34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Performance Evaluation of an Artificial Neural Network Automatic Spindle Detection System

Errikos M. Ventouras, Nicholas-Tiberio Economou, Ilia Kritikou, Hara Tsekou, Thomas J. Paparrigopoulos, and Periklis Y.Ktonas, *Senior Member, IEEE*

Abstract— **Sleep spindles are transient waveforms found in the electroencephalogram (EEG) of non-rapid eye movement (NREM) sleep. Sleep spindles are used for the classification of sleep stages and have been studied in the context of various psychiatric and neurological disorders, such as Alzheimer's disease (AD) and the so-called Mild Cognitive Impairment (MCI), which is considered to be a transitional stage between normal aging and dementia. The visual processing of whole-night sleep EEG recordings is tedious. Therefore, various techniques have been proposed for automatically detecting sleep spindles. In the present work an automatic sleep spindle detection system, that has been previously proposed, using a Multi-Layer Perceptron (MLP) Artificial Neural Network (ANN), is evaluated in detecting spindles of both healthy controls, as well as MCI and AD patients. An investigation is carried also concerning the visual detection process, taking into consideration the feedback information provided by the automatic detection system. Results indicate that the sensitivity of the detector was 81.4%, 62.2%, and 83.3% and the false positive rate was 34%, 11.5%, and 33.3%, for the control, MCI, and AD groups, respectively. The visual detection process had a sensitivity rate ranging from 46.5% to 60% and a false positive rate ranging from 4.8% to 19.2%.**

I. Introduction

Transient waveforms are present in the sleep electroencephalogram (EEG). Sleep spindles are one of the most distinctive transient waveforms found in the EEG of non-rapid eye movement (NREM) sleep. They are waxing and waning oscillations, of usually 1-2 sec duration, present predominantly in stage 2 of NREM sleep, but also persisting through slow wave sleep, i.e., stages 3 and 4 of NREM sleep. The waveform frequency ranges from around 11 to 16 Hz and its amplitude is mostly below 50 μV peak-to-peak in an adult [1,2]. In recent years there has been progress in elucidating the mechanisms generating spindles, although their functional significance is an ongoing research topic [3,4]. Spindles are used in the classification of sleep stages, since their presence constitutes one of the defining characteristics of stage 2 sleep [5]. Among other topics, sleep spindles have been investigated in the context of dementia, especially Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI), which is considered to be a transitional stage between normal aging and dementia, usually AD [6,7]. Sleep spindles are poorly formed, of lower amplitude, shorter duration and much less numerous in AD than in normal aging [8]. Concurrently, studies indicate possible sleep spindle involvement in cognition and learning [9].

The visual recognition and counting of spindles is a laborious and time-consuming task for whole-night sleep EEG recordings. Automatic spindle detection might help, but automatic detection is seriously hampered by the fact that spindles might be of low amplitude (a few microvolts), are often superimposed on much stronger slow-wave activity and co-exist with diffuse spindle-like rhythmic activity. These problems are compounded by the loose definition of spindles, the high variability in spindle characteristics between subjects and the lack of a reliable "gold" standard, apart from visual inspection, for benchmarking the performance of the proposed systems [2,10,11]. Nevertheless, a rich variety of techniques have been proposed [11]. The techniques applied comprise frequency and amplitude analysis [12-14], fuzzy detectors [11,13], Support-Vector Machine (SVM) classifiers [15], Matching Pursuit (MP) and wavelet techniques [16-18], as well as Artificial Neural Networks (ANN) [10,15,19].

In our previous work [10], the time-domain representation of band-pass (10.5-16 Hz) filtered EEG of a healthy adult subject was used as input to a feed-forward Multi-Layer Perceptron (MLP) ANN, without feature extraction other than that produced by the band-pass filtering. The ANN processed through its input layer successive 0.5 s-long windows of filtered EEG, down-sampled appropriately to 128 samples/sec. Each voltage sample in the input window was assigned to one input-layer neuron. The MLP was 3-layered, with 64, 30 and 2 neurons in the input, hidden and output layer, respectively. By training the network on characteristic examples of filtered EEG segments, the MLP provided acceptable classification results, bypassing the feature selection stage. In the present work, the MLP-ANN is evaluated in detecting spindles of both healthy controls as well as MCI and AD patients. An investigation is carried also concerning the visual detection process, taking into consideration the feedback information provided by the automatic detection system.

