Use of Deep Learning Techniques for Motor Events Detection in Polysomnographic Records

Inés Lloves García and Diego Álvarez Estévez

Faculty of Informatics, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: ines.lloves@udc.es, diego.alvareze@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.45*

Abstract: Sleep medicine deals with the diagnosis and treatment of sleep-related disorders. The diagnosis is carried out through the manual analysis and labeling of polysomnographic studies, which record various electrophysiological and pneumological signals of the patient throughout the night. This process involves the analysis of long duration signals, is complex, and demands considerable resources and time on the part of the clinical expert. The purpose of this proyect is the construction of automatic analysis algorithms that considerably reduce the analysis duration, reducing the manual workload, and minimizing possible human errors, providing repeatability and robustness. In particular, the objective is to use machine learning algorithms, based on Deep Learning techniques, for the identification and location of physiological events in these polysomnographic records. Specifically, the goal is to locate physiological events associated with involuntary motor movements that occur in the limbs, known as Limb Movements.

1 Introduction

Polysomnography is a study utilized to evaluate and diagnose sleep-related disorders. The polysomnographic recording (PSG) obtained, records multiple physiological signals during the night. The patient sleeps in a controlled environment, usually in a sleep laboratory, where electrodes and sensors are placed on various parts of the body to record these signals.

PSG is considered the gold standard for diagnosing sleep-related breathing disorders. Each recording lasts at least 8 hours, so manually analyzing each signal is a slow, time-consuming process, that generates high costs in human resources. This causes delays in diagnosis and increases the waiting lists, and therefore, motivates the research and development of automatic analysis tools for PSGs.

In the context of automatic detection of involuntary motor events, the reference signal is the Electromyogram (EMG) of the tibialis anterior muscles. EMG records electrical signals generated by muscle activity. Involuntary movements, known as Limb Movements (LM), often occur during sleep, especially during deep sleep phases, and can be identified by observing the muscle activity through an EMG signal.

Brief and repetitive episodes of LM during sleep are known as Periodic Limb Movements (PLM), which are found between the 85% and 95% of patients with Restless Legs Syndrome (Rye and Trotti, 2012), a neurological disorder characterized by an uncomfortable sensation in the legs and the urgent need to move them to relieve the sensation. These movements can disrupt sleep and cause insomnia problems. They have also been observed in other sleep-related neurological disorders, such as narcolepsy (Baker et al., 1986), sleep-related breathing disorders (Ancoli-israel et al., 1985), Parkinson's disease (Wetter et al., 2000), multiple system atrophy (Vetrugno et al., 2004) and REM sleep behavior disorder (Fantini et al., 2002). Therefore the importance of the detection and analysis of LM.

2 Material

For the development of this project, we have used PSGs annotated by clinical experts. Only the EMG signals from the PSG records were used. This signal is stored in two channels, with the signal originated from the right tibialis anterior muscle (EMG RAT) and from the left tibialis anterior (EMG LAT), which will be used as independent signals. Three different data sets have been used:

HMCP: comes from the Álvarez Estévez and Rijsman (2022) study. The study consists of 20 PSGs individually selected for 4 tasks. In this project, we have used the 5 PSGs selected in this study for the LM annotation task, which are individually annotated by 12 clinical experts. The EMG signals are stored in two separate channels for each leg, and are sampled at a frequency of 128 or 256 Hz.

MrOS: comes from the Osteoporotic Fractures in Men Study (MrOS) (Blank et al., 2005). 3,135 of the participants were recruited to take the MrOS sleep study, and underwent the recording of a complete PSG. A subset of 10 randomly selected PSGs has been used in this project. Each PSG includes annotations made by one clinical expert. All EMG signals are stored in two separate channels for each leg, and are sampled at a frequency of 64 Hz.

