Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts

Carvelli, Lorenzo; Olesen, Alexander N.; Brink-Kjær, Andreas; Leary, Eileen B.; Peppard, Paul E.; Mignot, Emmanuel; Sørensen, Helge Bjarup Dissing; Jennum, Poul

Published in:
Sleep Medicine

Link to article, DOI:
[10.1016/j.sleep.2019.12.032](https://doi.org/10.1016/j.sleep.2019.12.032)

Publication date:
2020

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Carvelli, L., Olesen, A. N., Brink-Kjær, A., Leary, E. B., Peppard, P. E., Mignot, E., Sørensen, H. B. D., & Jennum, P. (2020). Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts. *Sleep Medicine*, 69, 109-119. <https://doi.org/10.1016/j.sleep.2019.12.032>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Number of tables: 7

Number of figures: 6

Abstract word count: 242

Word count: 6485

Design of a deep learning model for automatic scoring of periodic and non-periodic leg movements during sleep validated against multiple human experts

Lorenzo Carvelli^{1,2}, Alexander N. Olesen^{1,2,4}, Andreas Brink-Kjær^{1,2,4}, Eileen B. Leary², Paul E. Peppard³, Emmanuel Mignot^{2*}, Helge B.D. Sørensen^{1*}, and Poul Jennum^{4*}

¹ Department of Health Technology, Technical University of Denmark, Kgs Lyngby, Denmark

² Stanford University Center for Sleep Sciences and Medicine, Palo Alto, CA, USA.

³ Department of Population Health Sciences, University of Wisconsin-Madison, WI, USA

⁴ Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark

*Shared last authorship

Corresponding author

Poul Jennum

Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark

Address

Email: poul.joergen.jennum@regionh.dk

Institution where work was performed

Department of Health Technology, Technical University of Denmark; Stanford University Center for Sleep Sciences and Medicine; Danish Center for Sleep Medicine.

ABSTRACT

Objective: The gold standard of scoring of leg movements (LMs) and periodic LMs (PLMS) in overnight polysomnography (PSG) studies is manual scoring, which is subject to inter-scorer variability. The objective of this study is to design and validate an end-to-end deep learning system for the automatic scoring of LMs and PLMS in sleep.

Methods: The deep learning system was developed, validated and tested, with respect to manual annotations by expert technicians on 800 overnight PSGs using a leg electromyography channel. The study includes data from three cohorts, namely, the Wisconsin Sleep Cohort (WSC), Stanford Sleep Cohort (SSC) and MrOS Sleep Study. The performance of the system was further compared against individual expert technicians and existing PLM detectors.

Results: The system achieved an F1 score of 0.83, 0.71, and 0.77 for the WSC, SSC, and MrOS cohorts, respectively. In a total of 60 PSGs from the WSC and the SSC scored by 9 expert technicians, the system performed better than 2 and comparable to 7 of the individual scorers with respect to a majority-voting consensus of the remaining scorers. In 60 PSGs from the WSC scored accurately for PLMS, the system outperformed 4 previous PLM detectors, which were all evaluated on the same data, with an F1 score of 0.85.

Conclusions: The proposed system performs better or comparable to individual expert technicians while outperforming previous automatic detectors. Thereby, the study validates fully automatic methods for scoring LMs in sleep.

Keywords: Leg Movements during Sleep, Periodic Leg Movements during Sleep, Polysomnography, Manual Scoring of Polysomnography, Automatic Event Detection.

1. INTRODUCTION

The gold standard for evaluating sleep and related phenomena is nocturnal polysomnography (PSG), a multi-modal system allowing acquisition of several physiological signals carried out during sleep. Among scored events of clinical interest, leg movements (LMs) and periodic leg movements during sleep (PLMS), a PSG feature often associated with Restless Legs Syndrome (RLS), are reported [1]. Although the pathological relevance of these events is not fully established, the presence of PLMS has been associated with deleterious cardiovascular outcomes [2,3,4], and contributes to disturbed sleep with resulting excessive daytime sleepiness, especially when PLMS are associated with arousals [5].

Although PLM scoring is often “computer assisted”, the current gold standard in clinical practice is manual scoring, which is carried out by an expert technician with subsequent review by a sleep physician. This task is associated with inter-scorer variability [6]. The World Association of Sleep Medicine (WASM) established specific criteria for scoring LMs: LM onset is defined as an increase of the electromyography (EMG) of at least $8\mu\text{V}$ above the resting baseline, while offset threshold is established as a drop in the signal to less than $2\mu\text{V}$ above baseline. The morphology of the EMG in LMs is required to contain a period ≥ 0.5 s with the median EMG amplitude $\geq 2\mu\text{V}$ above baseline [7]. Finally, once LMs have been annotated, those associated with sleep apnea events are removed, and PLMS series are scored [7].

Several methods designed to automatically score limb movements have been proposed. These include methods designed by Tauchmann and Pollmacher [8], Wetter et al. [6], Ferri et al. [9], Moore et al. [10] and Huang et al. [11]. In addition, Alvarez-Estevéz [12] implemented a method for detection and analysis of periodicity of limb movements that handles artifact presence and compares the results to some of the previous methods. Last, Stefani et al. [13] carry out the validation on 20 RLS patients and 20 controls of an algorithm for PLMS detection which is a feature of a commercial PSG software. These methods have been applied on a limited number of subjects and are built on rule-based detection methods rather than machine learning methods. A common issue of rule-based

methods is that these methods are not good at handling unique cases that can “fit the rule” but would be naturally excluded by a human scorer as unlikely. Examples of these limitations are exemplified by our prior work described in Moore et al. [10], where many adjustments to the rules needed to be made to consider unnoticed implicit rules such as increased baseline noise due to electrocardiography artifacts, large movement artifacts, etc. In contrast, machine learning algorithms learn to reject some artifactual detection by mimicking human-like scoring in cases that approximate scenarios on which the algorithm was trained.

The main goal of this work was to design, validate and test a robust, fast and automatic machine learning algorithm for scoring LMs, PLMS and calculating PLM indices. To ensure robustness, the algorithm was trained on large datasets collected at different locations and compared with consensus scoring and inter-scorer reliability obtained from multiple scorers.

2. MATERIAL AND METHODS

2.1. Data Description: We used the combined left/right anterior tibialis channels (LAT/RAT) to determine significant leg movements according to the 2007 AASM Manual for Scoring Sleep [14] with a slight modification, namely we did not consider limb movements in wake. PLMS are defined by the AASM 2007 as the consecutive sequence of four or more LMs whose inter-movement intervals are between 5 and 90 sec. This definition of PLM is consistent with the clinical guidelines set for by the World Association of Sleep Medicine in 2006 and the American Sleep Disorders Association in 1993 [7].