E. M. Ventouras is with the Medical Instrumentation Technology Department, Technological Educational Institute of Athens, Athens, 12210, Greece.

N. T. Economou, H. Tsekou, T. J. Paparrigopoulos, and P. Y. Ktonas are with the Sleep Research Unit, Psychiatry Clinic, Eginition Hospital, University of Athens, Athens, 11528, Greece.

I. Kritikou is with the Dept. of Psychiatry and Behavioral Sciences, Medical School, University of Crete, Heraklion, 71003, Greece and the Hershey College of Medicine, Penn State University, Herhsey, PA, 17033, USA (corresponding author, phone: ++1-717-395-2706; ikritikou@hmc.psu.edu).

II. Materials and Methods

The subject recordings in the present study were provided by the Sleep Research Unit of the Eginition Hospital in Athens. The polysomnograms were recorded with the Micromed/BrainQuick system.

The first set of recordings (Set1) consisted of 3 whole-night recordings from 3 healthy male control subjects (aged 26 years). The night sleep record of each subject was divided into three consecutive parts (thirds of the night). In each part, the longest stage 2 sleep period was selected, provided that it lasted for at least 10 minutes. Then a time frame, starting 5 min after the start of stage 2 and lasting 5 min and 5 sec, was selected for visual analysis. Cz recordings from one subject (sampling rate 256 Hz) and C3 recordings (sampling rate 512 Hz) from the other two subjects were scored by a polysomnographer, based on visual detection criteria described in [10]. After visual detection, the sampled EEG signal was band-pass filtered, using a 128-coefficient FIR filter, with 3dB cutoff frequencies set at 10.5 and 16 Hz. The MLP was trained using the methodology described in [10]. The set of biases and weights that was produced by the training is denoted TRA. It should be pointed that Set1 was used only for providing TRA. In the present study, no performance evaluation of the automated spindle detection system was carried on Set1.

The second set of recordings (Set2) was based on whole-night recordings from 4 healthy controls (3 male and 1 female, average age=67 years, S.D.=6.98 years), 7 patients suffering from MCI (7 females, average age=75.7 years, S.D.=5.8 years) and 2 patients suffering from AD (1 male and 1 female, aged 73 and 76 years, respectively). Recordings were performed with 512 Hz sampling rate. The recordings were scored in consensus by two human scorers, other from the scorer who scored Set1, according to the following criteria. "Bona fide" spindles were those that fitted unequivocally the criteria of [5], i.e., they presented a fusiform activity with frequency at 12 to 14 Hz with a duration of at least 0.5 sec. The number of those spindles was denoted as V_{BFS}. "Marginal" spindles where those that, although they kept their fusiform shape, had either duration from 0.4 to 0.5 sec or presented frequencies from 10 to 12 Hz or 12 to 15 Hz. Additionally, for "marginal" spindles to be accepted by the scorers, they had to be present at central or parietal electrodes. The number of those spindles was denoted as V_{MAR}. The relaxation of the visual detection criteria was deemed essential, since the study was designed to include recordings from patient populations, whose spindles are expected to be somewhat, or even severely, distorted. Suppose that visual detection indicated, in consensus scoring by two scorers, V spindles. The total number (V) of spindles detected in consensus by the scorers was the sum of the "bona fide" spindles, V_{BFS}, and the spindles that fitted the visual detection criteria marginally, V_{MAR}:

$$V=V_{BFS}+V_{MAR}. \quad (1)$$

After visual detection, the sampled EEG signal was band-pass filtered, as for Set1. The MLP was trained as described in [10]. The set of biases and weights that was produced by the training is denoted TRB.