WSC: comes from a follow-up of mortality in the population of the Wisconsin Sleep Cohort (WSC) (Young et al., 2008). Within the large sample, only a subset of patients had all PSG awakenings and LMs scored including the duration of the events. In this project, 16 PSG records were randomly selected from that subset. The EMG signals from these recordings are sampled at a frequency of 200 Hz, and only the right leg channel is recorded.

3 Methods

The task of detecting LM in EMG signals is approached as a supervised sequence-to-sequence classification problem, where the classifier is a deep learning model. The objective is to locate the start and the end point of each LM in the EMG signals, and to determine the specific channel in which the event takes place (EMG RAT/LAT).

3.1 Signal preprocessing

EMG signals can be contaminated by several types of noise, interferences and artifacts. To avoid possible misinterpretation of the data, a suitable filtering system is implemented to prepare the EMG signals. The filtering pipeline consists of a high-pass filter with a 15 Hz cut-off frequency, to remove low-frequency components not related to the muscle activity of interest; and a Notch filter centered on 50 Hz (Europe) or 60 Hz (North America) (according to each data set origin), to eliminate power line interference.

Signals can be recorded at different sampling frequencies. Among the three databases used, we have signals sampled at 64, 128, 200 and 256 Hz. In order to have all the data in a single homogeneous format, a resampling of the EMG LAT/RAT channels is performed at 128 Hz.

3.2 Sample processing

Once the EMG LAT and RAT signals have been preprocessed, the process of obtaining the samples, to train the models and subsequently to make predictions, is carried out. Training samples are pairs **(X, Y)**, where:

X: is a vector with the amplitude values of an interval of the input EMG signal. To obtain these vectors from the EMG signals, windowing is performed using a overlapping

sliding time window. A window length of 30 seconds is used (clinicians usually analyze these signals in 30 second intervals), and an overlap of the 50%.

Y: is a boolean vector that corresponds to the annotations of an expert indicating the presence or absence of LM in the signal interval. To obtain the label vectors, a resolution of 0.25 seconds is used, that is, every 0.25 seconds the signal has an associated label indicating the presence or absence of LM.

Training samples are obtained from the EMG signals of the PSGs of the HMCP dataset, while the MrOS and WSC datasets are utilized to evaluate the generalization of the models between different databases.

3.3 Training process

For the training process, samples are divided into three partitions:

- **Training partition**: 70% of the pairs (X, Y), are utilized for adjusting and learning the models.

- **Validation partition**: 15% of the pairs (X, Y), are utilized to monitor the learning of the models.

- **Test partition**: remaining 15% of the pairs (X, Y), are utilized to evaluate the performance of the trained models.

Different architecture models have been trained, they are presented in Section 4.

3.4 Prediction and annotation

To predict LM from the EMG LAT/RAT signals, the preprocessing and processing steps are carried out to obtain the X input vectors. Once the predictions (Y vectors) are obtained from the trained models, they are processed to transform predicted labels into annotations that follow the LM annotation criteria established by the World Association of Sleep Medicine (WASM) (Ferri et al., 2016).

3.5 Validation

Cohen kappa (Cohen, 1960) will be used as the main reference metric to measure and compare the quality of the algorithms. It is a useful metric because it considers the correct and the incorrect classification, and corrects the agreement due to chance. It is frequently used to test inter-rater agreement, and is the main reference metric for annotating events in PSG signals.

4 Deep learning approaches

Different architectures have been evaluated for the LM detection task. Table 1 shows a summary of the trained models and the Kappa values obtained by each of them in prediction of the test partition data.