The data used in this work include three representative subsamples of different research and clinical data, totaling 800 subjects. We selected recordings based on availability of LM annotations as well as diversity of demographics. This was done to maximize robustness of the deep learning method and to reduce risk of overfitting [15]. We included both PSG recordings acquired for research purposes (WSC, MrOS) and clinical data (SSC). Data available were characterized by one single merged channel for leg EMG, specifically the left/right anterior tibialis (LAT/RAT) EMG. In addition, we used the electrocardiography (ECG) channel to adaptively filter out electrical heart activity interference from the LAT/RAT EMG. Informed consent was obtained for all included subjects in parent epidemiological studies, and this study was approved by the cohort steering committees and the Stanford Institutional Review Board.

A subset of 275 subjects randomly sampled from the WSC were used. The WSC is a longitudinal study of the causes, consequences and natural history of sleep disorders. The WSC uses overnight, in-laboratory sleep studies carried out in a sample of 1500 Wisconsin state employees, assessed at four-year intervals [16]. LAT/RAT EMG was acquired through Ag/AgCl “Huggables” snap electrodes using a sampling frequency (f_s) of 100 Hz or 200 Hz. The signal was then pre-filtered through a band-pass filter with low cutoff at 10 Hz and a high cutoff at 30 Hz. The ECG signal was

pre-filtered using a band-pass of low and high cutoffs at 3 and 30 Hz respectively [16]. An independent subset of 30 subjects randomly sampled from the WSC were scored for LMs by five different trained sleep technicians from a pool of nine; these subjects were included in the test set. The scorers were specifically instructed to accurately score leg movements and their duration.

A total of 177 subjects sampled at random from the Stanford Sleep Cohort (SSC) were also included in this work. This data belongs to patients with diverse sleep disorders undergoing in-laboratory PSG. The LAT/RAT EMG was sampled at 128 or 256 Hz through Ag/AgCl electrodes. ECG had a f_s of 256 Hz. In both cases, pre-filtering was applied through a band-pass with cutoffs at 0.1Hz and 0.45 times the f_s of the signals. Analogously to the WSC test subset, additional LM annotations of the SSC data were performed by five human technicians for each recording in a random sample of 30 subjects from SSC independent of the training data.

MrOS is an ancillary study of the study Osteoporotic Fractures in Men Study [17,18]. Between 2000 and 2002, 5994 65-year-old (or older) community-dwelling men were enrolled at six clinical centers for baseline examination. 3135 of the participants underwent PSG acquisition between 2003 and 2005. Those acquisitions in the dataset are gathered under the label “Visit 1”, while a follow up PSG recording “Visit 2” (not used in this study) was conducted in the years 2009 through 2012 [17,18]. Importantly, MrOS data are characterized by the use of piezoelectric electrodes for EMG acquisition (“Compumedics” piezo leg leads), whereas Ag/AgCl patches are used for ECG. f_s were 64 Hz and 512 Hz for legs EMG and for ECG respectively. All data were pre-filtered with a high pass filter of 0.15 Hz cutoff. In this work, a representative sample of 348 of 3135 “Visit 1” subjects was used. The data were stored as two separate channels, one for each leg with respect to a ground reference. To align this data with the other two datasets, we recovered the LAT/RAT channel through subtraction.

For all three data cohorts, manual annotations of (non-)periodic leg movements were performed by registered sleep technicians.

Table I Demographic data from the three different cohorts used in this study. Mean $\pm$ standard deviation of the data are presented. The p-values in the table were obtained from the Kruskal-Wallis test, carried out to verify or reject the null hypothesis that the data from the three cohorts is drawn from the same distribution. * Periodic leg movements index not available due to absence of annotations † p-value obtained from chi-squared test.				
Cohort	WSC	SSC	MrOs	p-value
Age	54.3 $\pm$ 7.3	44.5 $\pm$ 14.3	76.2 $\pm$ 5.8	< 0.001
BMI	32.0 $\pm$ 6.8	27.1 $\pm$ 6.4	27.3 $\pm$ 3.8	< 0.001
PLMSi	- *	5.4 $\pm$ 11.2	42.3 $\pm$ 35.4	< 0.001
Sleep Efficiency	0.76 $\pm$ 0.14	0.79 $\pm$ 0.15	0.55 $\pm$ 0.12	< 0.001
REML [min]	133.8 $\pm$ 86.4	137.0 $\pm$ 84.6	105.5 $\pm$ 78.7	< 0.001
N1 (%)	0.10 $\pm$ 0.06	0.11 $\pm$ 0.09	0.07 $\pm$ 0.04	< 0.001
N2 (%)	0.65 $\pm$ 0.10	0.64 $\pm$ 0.46	0.63 $\pm$ 0.11	0.004
N3 (%)	0.10 $\pm$ 0.09	0.09 $\pm$ 0.09	0.11 $\pm$ 0.09	0.006
REM (%)	0.15 $\pm$ 0.07	0.17 $\pm$ 0.07	0.19 $\pm$ 0.07	< 0.001
Total Subjects	275	177	348	-
Frac. Males	0.51	0.55	1	< 0.001 †

In Table I, we report on the demographics of the three different datasets, including information on scored PLMS index (PLMSi), when available, as well as sleep efficiency, percentage of each manually annotated sleep stage and REM latency (REML). Data from all three datasets were manually annotated for LMs by trained technicians who scored regardless of association with other events, such as sleep apnea or arousals. Events taking place in wake were not scored.

Proper design of a deep learning system requires the data be split into a training set, used to build the model; a validation set, consisting of additional training data that is never directly fed to the al-

gorithm but is used for performance evaluation in the hyperparameter tuning phase; and a test set, used at the last stage to provide an unbiased estimate of performance [15]. Our splitting strategy for this project is reported in Table II. The criteria we tried to adhere to for performing such division were (i) maximization of training data, (ii) use of WSC and SSC data with multiple scorers' annotations as test data, (iii) balance between the three cohorts across subsets, (iv) random selection whenever possible. The demographic data of the test set are provided in Supplementary Table 1.

