FORKED RECURRENT NEURAL NETWORK FOR HAND GESTURE CLASSIFICATION USING INERTIAL MEASUREMENT DATA

Philipp Koch[*], *Nele Brügge*[*], *Huy Phan*[†], *Marco Maass*[*], *Alfred Mertins*[*]

[*] University of Lübeck, Institute for Signal Processing, Lübeck, Germany
[†] University of Kent, School of Computing, Canterbury, Kent, United Kingdom

ABSTRACT

For many applications of hand gesture recognition, a delay-free, affordable, and mobile system relying on body signals is mandatory. Therefore, we propose an approach for hand gestures classification given signals of inertial measurement units (IMUs) that works with extremely short windows to avoid delays. With a simple recurrent neural network the suitability of the sensor modalities of an IMU (accelerometer, gyroscope, magnetometer) are evaluated by only providing data of one modality. For the multi-modal data a second network with mid-level fusion is proposed. Its forked architecture allows us to process data of each modality individually before carrying out a joint analysis for classification. Experiments on three databases reveal that even when relying on a single modality our proposed system outperforms state-of-the-art systems significantly. With the forked network classification accuracy can be further improved by over $10\,\%$ absolute compared to the best reported system while causing a fraction of the delay.

***Index Terms*—** Inertial measurement unit, hand gesture recognition, recurrent neural network, multi-modal fusion

1. INTRODUCTION

In many applications such as virtual reality [1] and human machine interaction [2] the recognition of hand gestures is essential. The conditions these systems have to meet vary between applications. However, in general it is desirable for hand gesture recognition systems to be cheap and free of delay. Also many applications require a system that can be realised as a mobile device or an embedded system.

In many cases, especially in the medical field of prosthesis control [3] or the control of exoskeletons [4, 5], hand movements are classified on the base of surface electromyography (sEMG) signals since the signals can be unproblematically acquired by electrodes. Therefore, the research was mostly focused on new approaches to analyse sEMG signals. Most of the successful systems followed a classic pipeline. It starts with signal acquisition followed by preprocessing and windowing. Afterwards, hand crafted features are extracted for each window [6]. The resulting features are then presented to a standard classifier such as support vector machine or random forest that finally determines the hand movement [7, 8, 9]. Besides, techniques from the field of deep learning have recently evolved. For instance, recurrent neural networks (RNNs) have been used to classify sequences of feature vectors representing sEMG signals [10, 11]. Even with a compact network with a small number of parameters RNNs were shown to achieve outstanding results. In addition, with the usage of convolutional neural networks on raw sEMG signals, feature extraction and classification are jointly performed within one network and trained in an end-to-end fashion [12, 13]. This allows the features to be learned rather than being hand-crafted. These networks were shown to achieve state-of-the-art performance. In general, all of these approaches suffer from two main drawbacks. Firstly, in order to achieve satisfying performance, long windows ($100\,\mathrm{ms}$ and longer) are required that cause an undesired delay. Secondly, expensive electrodes are needed to acquire the signals.

In addition or even as a cheap alternative to electrodes, accelerometers or more complex IMUs can be used. An IMU usually has three sensor modalities: an accelerometer, a gyroscope, and a magnetometer. With IMUs attached to the skin of a subject's forearm, the movement of the arm, the deformation of the skin and the change of the electromagnetic field caused by the muscle contraction can be measured. Consequently, it enables gathering information about the arm movement and the finger movement indicated by the muscle movement. Prior works suggest that it is possible to determine hand movements given the signals of IMUs [14, 15] or a combination of sEMG signals and inertial measurement (IM) data [16, 17]. However, these approaches adhere to the standard pipeline, i.e. using long windows, relying on hand-crafted features and more conventional classification methods, resulting in long delays as discussed above.

The aim of this work is to develop a system for a cheap, IMU-based hand movement detection system that is expected to only cause minimal delay. Because RNNs have been proven to be effective for sequence analysis on many different temporal data types such as audio signals [18, 19] and electroencephalography data [20, 21] as well as sEMG based hand movement recognition, we employ them for decoding the hand gestures from a sequence of small windows ($5\,\mathrm{ms}$ long) of raw data. At first, a fairly simple RNN consisting of a

single RNN cell is presented. Its performance is evaluated for classifying hand movements given data of a single modality as well as for the analysis of the multi-modal IM data. To improve the efficiency in handling multi-modal data a tailored RNN architecture, namely forked RNN, is further proposed. At a lower level of the network, an RNN cell is designated to each modality (i.e. accelerometer, gyroscope, or magnetometer) to process the modality-specific input independently from the rest. Afterwards the outputs of the individual RNN streams are fused and commonly analysed by an RNN cell at a higher level of the network.