Table 1: Trained models summary

Model	Arquitecture	Parameters	Training samples	Cohen Kappa
Method 1	LSTM	9,217,880	only pairs (X,Y) from EMG RAT	0.6769
Method 2	LSTM	9,217,880	all pairs (X,Y)	0.7422
Method 3	LSTM	69,117,300	all pairs (X,Y)	0.7001
Method 4	BiLSTM	46,119,940	all pairs (X,Y)	0.6830
Method 5	1D CNN - BiLSTM	4,610,680	all pairs (X,Y)	0.7612
Method 6	1D CNN - BiLSTM	4,610,680	only pairs (X,Y) containing some LM	0.7219
Method 7	1D CNN - BiLSTM	4,610,680	pairs (X,Y) containing some LM and 10% of pairs not containing any LM	0.7391
Method 8	LSTM - BiLSTM	9,220,360	all pairs (X,Y)	0.7547

5 Results

We evaluated the performance of the trained models. Additionally, we evaluated the performance of the automatic annotation method of Álvarez Estévez (2016) (Polyman method), based on conventional signal processing methods, so as to compare its performance to our models. This method was designed with PSGs coming from the same medical center as HMCP.

5.1 HMCP

An inter-expert validation was carried out among the different HMCP clinicians, to know the reference measure of performance between experts. We measured the Kappa agreement between each clinical expert individually and the consensus formed by the rest of the experts, in the annotation of LM in every EMG signal. The average Kappa of agreement between experts was **0.7867**, with a deviation of $\pm$ **0.0920**.

Equally, the evaluation between the automatic annotation methods and the expert consensus was carried out (see Table 2) for each EMG signal. Additionally, a multiple comparison test was carried out to analyze the statistical significance of the differences between methods (see Figure 1).

Table 2: HMCP validation results

Model	Cohen Kappa
Method 1	0.7739 ± 0.0648
Method 2	0.8100 ± 0.0677
Method 3	0.7964 ± 0.0691
Method 4	0.7941 ± 0.0616
Method 5	0.8217 ± 0.0450
Method 6	0.6889 ± 0.3148
Method 7	0.8285 ± 0.0350
Method 8	0.8088 ± 0.0487
Polyman	0.8415 ± 0.0818

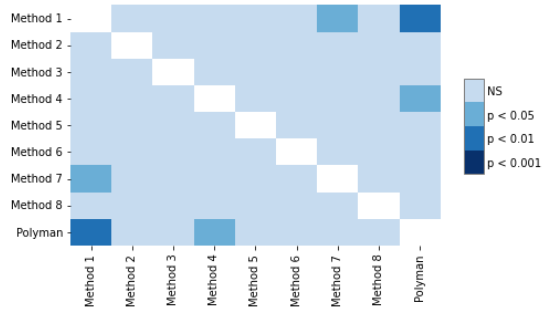

Figure 1: Multiple comparison test on HMCP

5.2 MrOS and WSC

Validation was carried out for the MrOS and WSC data sets, to evaluate the generalization capacity of the automatic annotation methods on data sets independent of the training one.

In this case, we only have the annotations of one clinical expert per EMG signal, therefore we measure the agreement between the methods and the expert (instead of a consensus). Tables 3 and 4 contain the corresponding Kappa means and standard deviations for MrOS and WSC respectively. Additionally, a multiple comparison test was carried out (see Figures 2 and 3).

Table 3: MrOS validation results

Model	Cohen Kappa
Method 1	0.4470 ± 0.1866
Method 2	0.2937 ± 0.1589
Method 3	0.2151 ± 0.1606
Method 4	0.2614 ± 0.1692
Method 5	0.2036 ± 0.1576
Method 6	0.4291 ± 0.1894
Method 7	0.3850 ± 0.1652
Method 8	0.2574 ± 0.1542
Polyman	0.1419 ± 0.2337