Table II Training-Validation-Test split per dataset.				
Subset	Training	Validation	Test	Total
WSC	235	10	30	275
SSC	137	10	30	177
MrOs	283	33	32	348

2.2. Algorithm Overview: Figure 1 describes the architecture of our algorithm. The LAT/RAT EMG channel is the input of the system we devised, although, as mentioned, in one cohort the signals were derived through subtraction (MrOS). Our system outputs the conditional probability of a LM given an input signal $P(LM|x)$. Before the signal is fed to the deep neural network shown in Figure 1, the signal is preprocessed to reduce any performance-compromising artifact contributions. Briefly, the architecture of the neural network is designed to perform successive algebraic manipulations of the input data, that combine linear and non-linear functions, that progressively transform the dimensionality of the input signal to map it into a probability space with a resolution of 0.5 s. Post-processing of output probability predictions is performed to eliminate segments with no relevance, such as wake epochs. Performance is evaluated relative to the ground truth of preexisting annotations, and with respect to the multiple scorers.

Figure 1 - Deep neural network used for leg movement detection. The network takes a matrix containing 32-time steps of 0.5s as input and manipulates it to obtain a label for each time step as output. The network is made up of two convolutional layers, one feedforward followed by a long short-term memory cell, two fully connected that feed their output into a sigmoid function that returns the probability of leg movement. The figures that follow are a more detailed representation of the dimensionality transformations occurring in the network, including the minibatch size dimension. What we represent in this current figure is the path followed by one single example.

2.3. Signal Preprocessing pipeline: Briefly, the merged LAT/RAT EMG signal was obtained through subtraction after applicable high pass filtering on each separate channel. ECG artifacts in the LAT/RAT were removed using an adaptive recursive least squares filter. Due to varying sampling frequencies for the LAT/RAT EMG across datasets and within the same dataset for WSC and SSC, we chose 100 Hz as the reference f_s and resampled all channels, where needed, at that frequency.

2.3.1. High-Pass Filtering: AASM guidelines recommend the use of a band-pass filter with low-frequency cutoff at 10 Hz and high-frequency cutoff at 100 Hz for EMG [14]. Because some of our data had sampling frequencies as low as 64 Hz, we simply applied a high-pass filter with cutoff at 10 Hz. The specific implementation was a 4th order infinite impulse response (IIR) Butterworth filter. To overcome issues linked to group delay, we applied the designed filter with a zero-phase filtering procedure, as employed in Elci et al. [19].

2.3.2. Adaptive Filtering of ECG Interference: ECG artifacts within the LAT/RAT EMG are a

major source of interference [10], which were removed using adaptive filters inspired by Moore et al. [10]. An analogous method was employed in He et al. [20] for removal of ocular interference in the electroencephalogram. The filter was based on a recursive least square algorithm due to computational efficiency and effectiveness. A configuration with $\lambda = 0.995$ and filter order of $p = 4$ was selected based on previous work [10].

2.3.3. Resampling: Resampling was needed to input LAT/RAT EMG signals with the same f_s configuration while allowing to consistently have a resolution of 0.5 s. LAT/RAT EMG can have f_s higher, equal or lower than reference frequency we selected for the system, namely 100 Hz. The resampling step is performed, where necessary, by means of (i) the application of an anti-aliasing low pass filter followed by down sampling of the signal ($f_s > 100$ Hz) or (ii) up sampling the signal through interpolation ($f_s < 100$ Hz).

2.3.4. Normalization: Normalization of LAT/RAT EMG was crucial to enhance training efficiency. A narrower distribution of data effectively speeds up training and is suggested for deep learning applications [15]. We normalized the filtered and resampled LAT/RAT EMG through subject standardization, so that each subject EMG was distributed centered at zero, with a standard deviation of one.

2.3.5. Creation of Examples for the Network: The deep neural network was fed using examples created by slicing the entire recording of each subject into segments of 16 s, chosen for computational convenience, being 16 a power of 2. Such slices were then rearranged to form 3D matrices with consecutive time-steps of 0.5 s stacked together. Consequently, the dimensionality of each example was $32 \times 50 \times 1$. 32 corresponds to the number of 0.5 s time-steps contained within the 16 s segment, 50 is the number of samples in 0.5 s, coming from a f_s of 100 Hz, and 1 represents the number of channels, as it is conventional when describing, for instance, color images. In our

case we had one single channel for the example matrices.

2.4. Deep Neural Network Design: The deep neural network used for determining $P(LM|x)$ was characterized by convolutional layers, feedforward layers and a recurrent layer. First, convolutional layers were applied on the filtered signal, in the form of matrix examples. They extracted relevant information from the “raw” data through a procedure that can, somehow, be related to an “automatic feature extraction”, in which the model learns the relevant features of the signal when it is fed with the data. After being mapped into a transformed feature space, the data undergoes a feedforward layer, which feeds its outputs into a recurrent layer, constituted by a long short-term memory (LSTM) cell, a specific kind of gated recurrent unit that overcomes the issue of vanishing and exploding gradient that usually affects recurrent networks [15]. Two more feedforward layers followed, mapping the data into one single number for each of the 32 time-steps constituting the input matrix. These correspond to $P(LM|x)$ for each single time-step, resulting into a resolution of 0.5 s. An overview of the network we designed is presented in Figure 1. The structure of the network was determined through a trial-and-error approach. We started with a simple structure and progressively added complexity, both in terms of architecture (specifically memory components) and in terms of computational effort. The constraints we had were solely determined by input and output shapes. We implemented a random search procedure for hyperparameter selection (e.g. the number of hidden units in the feedforward layers, L2 regularization scale, learning rate). Manual tuning of hyperparameters or grid search were deemed impractical for our application, as pointed out by Goodfellow et al. [15]. The network was fed with minibatches of examples (i.e. subsets of data randomly drawn from the training set). The minibatch size is one additional hyperparameter tuned through random search.

Figure 2 - Dimensionality transformations throughout the network. In each panel, the red box points out the part of the overall network corresponding to the layer(s) represented. The recurring 64 corresponds to the minibatch size used in training, while 32 is the number of time steps.

Panel A - Convolutional layers taking the signal as input and mapping it into a set of 31×50 features for each time step. That is achieved by first using 20 convolutional kernels of size 10×1 with zero padding to keep the last dimension constant and stride 1, then 50 kernels of size 20×1 without zero padding and stride 1. The latter shrink the width of the inputs. Both layers use biases and rectified linear unit as activation and batch normalization is applied after each of them.

Panel B - The output of the convolutional layers is flattened, preserving the number of time steps and the batch size. In this step, the dimension of the data goes from $64 \times 32 \times 31 \times 50$ to $64 \times 32 \times 1550$. The flattened input goes through the first dense layer made up of 128 units. That changes the last dimension to the number of units in the feedforward layer, resulting in $64 \times 32 \times 128$.