To show the effectiveness of the proposed approach, experiments were conducted on three databases. Results indicate that even when using just a single sensor modality, we can achieve a new state-of-the-art performance for all databases. With the forked RNN and the multi-modal IM data, the current state-of-the-art systems using IM data and sEMG signals are outperformed by more then $10\,\%$ absolute. Furthermore, as the proposed systems use $5\,\mathrm{ms}$ long windows, the induced delay is only a fraction of that caused by the other systems.

2. JOINT ANALYSIS OF MULTI-MODAL IM DATA

In contrast to the commonly used strategies for determining hand movements based on sEMG as well as IM data, the proposed systems do not rely on hand-crafted features. Instead, raw IM data are given to the network and features are learned automatically in a data-driven way within the network. The analysis of a single short window prevents the full exploration of the sequential nature of the IM data. Therefore, sequences of consecutive windows are analysed using an RNN-based approach. The used network architecture is illustrated in Fig. 1. The RNN allows us to process the data at the current time step with respect to the inputs of the preceding time steps. The used networks consist of a single RNN cell that receives all data collected by all IMUs at each time step. We chose the long-short term memory (LSTM) cell [22] as the RNN cell. Its hidden layer can be described as $(\boldsymbol{h}_t, \boldsymbol{C}_t) = \mathcal{H}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{C}_{t-1})$ with $\boldsymbol{x}_t$ being the current input, $\boldsymbol{h}_{t-1}$ representing the output of the previous time step, and $\boldsymbol{C}_{t-1}$ denoting the cell state also of the previous time step. The update of the cell state is given by

$$\boldsymbol{C}_t = \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t \tag{1}$$

with $*$ being the Hadamard product, the gates $\boldsymbol{f}_t$ and $\boldsymbol{i}_t$ defined as

$$\boldsymbol{f}_t = \mathrm{sigm}(\boldsymbol{W}_f(\boldsymbol{h}_{t-1} \oplus \boldsymbol{x}_t) + \boldsymbol{b}_f), \tag{2}$$

$$\boldsymbol{i}_t = \mathrm{sigm}(\boldsymbol{W}_i(\boldsymbol{h}_{t-1} \oplus \boldsymbol{x}_t) + \boldsymbol{b}_i), \tag{3}$$

and

$$\tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_C(\boldsymbol{h}_{t-1} \oplus \boldsymbol{x}_t) + \boldsymbol{b}_C) \tag{4}$$

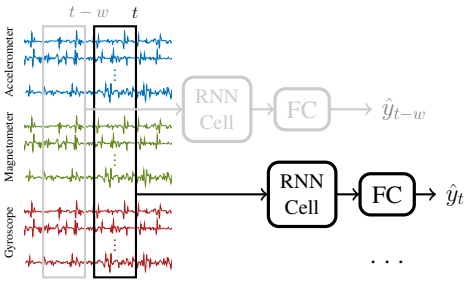

Fig. 1. Illustration of the RNN architecture for early fusion.

with $\boldsymbol{W}$ being weight matrices, $\boldsymbol{b}$ denoting biases, and $\oplus$ representing vector concatenation. The output of the LSTM cell is then given by

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \tag{5}$$

where $\boldsymbol{o}_t$ is the output gate defined as

$$\boldsymbol{o}_t = \mathrm{sigm}(\boldsymbol{W}_o(\boldsymbol{h}_{t-1} \oplus \boldsymbol{x}_t) + \boldsymbol{b}_o). \tag{6}$$

It is obvious that this kind of network produces an output for each input. Every output is classified by a tied fully connected layer that is defined as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \tag{7}$$

with $\boldsymbol{x}$ the input vector, $\boldsymbol{W}$ and $\boldsymbol{b}$ being a trainable weight matrix and a bias vector, respectively. Since the aim is a classification of hand movements, the softmax function is used as nonlinear activation function.

3. PARTIALLY INDIVIDUAL ANALYSIS OF MULTI-MODAL IM DATA

Because the three sensor types of an IMU are acquiring very different signals, it is reasonable to have an individual preprocessing or feature extraction for each modality. Therefore we propose an approach where a preprocessor is dedicated to each modality. Mid-level fusion of different preprocessed data streams is then carried out, followed by joint analysis and classification.