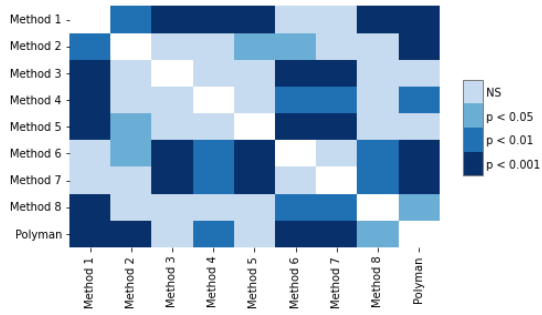

Figure 2: Multiple comparison test on MrOS

Table 4: WSC validation results

Model	Cohen Kappa
Method 1	0.7604 ± 0.0943
Method 2	0.7583 ± 0.0987
Method 3	0.7509 ± 0.1084
Method 4	0.7234 ± 0.1117
Method 5	0.7689 ± 0.0929
Method 6	0.7459 ± 0.1106
Method 7	0.7732 ± 0.0943
Method 8	0.7546 ± 0.0960
Polyman	0.4437 ± 0.3170

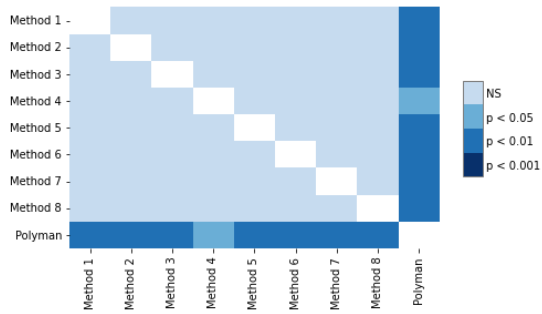

Figure 3: Multiple comparison test on WSC

6 Discussion

Firstly, analyzing the results of the HMCP data set (see Section 5.1), clinical experts obtained an average Kappa agreement of 0.7867 ± 0.0920, while six of our models (Method 2, Method 3, Method 4, Method 5, Method 7 and Method 8) and the Polyman method exceed this value and were more stable in deviation (see Table 2).

On the other hand, the Polyman method obtained the highest Kappa on average over HMCP, although our models have greater stability. However, in the multiple comparison test (see Figure 1), the Polyman method was only significantly different from the Method 1 and Method 4 models. Among our models, Method 7 obtained the highest Kappa and the lower deviation, although the multiple comparison test indicates that it is only significantly different from the Method 1 model, therefore, among the rest of the models, the computationally less expensive method (with fewer training parameters), which are Methods 5, 6 or 7, are preferred.

Analyzing Table 3 we can observe that in the MrOS set, that contains signals with worse quality than the training data (signals were sampled only at 64 Hz), all our models obtain better performance and are more stable than the Polyman method. The multiple comparison test (see Figure 2) indicates that the difference in performance between the Polyman method and our models is significant except for Method 3 and Method 5 models. Among our models, the one that obtained the highest Kappa was Method 1, which turned out to be significantly equivalent in performance to the Method 6 and Method 7 models, therefore, again, the computationally less expensive models are preferred, Methods 6 or 7.

Analyzing Table 4 we can observe that in the WSC set, that contains signals with comparable quality to the training data, all our models obtain better performance and are more stable than the Polyman method. Furthermore, the multiple comparison test (see Figure 3) indicates that the difference in performance between the Polyman method and each of our models is significant. Finally, the performance differences between our models are not statistically significant, so again, the computationally less expensive models are preferred, Methods 5, 6 or 7.

7 Conclusions

We implemented a process for automatic annotation of involuntary motor movements in the limbs, on EMG signals recorded in a PSG. The method includes the processes of preparing the EMG signals for prediction and the subsequent annotation of the detected events.

For the automatic detection of LM, the applicability and suitability of different deep learning architectures was investigated. Eight different deep learning models were trained: three with an LSTM architecture, one with a BiLSTM architecture, three 1D CNN-BiLSTM, and one LSTM-BiLSTM.

The models obtained solid performance, better or comparable to the performance of HMCP clinical experts, and above all, they were more stable, which is highly desirable to eliminate or minimize subjectivity and possible human errors.