Panel C - The first transformation shown is the long short-term memory layer of the neural network. The output of the first fully connected layer, of size $64 \times 32 \times 128$, goes through the long short-term memory cell. The innermost dimension is changed to the number of units within such cell, transforming data into $64 \times 32 \times 100$. The second transformation displayed corresponds to the second fully connected layer, containing 128 units. That outputs $64 \times 32 \times 128$ data.

Panel D - The output of the second fully connected layer goes through a hidden layer with one single fully connected unit, then a sigmoid activation is applied to map each number to a probabilistic $[0;1]$ range. The output of this operation returns a dimensionality of $64 \times 32 \times 1$.

2.4.1. Convolutional Layers: The first part of the neural network used for LM detection was made up of two convolutional layers. Such choice is convenient since these are adept at handling array-like data and have computational efficiency generated by parameter-sharing. In the first layer, convolution was performed using 20 filters whose size is 10×1 using a stride of 1. Some zero padding was applied to keep constant the innermost dimension of the input. The activation function was the rectified linear unit (ReLU) and biases were used. The output of this convolutional layer went through batch normalization, a practice used to avoid the issue of covariate shift in deep networks [15]. The second convolutional layer was characterized by 50 filters of size 20×1 , employed with a stride of 1 and no zero padding. The activation function is the ReLU and biases were used in this case as well. Batch normalization took place on the output. A visual description of these two convolutional layers is contained in Figure 2, Panel A. The input of the convolutional layers are 3D matrices. The convolution operation was carried out using filters with a width of 10 and 20, respectively, for the first and second layers, but in both cases, these were characterized by a height of 1. This allows to be processed the data in a way that coincides with 1D convolution so that only time information flows through the network. The use of 3D matrices, on the other hand, was in line with computational efficiency given that we used the deep learning framework TensorFlow for this work. Specifically, we employed the Conv2D function to help us training in minibatches.

2.4.2. First Feedforward Layer: The first step after application of the two convolutional layers was a flattening of their output, that squeezes the two innermost dimensions into one. This operation leaves unchanged the number of time-steps (32) as well as the batch size (64) but compresses the dimensionality of the feature map of each time-step into one. That step is needed to make the data compatible with the application of a feedforward layer. This component is constituted by 128 units with ReLU activation. The regularization methods used were L2, with a regularization scale of 0.29586 and dropout, with a keep-probability of 0.56893. The output of the feedforward layer un-

dergoes batch normalization. The fully connected layer, in terms of dimensionality transformations, only operates on the innermost dimension, turning it into the number of units it contains. Its output is consistently $64 \times 32 \times 128$. How dimensions were manipulated from the output of the convolutional layer to the output of the first feedforward one, is depicted in Figure 2, Panel B.

2.4.3. LSTM Layer: LSTM cells are gated recurrent units first described in Hochreiter and Schmidhuber [21] that overcome the issue of vanishing or exploding gradient that usually challenges the training of recurrent networks. In this work, we used the implementation found in Graves et al. [22] for such component. The LSTM layer processes data with a time series-like structure, so that the temporal information of some past input $x_{n-\tau}$ is exploited for outputting a more accurate prediction y_n . Such layer is shown as the first transformation in Figure 2, Panel C and contains 100 hidden units in its gates, therefore the data dimensionality is turned from $64 \times 32 \times 128$ into $64 \times 32 \times 100$.

2.4.4. Second Feedforward Layer: The output of the LSTM layer undergoes the manipulations of the second fully connected layer of the system, which was characterized by the same architectural choices as the first one: 128 units, ReLU activation, L2 regularization with a scale of 0.29586, dropout with keep-probability of 0.56893. This layer changes the shape of the data from $64 \times 32 \times 100$ to $64 \times 32 \times 128$, as visible in the second operation in Figure 2, Panel C. This feedforward layer's output is then subject to batch normalization.

2.4.5. Probability Layer: The last layer of the network was what we called here a probability layer, a fully connected layer with a single unit and a sigmoid activation. Such layer presents no regularization methods. The aim of this layer is to map the 128 features coming from the second dense layer of the system, belonging to each of the 32 time steps of each example in the minibatch, into a single number corresponding to the probability of having a LM in that specific time step. The di-

mensions of the data are consistently transformed from $64 \times 32 \times 128$ to $64 \times 32 \times 1$ (see Figure 2, Panel D). The activation function is sigmoid, taking real numbers to map them into a $[0,1]$ probabilistic range.

2.4.6. Cost Function: Weighted cross entropy was selected as the cost function of the network. The weight of the positive examples was determined as a random search hyperparameter and set to 3.8183. This choice handles the class imbalance problem (most of the examples have non-LM labels). Equation 1 describes the cost function we used.

$$J(\theta) = \sum_{i=1}^N p \cdot y^i \{ -\log(\sigma(\hat{y}^i)) \} + (1 - y^i) \cdot \{ -\log(1 - \sigma(\hat{y}^i)) \} \quad (\text{Equation 1})$$

Where p is the weight to positive observations, y the annotated label, $\hat{y}$ the predicted label and θ the model's parameters.

As we used L2 regularization for the first and second feed forward layers, the complete expression of the cost function includes regularization terms, as shown by Equation 2.

$$J^*(\theta) = J(\theta) + \frac{\alpha}{2} ||\mathbf{W}_1||_2^2 + \frac{\alpha}{2} ||\mathbf{W}_2||_2^2 \quad (\text{Equation 2})$$

Where $\mathbf{W}_1$ and $\mathbf{W}_2$ refer to the weight matrices of the two dense layers and α is the regularization scale.

2.4.7. Other Architectural Choices: Other architectural choices of this work include the optimization algorithm chosen to tackle the problem, which in our case was Adam, an algorithm first described in Kingma and Ba [23]. This consists in computing the momentum of the rescaled gradients and it is suggested as one of the default choices for deep networks [15].

The learning rate, a crucial hyperparameter, was tuned to 10^{-4} , the recommended default choice for the Adam optimizer is 10^{-3} .

Initializers for weights, including kernels of the convolutional layers, were *Xavier initializers*, proved to be a robust choice for enhancing training [24]. Biases were simply initialized to 0.

The network was trained using early stopping, which is an algorithm that ceases training in case the validation error does not decrease below its lowest recorded value for an established amount of iterations called *patience*; selected weights and biases for the model are then those that generated the lowest validation error. Post processing of predictions consisted of hard coding to 0 all the output from the epochs scored as wake. This choice might be debatable, for the WASM guidelines specify that PLM trains can include events in wake [7]. On the other hand, we sought consistency with the multiple scorers, for these technicians were specifically instructed to score *all* LMs during sleep and neglect events occurring in wake.