Performance evaluation of the MLP was accomplished using the output value O(t), $0 \le O(t) \le 1$, t corresponding to the time samples of the visually scored EEG recordings. The O(t) curve was divided into parts that had value greater or lower than a threshold value V_T. The parts that had a value greater than V_T were denoted as "peaks". We used two criteria for checking whether a spindle presence was indicated by the ANN output. According to the "soft" criterion (SC), the ANN provided a spindle indication (SI) when a peak existed in the O(t) curve. According to the "hard" criterion (HC), a spindle presence was indicated only when the peak duration was greater than P_D sec. The SIs were automatically computed by the system [20-22].

Next, SIs were divided into 4 categories. Category 1 included the SIs which matched visually detected spindles. Their number was denoted by N_V. Category 2 included SIs in EEG segments where no spindles were detected by the human scorers. Nevertheless, on a second inspection based on the SIs, those EEG segments unequivocally fitted the visual detection criteria, i.e., corresponded to bona fide spindles that were missed by the scorers. Their number was denoted by N_{BFS}. Category 3 included SIs in EEG segments where no spindles were detected by the human scorers and where the EEG morphology also unequivocally did not fit the visual detection criteria for spindles. These SIs (whose number was denoted by N_{sFP}) could be denoted as "serious" false positives of the MLP. Finally, category 4 included SIs in EEG segments where no spindles were detected by the human scorers and at the same time the EEG morphology fitted only marginally the visual detection criteria for spindles. Their number was denoted by N_{MAR}. If N was the total number of SIs then:

$$N= N_V+N_{BFS}+ N_{sFP}+N_{MAR}. \quad (2)$$

In the present study, 3 sleep EEG segments, belonging to stage 2 of sleep, were used for implementing the performance evaluation of the MLP. Each segment belonged to one subject from the 3 subject groups used in Set2 and was not previously used in the training process. 2 MLP outputs ("runs") were computed, for each segment, one using the weights and biases computed by TRA and the second using the weights and biases computed by TRB. For each run, the respective quantities of SIs were denoted by N_X,$N_{V/X}$,$N_{BFS/X}$, $N_{sFP/X}$ and $N_{MAR/X}$, x=A or B, indicating the respective "run".

It is well known that high levels of inter-rater and intra-rater variability might be present in the scoring process [10]. A crucial quantity used for computing the performance of an automatic spindle detection system is the quantity of spindles that are considered to be present in each of the EEG segments used for the performance evaluation. This quantity (that will be denoted as E in the following) is affected by the scoring variability. On the other hand, the MLP provides SIs and their careful examination, as exposed in the categorization of SIs into 4 categories (see above), might provide some help in alleviating the problems related to visual scoring. For example, the information present in the SIs helped in detecting segments that included bona-fide spindles which

were nevertheless missed by the scorers (i.e., category 2 SIs). In light of the above considerations, in the present study, E was not taken as equal to the number of spindles initially detected by the human scorers (i.e., $V=V_{BFS}+V_{MAR}$), but was computed as described in the following. Firstly, in E were included the spindles that were visually detected by the scorers and fitted unequivocally the visual detection criteria, V_{BFS}. Secondly, from the spindles that were visually detected by the scorers and fitted the visual detection criteria marginally, V_{MAR}, only those that were also detected by the MLP in at least one of the 2 runs (whose number was denoted as $V_{MAR(1)}$) were included in E. The rest of the spindles that were visually detected by the scorers and fitted the visual detection criteria marginally (whose number was denoted as $V_{MAR(2)}=V_{MAR}-V_{MAR(1)}$) were not included in E, but were classified as visual false positives (VFP) in the performance evaluation of the visual detection process (see below). Thirdly, in E were also included those EEG segments that corresponded to category 2 SIs in at least one of the 2 runs. We denote the number of such EEG segments as $N_{BFS}(A \cup B)$. Finally, in E were also included those EEG segments that corresponded to category 4 SIs in at least one of the 2 runs. We denote the number of such EEG segments as $N_{MAR}(A \cup B)$.