Obviously the short windows used in this work cannot cover all distinctive temporal patterns that vary in length. To detect these long patterns the feature extraction should incorporate the temporal context of the windows. Using RNN cells as preprocessors allows us to incorporate information of previous windows within the feature extraction for the current window. The network architecture is illustrated in Fig. 2. Let $\boldsymbol{x}_t$ denote all data of all IMUs corresponding to the current time step t. The input $\boldsymbol{x}_t$ is now split into three vectors $\boldsymbol{x}_t^{acc}$, $\boldsymbol{x}_t^{\mathrm{gyro}}$, and $\boldsymbol{x}_t^{\mathrm{magn}}$ containing all data corresponding to accelerometer, gyroscope, and magnetometer, respectively. Each input vector is presented to its corresponding RNN cell. Consequently we have

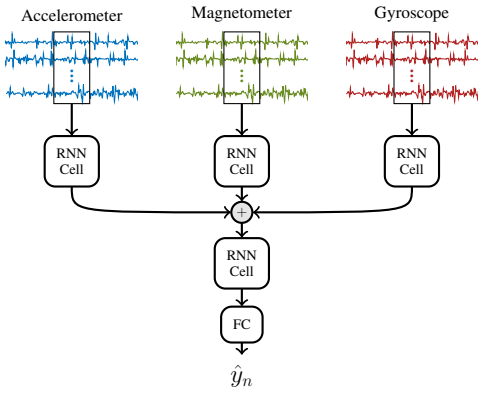

Fig. 2. Illustration of the forked network architecture for multi-modal fusion.

$\mathcal{H}^{\text{acc}}(\boldsymbol{x}_t^{\text{acc}}, \boldsymbol{h}_{t-1}^{\text{acc}}, \boldsymbol{C}_{t-1}^{\text{acc}})$, $\mathcal{H}^{\text{gyro}}(\boldsymbol{x}_t^{\text{gyro}}, \boldsymbol{h}_{t-1}^{\text{gyro}}, \boldsymbol{C}_{t-1}^{\text{gyro}})$, and $\mathcal{H}^{\text{magn}}(\boldsymbol{x}_t^{\text{magn}}, \boldsymbol{h}_{t-1}^{\text{magn}}, \boldsymbol{C}_{t-1}^{\text{magn}})$. The upper index indicates that a different set of weights and biases is learned for each RNN cell.

After this first stage the resulting outputs $\boldsymbol{h}_t^{\text{acc}}$, $\boldsymbol{h}_t^{\text{gyro}}$, $\boldsymbol{h}_t^{\text{magn}}$ are fused to

$$\boldsymbol{x}_t^{\text{fused}} = \boldsymbol{h}_t^{\text{acc}} \oplus \boldsymbol{h}_t^{\text{gyro}} \oplus \boldsymbol{h}_t^{\text{magn}} \tag{8}$$

for the comprehensive analysis over all modalities. For this purpose another RNN cell $\mathcal{H}^{\text{fused}}(\boldsymbol{x}_t^{\text{fused}}, \boldsymbol{h}_{t-1}^{\text{fused}}, \boldsymbol{C}_{t-1}^{\text{fused}})$ is employed. Finally, in analogy to the network described in the previous section its output $\boldsymbol{h}_t^{\text{fused}}$ is presented to a fully-connected layer with softmax activation for classification.

4. TRAINING AND EVALUATION OF NETWORKS

For all presented networks the training and testing procedure are the same. During training, sequences of fixed length (1 s) are extracted from the training examples. The sequences are presented to the network that classifies each window of a sequence. The network's error is estimated using only the classification $\hat{\boldsymbol{y}}_T$ of final window T of the presented sequence. As loss function for the optimization of the network, we use the cross-entropy

$$E\left(\Theta|\boldsymbol{X}, \boldsymbol{y}_T\right) = -\boldsymbol{y}_T \log\left(\hat{\boldsymbol{y}}_T\left(\Theta|\boldsymbol{X}\right)\right) \tag{9}$$

where $\boldsymbol{X}$ denotes the input sequence and $\boldsymbol{y}_T$ the ground truth corresponding to the final time step T represented as one-hot encoded vector. The network's parameters are denoted by Θ.

In contrast to training, for testing each test example is represented as a single sequence. Consequently, the lengths of the sequences vary. However, the sequences are presented to the network that assigns a class label to each window of every sequence. The performance of the network is evaluated by comparing the estimated class for every window of all sequences with its corresponding ground truth and calculating the accuracy.

5. EXPERIMENTS

5.1. Database

We conducted experiments on three publicly available databases: DB2, DB3, and DB7 published along the Ninapro project [7, 16]. Their general aim is to provide data for the determination of hand gestures on the base of biosignals.