2.5. Performance Evaluation: Performance evaluation with respect to preexisting annotations took place through the *event-wise* receiver operating characteristic curve (ROC) analysis. By event-wise, we mean that *any* overlap between predicted and annotated events was considered as a true positive (TP). On the contrary, false positives (FP) and false negatives (FN) were, respectively, predicted but not manually annotated events and LMs scored by the human technician but not predicted by the model. Evaluation took place through precision (or positive predictive value) and recall (or sensitivity) metrics. Resulting F1 score was considered most suited for the selection of the optimal detection threshold. To do this, ROC analyses were carried out on the validation set and an optimal threshold selected. This optimal F1 threshold was then applied to the test data, which included the multiscorer cohort analysis.

2.6. Multiscorer Analysis: Ideal comparisons of human scorer and deep learning algorithm performance must be carried out with respect to a consensus of scorers, assumed to be the ground truth. For scorers, however, each of them must be compared to the consensus of the others in a *leave-one-technician-out* analysis. As each recording in the WSC and SSC test subsets was scored by five dif-

ferent trained technicians, each recording was used to compare performance of one technician at the time with the deep learning algorithm performance using a consensus based on the annotations of the other four technicians (see Figure 3). This was obtained by establishing a majority m , in our case we selected $m = 2$, and by including in the consensus with a weight of 1 only the events scored by at least m technicians. One might dispute the choice of setting $m = 2$. We proceeded in this way because we observed a significant tendency of the technicians to underscore LMs, most likely because there are so many LMs to score in comparison to only PLMS that fatigue can occur. We carried out such comparison both event-wise (i.e. considering any overlap between scored event in the evaluated scoring and in the consensus) as well as time step-wise (i.e. considering the specific time steps of 0.5 s). Additionally, for the event-wise analysis, we addressed the potential issue of artificially overcounting LMs: for instance, if our algorithm had detected two LMs within the time interval of one single LM present in the consensus, only one TP would have been counted. The metrics used for this analysis were precision, recall and F1 score.

Figure 3 - *Leave-one-technician-out* analysis for one single sleep study scored by five technicians from the pool of nine. First, one of the five scorers is put aside, the one in blue, and the remaining four, in green, are used in the following step to obtain the consensus. The performance of the technician left out is then compared against the deep learning model's one, considering the other four technicians' consensus as ground truth.

2.7. Performance in comparison to rule-based detectors: Prior detectors have used rule-based decision in the scoring of LM or PLMS. One of the most performant of such detector is the Stanford Periodic Leg Movement Detector (SPLMD) described by Moore et al. [10]. In a final analysis, we compared performance of our deep learning detector with that of other detectors [6], [8], [9] including the SPLMD [10] in a subset of 60 subjects from the WSC. The comparison follows the same approach as described by Moore et al. [10], in which LMs co-occurring with respiratory events were excluded to comply with the AASM guidelines for PLMS [14]. Specifically, LMs were excluded if they overlapped with an exclusion windows at $[-5.0, 0.5]$ or $[-0.5, 5.0]$ seconds at the onset and offset of respiratory events, respectively [10].

3. RESULTS

3.1. Machine Learning Based PLMS detection: Results of this work were obtained using the described network with weights and biases determined by training on our datasets with the use of early stopping. Figure 4 includes real-life examples of the network’s output. Both show the model’s predictions with respect to the LAT/RAT EMG input (and the hypnogram with the sleep stages of the subject). Particularly, Figure 4 Panel B emphasizes the robustness of our method in relation to ECG interference, showing a case in which, although substantial electrical heart activity overlaps to LAT/RAT EMG signal and adaptive filtering preprocessing fails in filtering the interference properly, the deep neural network still makes accurate predictions. Code, weights and model for our machine learning based PLM detector will be available under a common license rule at <https://github.com/Stanford-STAGES/lm-detector> upon publication of this paper.

3.2. Performance on the three cohorts: Performance of our deep learning algorithm on the validation set for LM detection (without periodicity evaluation) is shown in Figure 5. The displayed ROC curve shows that the threshold that maximizes F1 score is 0.81. This threshold was therefore applied to obtain precision, recall and F1 relative to LMs without periodicity on the test set and corresponding statistics reported in Table III. These detections were next submitted to WASM rules for detection of PLMS to obtain metrics shown in Table IV.

Table III□□Performance of deep learning model on the three test subsets for leg movements.			
Subject Cohort	WSC	SSC	MrOs
Precision	0.79	0.70	0.73
Recall	0.87	0.72	0.83
F1 score	0.83	0.71	0.77

Table IV Performance of deep learning model on the three test subsets for periodic leg movements during sleep.			
Subject Cohort	WSC	SSC	MrOs
Precision	0.80	0.61	0.73
Recall	0.83	0.78	0.82
F1 score	0.81	0.69	0.77

3.3. Performance in comparison to human scoring: The results of our multiscorer analysis using majority voting consensus as ground truth are shown in Table V (event-wise analysis) and in Table VI (time step-wise analysis). In both comparisons, the model performs significantly better than two scorers, while no mean difference in F1 score is seen for the remaining seven scorers.

3.4. Performance in comparison to rule-based detectors: The results from the PLMS performance comparison between our machine learning based method and the previous algorithms on a subset of 60 WSC test subjects are reported in Table VII. Our algorithm shows high recall (second only to Wetter [6]) and a high precision (secondary only to Moore [10]), while achieving the best overall performance with a F1 score of 0.85. Furthermore, the Bland-Altman plot of the PLMi obtained through manual annotations PLMS and through the events predicted by the model is shown in Figure 6.

Table V Results of the multiscorer analysis obtained using the majority voting consensus and precision, recall and F1 metrics, evaluating event-wise. The model significantly outperforms two manual scorers while performing comparable to the remaining 7 with respect to the majority voting consensus in terms of F1 score. A two-sample t-test was used to calculate p -values with the null hypothesis being equal means while assuming equal but unknown variances, and those below 0.05 are highlighted as significant.