The performance of our deep learning models was compared to the automatic annotation method of Álvarez Estévez (2016), based on conventional signal processing methods. The results indicated that methods based on deep learning are more robust and stable facing with signals from new data sources, with different qualities, configurations, recording processes, and in which clinical experts may have slight discrepancies in the annotation criteria.

Finally, models with more complex architectures, and therefore with higher computational cost, did not present significant advantages in performance compared to simpler and more efficient models.

The implemented process could be used to largely reduce the costly manual work of LM annotation, and thus, enable the diagnosis and treatment of many people that are not diagnosed due to limited human resources.

Acknowledgments

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

S. Ancoli-israel, D. F. Kripke, W. Mason, and O. J. Kaplan. Sleep Apnea and Periodic Movements in an Aging Sample1. *Journal of Gerontology*, 40(4):419–425, 07 1985.

T. L. Baker, C. Guilleminault, G. Nino-Murcia, and W. C. Dement. Comparative Polysomnographic Study of Narcolepsy and Idiopathic Central Nervous System Hypersomnia. *Sleep*, 9 (1):232–242, 03 1986.

J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, L. Harper, J. P. Johnson, E. Mitson, and R. R. Delay. Overview of recruitment for the osteoporotic fractures in men study (mros). *Contemporary Clinical Trials*, 26(5):557–568, 2005. URL *https://www.sciencedirect.com/science/article/pii/S1551714405001060*.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

M. Fantini, M. Michaud, N. Gosselin, G. Lavigne, and J. Montplaisir. Periodic leg movements in rem sleep behavior disorder and related autonomic and eeg activation. *Neurology*, 59(12): 1889–1894, 2002. URL *https://n.neurology.org/content/59/12/1889*.

R. Ferri, S. Fulda, R. Allen, M. Zucconi, O. Bruni, S. Chokroverty, L. Ferini-Strambi, B. Frauscher, D. Garcia-Borreguero, M. Hirshkowitz, B. Högl, Y. Inoue, A. Jahangir, M. Manconi, C. Marcus, D. Picchietti, G. Plazzi, J. Winkelman, and R. Zak. 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 groups (irlssg and eurlssg). *Sleep Medicine*, 26:86–95, 2016. URL *https://www.sciencedirect.com/science/article/pii/S1389945716302398*.

D. B. Rye and L. M. Trotti. Restless legs syndrome and periodic leg movements of sleep. *Neurologic Clinics*, 30(4):1137–1166, 2012. URL *https://www.sciencedirect.com/science/article/pii/S073386191200045X*. Sleep Disorders.

R. Vetrugno, F. Provini, P. Cortelli, G. Plazzi, E. M. Lotti, G. Pierangeli, C. Canali, and P. Montagna. Sleep disorders in multiple system atrophy: a correlative video-polysomnographic study. *Sleep Medicine*, 5(1):21–30, 2004. URL *https://www.sciencedirect.com/science/article/pii/S1389945703001953*.

T. Wetter, V. Collado-Seidel, T. Pollmacher, A. Yassouridis, and C. Trenkwalder. Sleep and periodic leg movement patterns in drug-free patients with parkinson's disease and multiple system atrophy. *Sleep*, 23:361–367, 05 2000.

T. Young, L. Finn, P. Peppard, M. Szklo-Coxe, D. Austin, F. Nieto, R. Stubbs, and K. Hla. Sleep disordered breathing and mortality: Eighteen-year follow-up of the wisconsin sleep cohort. *Sleep*, 31:1071–8, 09 2008. URL *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2542952/*.

D. Álvarez Estévez. A new automatic method for the detection of limb movements and the analysis of their periodicity. *Biomedical Signal Processing and Control*, 26:117–125, 2016. URL *https://www.sciencedirect.com/science/article/pii/S1746809416300015*.

D. Álvarez Estévez and R. M. Rijsman. Computer-assisted analysis of polysomnographic recordings improves inter-scorer associated agreement and scoring times. *PLOS ONE*, 17 (9):1–19, 09 2022.