The acquisition of all three databases followed mainly the same protocol. For the experiments the subjects were equipped (if possible) with 12 electrodes with integrated accelerometers or IMUs. The subjects were asked to perform a number of different hand movements of which each one was repeated six times.

In DB2 and DB3, Delsys® Trigno™Wireless sensors were used to acquire the data. Besides the sEMG signal, each sensor provides the data of a tri-axial accelerometer sampled at 148 Hz. This data are upsampled to 2 kHz to match the sampling frequency of the sEMG signal. In both databases, DB2 and DB3, the subjects performed 50 different hand gestures. However, DB2 and DB3 differ in their subjects. While DB2 contains experiments of 40 able-bodied persons, DB3 includes experiments of 11 amputees.

For DB7, sensors of Delsys® Trigno™IM Wireless System were used. The sensors were used to collect both the sEMG data at a sampling frequency of 2 kHz and the raw signals of an IMU with 9-degree-of-freedom (tri-axial accelerometer, gyroscope, magnetometer) sampled at 128 Hz. The IM data were upsampled to 2 kHz. During the experiments for DB7, the subjects performed a subset of 40 gestures instead of all gestures as in DB2 and DB3. The seventh database contains the data of 20 able-bodied persons and 2 amputees.

In this work DB7 was used to evaluate the suitability of IM data for determining hand gestures. DB2 and DB3 were included to investigate whether the hand gesture recognition works for different sensors and for different subjects like amputees.

5.2. Preprocessing

The data was split into training and test data according to the original suggestions for the database. In order to prepare the signals for the analysis by the networks, the signals of each axis of each sensor were normalized individually by subtracting the mean and dividing by standard deviation. The necessary statistics were calculated only on the train data. For classification the signals were split into consecutive 5 ms long windows.

5.3. Results

The accuracies reported in the following are average accuracies calculated across the subjects of the corresponding database.

Table 1. Results on DB7 obtained by the RNN-based on a LSTM with 256 state size using Accelerometer (Acc.), Gyroscope (Gyro.), and Magnetometer (Magn.) data.

Group	Acc.	Gyro.	Magn.
Able-bodied	89.0 %	86.6 %	86.5 %
Amputated	83.9 %	81.6 %	82.6 %

Table 2. Accuracy comparison of accelerometer and sEMG. The reported results for accelerometer were obtained by a network based on a single LSTM cell with a state of size 256. The results were achieved on 100 ms long windows that were represented by a feature vector and classified by an RNN.

Database	Accelerometer (LSTM (256))	sEMG [11]
DB2	80.4 %	78.0 %
DB3	67.7 %	55.3 %

5.3.1. Individual Sensors

Firstly, we evaluated the possibility of recognising hand movements based on a single modality. Therefore, a network described in Section 2 with an LSTM cell and a state size of 256 was used. The results on DB7 of this network for the different kinds of input data are shown in Table 1. These results suggest that, on each of the three modalities, the RNN outperforms the state-of-the-art results reported in [16], where 256 ms long windows and a standard classification scheme were used. Considering the accelerometer data, our obtained results surpass those in [16] by more than 15 % absolute for able-bodied subjects and by about 25 % absolute for the amputees. The difference is even more striking for the gyroscope data where the approach in [16] achieved less then 10 % accuracy for both groups of subjects. For the magnetometer the difference is not that drastic but our RNN still outperforms the reported results by far. Even when considering the best results in [16] (82.7 % and 77.8 % accuracy for able-bodied and amputated subjects) obtained by using all sEMG and IM data and windows of length 256 ms our results are still better. Thus, the RNN-based approach appears to be efficient for the problem.

To show that hand gestures can be determined well we tested our system on DB2 and DB3, too. In Table 2 the results are presented and compared with, to the best of our knowledge, the best reported results on those datasets [11]. These results were obtained with an RNN on sEMG data that were segmented into 100 ms windows and then represented by a feature vector. As can be seen, even with significantly shorter windows, our proposed system outperforms the state-of-the-art sEMG-based hand movement classification system.

Consequently, RNN-based networks seem to be efficient for analysing IM data especially in the context of hand gesture classification. RNNs allow us to use very small signal windows and are able to extract information from the gyroscope where other methods are struggling.

Table 3. Comparison of different fusion approaches on DB7. For the network described in Section 2 a LSTM cell with a state of size 256 was used. In the forked RNN (see Section 3) LSTM cells with a state size of 256 were used.