Metric	Precision	Recall	F1 score	p -val (F1)
Technician A	0.645 ± 0.307	0.599 ± 0.325	0.568 ± 0.312	0.597
Deep Learning model VS A	0.576 ± 0.252	0.732 ± 0.216	0.603 ± 0.218	
Technician B	0.885 ± 0.153	0.478 ± 0.249	0.588 ± 0.244	0.0832
Deep Learning model VS B	0.657 ± 0.228	0.765 ± 0.166	0.681 ± 0.181	
Technician C	0.596 ± 0.246	0.748 ± 0.298	0.586 ± 0.280	0.562
Deep Learning model VS C	0.568 ± 0.246	0.786 ± 0.190	0.622 ± 0.213	
Technician D	0.525 ± 0.274	0.777 ± 0.222	0.590 ± 0.262	0.682
Deep Learning model VS D	0.594 ± 0.291	0.762 ± 0.163	0.616 ± 0.260	
Technician E	0.511 ± 0.269	0.735 ± 0.238	0.556 ± 0.267	0.656
Deep Learning model VS E	0.530 ± 0.270	0.801 ± 0.156	0.584 ± 0.231	
Technician F	0.874 ± 0.199	0.281 ± 0.237	0.365 ± 0.253	6.23e-06
Deep Learning model VS F	0.612 ± 0.247	0.766 ± 0.172	0.649 ± 0.213	
Technician G	0.667 ± 0.227	0.693 ± 0.244	0.635 ± 0.197	0.331
Deep Learning model VS G	0.550 ± 0.262	0.729 ± 0.196	0.585 ± 0.222	
Technician H	0.713 ± 0.278	0.594 ± 0.262	0.604 ± 0.251	0.489
Deep Learning model VS H	0.613 ± 0.237	0.755 ± 0.158	0.642 ± 0.201	
Technician I	0.667 ± 0.286	0.270 ± 0.262	0.322 ± 0.286	8.86e-09
Deep Learning model VS I	0.651 ± 0.204	0.788 ± 0.157	0.692 ± 0.160

Figure 5 - Left panel: Receiving operator curve obtained on the entire validation set (black line), together with the ones obtained for each individual validation subset (blue line for Wisconsin Sleep Cohort, red line for Stanford Sleep Cohort, green line for MrOS). The point corresponding to the optimal threshold (i.e. the one that maximizes the F1 score on the entire validation set) is plotted in cyan. Right panel: F1 score plotted against the threshold used for thresholding the probability output by the model for the whole validation set (in black) as well as for the single validation subsets (same color code as the left panel). The point maximizing the F1 score is plotted in cyan.

Figure 6 - Bland-Altman plot for periodic leg movements index of the subset of 60 Wisconsin Sleep Cohort subjects used in the comparison with the previous algorithms [5], [6], [7], [8].

Table VI Results of the multiple scorers analysis obtained using the majority voting consensus and precision, recall and F1 metrics, evaluating time step-wise. The model significantly outperforms two manual scorers while performing comparable to the remaining seven with respect to the majority voting consensus in terms of F1 score. A two-sample t-test was used to calculate p -values with the null hypothesis being equal means while assuming equal but unknown variances, and those below 0.05 are highlighted as significant.

Metric	Precision	Recall	F1 score	p -val (F1)
Technician A	0.582 ± 0.293	0.495 ± 0.271	0.480 ± 0.259	0.381
Deep Learning model VS A	0.580 ± 0.245	0.568 ± 0.211	0.530 ± 0.193	
Technician B	0.808 ± 0.187	0.407 ± 0.207	0.513 ± 0.211	0.184
Deep Learning model VS B	0.646 ± 0.215	0.568 ± 0.187	0.576 ± 0.166	
Technician C	0.438 ± 0.193	0.654 ± 0.284	0.462 ± 0.234	0.119
Deep Learning model VS C	0.548 ± 0.229	0.619 ± 0.198	0.545 ± 0.187	
Technician D	0.487 ± 0.242	0.648 ± 0.215	0.523 ± 0.214	0.939
Deep Learning model VS D	0.578 ± 0.279	0.570 ± 0.171	0.527 ± 0.210	
Technician E	0.497 ± 0.232	0.588 ± 0.225	0.500 ± 0.214	0.833
Deep Learning model VS E	0.522 ± 0.237	0.607 ± 0.181	0.510 ± 0.179	
Technician F	0.743 ± 0.232	0.241 ± 0.206	0.310 ± 0.222	1.37e-05
Deep Learning model VS F	0.593 ± 0.232	0.566 ± 0.182	0.547 ± 0.185	
Technician G	0.428 ± 0.191	0.677 ± 0.235	0.482 ± 0.169	0.404
Deep Learning model VS G	0.536 ± 0.249	0.579 ± 0.180	0.517 ± 0.179	
Technician H	0.558 ± 0.240	0.525 ± 0.201	0.505 ± 0.198	0.244
Deep Learning model VS H	0.590 ± 0.208	0.591 ± 0.157	0.557 ± 0.166	
Technician I	0.470 ± 0.216	0.261 ± 0.256	0.270 ± 0.236	2.04e-09
Deep Learning model VS I	0.639 ± 0.184	0.604 ± 0.171	0.598 ± 0.143	

Table VII Results of the comparison between the previous algorithms and our machine learning based method on a subset of 60 Wisconsin Sleep Cohort subjects [8].

Metric	Tauch. [8]	Wett. [6]	Ferri [9]	Moore [10]	Moore SNR+ [10]	New ML method
Precision	0.79	0.47	0.62	0.88	0.85	0.81
Recall	0.24	0.96	0.85	0.60	0.73	0.90
F1 score	0.37	0.63	0.72	0.71	0.79	0.85

4. DISCUSSION

In this work, we design and validate a deep learning model to score LMs and PLMS that achieved robust performance in comparison to manual scoring, showing F1 score of 0.770 ± 0.049 for LM and 0.757 ± 0.050 for PLMS detection on three different datasets. To achieve this performance, these datasets were purposefully scored for LMs and not only for PLMS using experienced polysomnography technicians. LM annotations underwent an evaluation of periodicity (implemented according to the WASM periodicity rule) and used as ground truth for PLMS. Table III and Table IV show that recall is higher than precision, signifying that the model tends to be very inclusive in assigning probability of LMs. A risk of a highly inclusive scoring is an increased number of FP, which might be caused by a noisy signal or a minor muscular event, e.g. motor unit potentials or short muscle activations/twitches.

We used the detector in 60 PSGs drawn from WSC and SSC, each scored by five independent technicians drawn from a pool of nine. Comparing the performance of the technicians and of the machine learning algorithm with consensus scoring, we found that our algorithm outperformed two of nine technicians while performing comparable to the remaining seven technicians in an event-

wise analysis and in all cases when considering 0.5s time-steps. The results of the multiple scorers' analysis prove the major point that we would like to make in this work: Human scoring may present high inter-scorer variability, which may be avoided using automatic scoring. Inter-scorer agreement is rather low for LMs in the analysis we carried out on the test sets with multiple annotations from WSC and SSC.

