learning for electromyographic hand gesture signal classification using Transfer Learning," ArXiv e-prints, Jan. 2018.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in 2016 Intern. Conf. on Learning Representations (ICLR), 2016.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," ArXiv e-prints, Dec. 2015.

[29] Z. Yang et al., "Hierarchical Attention Networks for document classification," in Proceedings of NAACL-HLT, pp. 1480–1489, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 2015 Intern. Conf. on Learning Representations (ICLR), 2015.

[31] T. T. Um et al., "Data augmentation of wearable sensor data for parkinson's disease monitoring using Convolutional Neural Networks," in Proceedings of the 19th ACM Intern. Conf. on Multimodal Interaction - ICMI 2017, vol. 517, pp. 216–220, 2017.