Method	able-bodied	amputated
[16]	81.7 %	77.7 %
LSTM (256)	92.1 %	89.7 %
Forked RNN	93.4 %	89.3 %

5.3.2. Sensor Fusion

In a final experiment, we examined whether it is possible to gain performance by the joint analysis of the multi-modal data of IMUs. The RNNs described in Section 2 with state sizes of 256 were used for the joint analysis of all IM data. Furthermore, a forked RNN network proposed in Section 3 utilising LSTM cells with a state size of 256 was also evaluated. The results obtained for DB7 are shown in Table 3. As can be seen, both of our proposed systems outperform the prior work in [11] by more than 10 % for both amputees and healthy subjects. This fact emphasizes that RNNs can analyse IM data with outstanding accuracy while providing a fast reaction time due to the possibility of using small windows.

Moreover, the forked RNN achieves better performance compared to the single RNN cell for healthy subjects. This result does not hold for the amputees but since only two amputees are included in the database the results on this group are not too reliable. Overall, the proposed mid-fusion approach within an end-to-end trained network seems to be favourable. Intuitively it seems to be suitable for the data to have an individual preprocessing unit for each modality.

6. CONCLUSIONS

We have shown that it is possible to decode over 50 hand gestures from raw IM data. Compared with state-of-the-art systems using often at least 100 ms long windows the proposed networks allow for a quick response to hand gestures and a minimal delay due to the usage of 5 ms long windows. The experimental results revealed that our proposed systems outperform the state-of-the-art systems on each individual modality. Furthermore, the proposed forked RNN was shown to be beneficial for multi-modal data analysis since the signals of each modality can be preprocessed in individual RNN cells within an end-to-end trained network. An accuracy gain of 10 % absolute over the state-of-the-art performance can be achieved by the forked RNN.

7. REFERENCES

[1] Fabricio Muri, Celina Carbajal, Ana M Echenique, Hugo Fernndez, and Natalia M Lpez, "Virtual reality upper limb model controlled by emg signals," *J. Phys. Conf. Ser.*, vol. 477, 2013.

[2] Juan Cheng, Chen Xiang, Zhiyuan Lu, Kongqiao Wang, and Minfen Shen, "Key-press gestures recognition and interaction based on semg signals," in *Proc. Int. Conf. Multimodal Interact. and Mach. Learn. Multimodal Interact.*, 2010.

[3] C. Cipriani, F. Zaccone, S. Micera, and M. C. Carrozza, "On the shared control of an emg-controlled prosthetic hand: Analysis of userprosthesis interaction," *IEEE Trans. Robot.*, vol. 24, no. 1, pp. 170–184, 2008.

[4] J. Rosen, M. Brand, M. B. Fuchs, and M. Arcan, "A myosignal-based powered exoskeleton system," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 3, pp. 210–222, 2001.

[5] K. Kiguchi and Y. Hayashi, "An emg-based control for an upper-limb power-assist exoskeleton robot," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1064–1071, 2012.

[6] 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.

[7] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. Mittaz Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, no. 140053, 2014.

[8] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and emg sensors," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 6, pp. 1064–1076, 2011.

[9] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–854, 2003.

[10] Philipp Koch, Huy Phan, Marco Maass, Fabrice Katzberg, and Alfred Mertins, "Recurrent neural network based early prediction of future hand movements," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, July 2018.

[11] Philipp Koch, Huy Phan, Marco Maass, Fabrice Katzberg, Radoslaw Mazur, and Alfred Mertins, "Recurrent neural networks with weighting loss for early prediction of hand movements," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, September 2018.

[12] M. Atzori, Cognolato M, 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, no. 9, 2016.

[13] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, "Gesture recognition by instantaneous surface emg images," *Sci. Rep.*, vol. 6, no. 36571, 2016.

[14] Christoph Amma, Marcus Georgi, and Tanja Schultz, "Airwriting: A wearable handwriting recognition system," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 191–203, 2014.

[15] B. Hartmann and N. Link, "Gesture recognition with inertial sensors and optimized dtw prototypes," in *IEEE Trans Syst. Man Cybern. B Cybern.*, 2010.

[16] A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements," *J. Neuroeng. Rehabil.*, vol. 14, no. 71, 2017.

[17] Marcus Georgi, Christoph Amma, and Tanja Schultz, "Recognizing hand and finger gestures with imu based motion and emg based muscle activity sensing," in *Proc. Biomed. Eng. Syst. Technol. Int. Jt. Conf. BIOSTEC*, 2015.

[18] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio scene classification with deep recurrent neural networks," in *Proc. Interspeech*, 2017.

[19] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016.

[20] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *Proc. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2018.

[21] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *arXiv Preprint arXiv:1809.10932*, 2018.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.