For automatic detection of leg movements, many rule-based LM detectors have been implemented and are often used in clinical practice prior to conformation through manual inspection. Moore et al. carried out a comparison between the newly implemented algorithms and the previous ones for PLMS detection on a subset of 60 test subjects from WSC, the results of which are reported in table VII [10]. Our proposed method achieved the highest F1 score of the examined algorithms, thereby outperforming previous rule-based methods. The deep learning approach used in our method allows the model to learn complicated non-linear patterns unlike the preceding rule-based approaches. Previous works were based on criteria including voltage thresholding and time bridging [6], features such as spikes repetition rate, average amplitude and duration [8], techniques such as rectification followed by double thresholding [9]. Also, signal processing methods including adaptive filtering of ECG interference (the same as the one used in this work), noise floor detection and consequent adaptation of the threshold used for the EMG were employed [10]. In this study, several data cohorts were included with different acquisition equipment, demographics and scorings technicians. Therefore, we believe that our methods are more robust and have better prospects of generalizing to new data compared to rule-based methods developed on single cohorts.

The metrics used in this evaluation were influenced by the imbalance between positive observations (i.e. LMs) and negative observations (i.e. segments without LMs). The latter were the vast majority of all subjects' recordings. Consequently, the use of certain metrics, such as accuracy, were excluded, since they would be biased by the high number of true negatives.

Some limitations of this study deserve to be discussed more in detail. The inclusion of the WSC and the SSC data, whose acquisition and scoring followed the clinical practice at the time of recording,

introduced three substantial limitations to the algorithm. Firstly, the scoring criteria have been updated since the time of scoring [7]; secondly, in some of the recordings, there was no annotation of the events in wake epochs; thirdly, the use of a merged left/right anterior tibialis EMG is not current clinical practice. The ideal way to address the first two issues would have been re-scoring the subjects whenever needed, which, unfortunately, was not feasible given the timeline of this study. Regarding the issue of merged legs channels, this limitation might be overcome through re-training of our network for LM detection on newly-scored, separate-legs data. Alternatively, we could have discarded the WSC and SSC data. However, we felt that would have limited the diversity and the size of our dataset.

As briefly mentioned in Section 2.4, the network construction took place through a trial-and-error approach which, of course, is not ideal, as it is hard to justify why our specific network architecture is a better choice than any other. Deeper or shallower networks could present increased performance or computational advantages, but we settled with the structure that seemed a fair trade-off between the two. An interesting investigation could be focusing on increasing performance through exploring more complex and deeper network structures.

Additionally, the achieved performance can be positively impacted by increasing the number of subjects included in the study. LMs are events whose occurrence can vary substantially from subject to subject and it is possible that 655 subjects from three different datasets is not enough data for this specific application.

Furthermore, it could be argued that our choice of filtering according to the AASM guidelines and the addition of the adaptive filtering module to cancel ECG interference was rather conservative. In fact, a system that bypasses filtering of the data can perform as well, or perhaps better, than the system we designed.

The results show that our method outperforms previous rule-based methods and performs comparable or better than individual human scorers, thereby avoiding the inter-scorer variability issue. Thereby, we expect that this method can be adapted with similar results for revised scoring

definitions and to function with multiple leg EMG channels. Based on these results, we would further like to stress is that automatic algorithms can provide substantial support to the task of scoring of sleep recordings, reducing the manual work of the sleep technicians and minimizing the inter-scorer variability issue.

5. CONCLUSIONS

This work presents a novel deep learning-based method for detection of LMs and PLMS that can be used as a support tool for manual scorers in research and in clinical practice. The dataset used was built from subsets of three different cohorts: WSC, SSC and MrOS, totaling 800 subjects. These have heterogeneous demographics, different sampling frequencies, methods of acquisition, scorers and purposes (research and clinical).

Signal preprocessing following the AASM guidelines was performed, together with cleaning of the ECG interference through adaptive filtering. The data were used to design and validate a neural network for LMs detection, without considering periodicity, and the evaluation of that was added on top of the LMs predictions, following the WASM guidelines.

A comparison against human scoring and subsequent analysis of inter-scorer variability was carried out, using a subset of data scored by multiple technicians. Our method achieved a performance better or comparable to that of expert technicians.

We further compared our method with previously existing algorithms, which yielded state of the art results for our algorithm.

In conclusion, the algorithm can be used as a tool for support and assist the manual scorer in the research or clinical field. The designed algorithm may have significant implications in future sleep evaluation studies to reduce analysis time.

6. DECLARATION OF INTEREST

7. ACKNOWLEDGEMENTS

L. Carvelli has received funding from Otto Mønstedts Fond; W. Demant og Hustru Ida E, Fond; Marie og M.B. Richters Fond; Reinholdt W. Jorck og Hustrus Fond.

A. N. Olesen has received funding from The Klarman Family Foundation; Technical University of Denmark; University of Copenhagen; Reinholdt W. Jorck og Hustrus Fonden; Otto Mønstedts Fond; Stibofonden; Knud Højgaards Fond; Augustinus Fond; and Vera og Carl Johan Michaelsens Fond.

REFERENCES

- [1] Ferri R, Koo BB, Picchiatti DL and Fulda S. Periodic Leg Movements During Sleep: Phenotype, Neurophysiology, and Clinical Significance. PLoS One, 31:29–38, December 2017.
[doi: 10.1016/j.sleep.2016.05.014](https://doi.org/10.1016/j.sleep.2016.05.014)
- [2] May AM, Blackwell T, Stone KL, Cawthon PM, Sauer WH, Varosy PD, Redline S, Koo BB, Mehra R; Osteoporotic Fractures in Men (MrOS) Study Group. Longitudinal relationships of periodic limb movements during sleep and incident atrial fibrillation. Sleep Med. 2016 Sep; 25:78-86.
[doi: 10.1016/j.sleep.2016.08.009](https://doi.org/10.1016/j.sleep.2016.08.009)
- [3] Yumino D, Wang H, Floras JS, Newton GE, Mak S, Ruttanaumpawan P, Parker JD, Bradley TD; Relation of Periodic Leg Movements During Sleep and Mortality in Patients With Systolic Heart Failure. The American Journal of Cardiology, 107(3):447-451. [doi: 10.1016/j.amjcard.2010.09.037](https://doi.org/10.1016/j.amjcard.2010.09.037)
- [4] Skomro R, Silva R, Alves R, Figueiredo A, Lorenzi-Filho G; The prevalence and significance of periodic leg movements during sleep in patients with congestive heart failure. Sleep Breath. 2009 Mar;13(1):43-7. [doi: 10.1007/s11325-008-0207-5](https://doi.org/10.1007/s11325-008-0207-5).
- [5] Hardy De Buisseret FX, Mairesse O, Newell J, Verbanck P, Neu D. While Isolated Periodic Limb Movement Disorder Significantly Impacts Sleep Depth and Efficiency, Co-Morbid Restless Leg Syndrome Mainly Exacerbates Perceived Sleep Quality. Eur Neurol. 2017; 77(5-6):272-280.
[doi: 10.1159/000471920](https://doi.org/10.1159/000471920)
- [6] Wetter TC, Dirlich G, Streit J, Trenkwalder C, Schuld A and Pollmacher T. An Automatic Method for Scoring Leg Movements in Polygraphic Sleep Recordings and Its Validity in Compari-

son to Visual Scoring. *Sleep*, 27(2):324–328, March 2004.

[doi: 10.1093/sleep/27.2.324](https://doi.org/10.1093/sleep/27.2.324)

[7] Ferri R, Fulda S, Allen R, Zucconi M, Bruni O, Chokroverty S, Ferini-Strambi L, Frauscher B, Garcia-Borreguero D, Hirshkowitz M, Hogl B, Inoue Y, Jahangir A, Manconi M, Marcus CL, Picchietti DL, Plazzi G, Winkelmann JW, and Zak RS. The official World Association of Sleep Medicine (WASM) 2016 standards for recording and scoring leg movements in polysomnograms developed by a joint task force from the International and the European Restless Legs Syndrome Study Group (IRLSSG and EURLSSG). *Sleep Medicine*, 26(2):86–95, October 2016.

[doi: 10.1016/j.sleep.2016.10.010](https://doi.org/10.1016/j.sleep.2016.10.010)

[8] Tauchmann N and Pollmacher T. Automatic detection of periodic leg movements (PLM). *Journal of Sleep Research*, 5(4):273–275, 1996.

[doi: 10.1111/j.1365-2869.1996.00273.x](https://doi.org/10.1111/j.1365-2869.1996.00273.x)

[9] Ferri R, Zucconi M, Manconi M, Bruni O, Miano S, Plazzi G and Ferini-Strambi L. Computer-Assisted Detection of Nocturnal Leg Motor Activity in Patients with Restless Legs Syndrome and Periodic Leg Movements during Sleep. *Sleep Medicine*, 28(8):998–1004, August 2005.

[doi: 10.1093/sleep/28.8.998](https://doi.org/10.1093/sleep/28.8.998)

[10] Moore H, Leary E, Lee SY, Carrillo O, Stubbs R, Peppard P, Young T, Widrow B and Mignot E. Design and Validation of a Periodic Leg Movement Detector. *PLoS One*, 12(9):1–30, December 2014.

doi: 10.1371/journal.pone.0114565

[11] Huang AS, Skeba P, Yang MS, Sgambati FP, Earley CJ and Allen RP. MATPLM1, A MATLAB script for scoring of periodic limb movements: preliminary validation with visual scoring. *Sleep Medicine*, 16(4):1541–1549, April 2015.

doi: 10.1016/j.sleep.2015.03.008

[12] Alvarez-Estevéz D, Wahid D, Rijsman RM. Validation of an automatic scoring algorithm for the analysis of periodic limb movements according to the WASM2016 guidelines. *Sleep Medicine* 40, December 2017.

doi: 10.1016/j.sleep.2017.11.032

[13] Stefani A, Heidbreder A, Hackner H, and Högl B. Validation of a leg movements count and periodic leg movements analysis in a custom polysomnography system. *BMC Neurol.* 2017; 17: 42.

doi: 10.1186/s12883-017-0821-6

[14] Berry RB, Brooks R, Gamaldo CE, Harding SM, Lloyd RM, Marcus CL and Vaughn BV. The AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specifications. Darien, Illinois, American Academy of Sleep Medicine, 2012.

url: <https://aasm.org/clinical-resources/scoring-manual/> (last visited: June 17th, 2019)

[15] Goodfellow I, Bengio Y and Courville A. *Deep Learning*. The MIT Press, Cambridge, Massachusetts, 2016.

url: <http://www.deeplearningbook.org> (last visited: June 17th, 2019)

[16] Young T, Finn L, Peppard P, Szklo-Coxe M, Austin D, Nieto FJ, Stubbs R and Hla KM. Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort.

Sleep, 31(8):1071–1078, 2008.

url: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2542952/> (last visited: June 17th, 2019)

[17] Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Ensrud KE, Stefanick ML, Laffan A and Stone KL. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the Osteoporotic Fractures in Men Sleep Study. *Journal of the American Geriatrics Society*, (12):2217–2225, 2011.

doi: [10.1111/j.1532-5415.2011.03731.x](https://doi.org/10.1111/j.1532-5415.2011.03731.x).

[18] Blank JB, Cawthon PM, Carrion-Petersen ML, Harper L, Johnson JP, Mitson E and Delay RR. Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemporary Clinical Trials*, (5):557–568, 2005.

doi: [10.1016/j.cct.2005.05.005](https://doi.org/10.1016/j.cct.2005.05.005)

[19] Elci H, Longman RW, Phan JN, Juang MQ and Ugoletti R. Simple Learning Control Made Practical by Zero-Phase Filtering: Applications to Robotics. *IEEE Transactions on Circuits and Systems*, 49:753–767, 2002.

doi: [10.1109/TCSI.2002.1010031](https://doi.org/10.1109/TCSI.2002.1010031)

[20] He P, Wilson G and Russel C. Removal of Ocular Artifacts from Electro-Encephalogram by Adaptive Filtering. *Medical and Biological Engineering and Computing*, (42):407–412, 2004.

doi: [10.1007/BF02344717](https://doi.org/10.1007/BF02344717)

[21] Hochreiter S and Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 9:1735–1780, 1997.

doi: 10.1162/neco.1997.9.8.1735

[22] Graves A, Mohamed AR, and Hinton G. Speech Recognition with Deep Recurrent Neural Networks. *ICASSP2013*, pages 6645–6649, 2013.

doi: 10.1109/ICASSP.2013.6638947

[23] Kingmaand DP, Ba JL. Adam: A Method for Stochastic Optimization. *ICLR 2015* 1–15, 2015.

url: arXiv:1412.6980 (last visited: June 17th, 2019)

[24] Glorot X and Bengio Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *PMLR*, 9:249–256, 2010.

url: <http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf> (last visited: June 17th, 2019)