According to the above categorization, the sensitivity of the MLP for run x (x=A or B) was computed as:

$$S_X=H_X/E*100 \quad (3)$$

where

$$H_X = N_{V/X}+N_{BFS/X}+N_{MAR/X} \quad (4)$$

$$E= V_{BFS}+V_{MAR(1)}+N_{BFS}(A \cup B)+N_{MAR}(A \cup B) \quad (5)$$

The false positive rate of the MLP for run x (x=A or B) was computed as

$$FP_X = N_{sFP/X}/N_X*100 \quad (6)$$

It was stated previously that E might provide a better indication, as compared to V, about the quantity of spindles that are present in each of the EEG segments used for the performance evaluation. Following the rationale used for computing E, in the present study, in addition to the performance evaluation of the MLP as expressed by S_X and FP_X, an evaluation was attempted for the visual detection process. The sensitivity of the visual detection process can be considered as:

$$S_{VIS}=(V_{BFS}+V_{MAR(1)})/E*100 \quad (7)$$

The false positive rate of the visual detection process is quantified as:

$$FP_{VIS}=VFP/V=V_{MAR(2)}/(V_{BFS}+V_{MAR})=$$

$$V_{MAR(2)}/(V_{BFS}+V_{MAR(1)}+V_{MAR(2)})*100, \quad (8)$$

where the visual false positives (VFP) are those spindles that were visually detected, fitted the visual detection criteria marginally and were not detected by the MLP in any of the 2 runs.

It is worth noting that the above procedure, both for the MLP and the visual detection process performance evaluation, can be, in principle, extended to include the information from more than 2 "runs".

III. Results

For the training of the MLP using Set1, 18 spindle segments, containing only spindle activity, and 18 segments free of spindle activity were used. The spindle segments, in order to be selected, had to correspond to "bona fide" spindles, with spindle amplitude of at least 10μV. The segments were selected proportionally from the 3 subjects of Set1 and the 3 parts of the night.

For the training of MLP using Set2, 6, 11 and 6 spindle segments were used, containing only spindle activity, from the healthy, MCI and AD group recordings, respectively, as well as 5, 10 and 2 segments, free of spindle activity, from the healthy, MCI and AD groups, respectively. Spindle segments were selected for training with the same criteria as those for Set1. Various electrode positions were used in the analysis, such as F4, C4, P3, P4, O1 and O2. This was so in order to make the training of the MLP classifier independent of a specific location. This was expected to provide classification results that would have been more representative of what would happen when the system is used, after learning, for detecting spindles at various electrode positions.

The visual detection process for Set2 indicated 21 spindles for the control subject segments, 26 spindles for the MCI subject segments and 38 spindles for the AD subject segments. Two "runs" were performed on segments from Set2, one for TRA ("run A"), with V_T=0.5 and P_D=0.3 sec, and another for TRB ("run B"), with V_T=0.9 and P_D=0.5 sec. The values of the thresholds were selected after preliminary sets of "testing runs" were performed, for weights and biases values as produced by TRA and TRB. Threshold optimization was implemented only for the purpose of the present performance evaluation. In future use of the system, for AD and MCI patients, thresholds might be permanently fixed to the values selected in the present work for TRB. Measure E had the values 43, 37 and 60 for the control subject, the MCI patient and the AD patient, respectively. Performance evaluation results for the MLP are given in Table I, for runs A and B. Performance evaluation results of the visual spindle detection process are given in Table II.

IV. Discussion

The performance evaluation results presented in Table I provide a clear distinction between the results for runs A and B, concerning sensitivity values. Sensitivity is much higher for run B, for each of the subject groups. In run A the MLP used was trained on data from control subjects only, different from the control subjects used in testing of the MLP. In run B the MLP used was trained on data from both control and patient subjects. Therefore, it might be conjectured that the improvement of the performance of the MLP in run B is due to the fact that it has been trained on both patient and on control data, and to the fact that the training data were from

TABLE I. PERFORMANCE EVALUATION FOR MLP FOR RUNS A AND B

	Run A						
Subjects	$N_{V/A}$	$N_{BFS/A}$	$N_{sFP/A}$	$N_{MAR/A}$	H_A	S_A	FP_A
Controls	17	4	5	6	27	62.8	15.6
MCI	8	2	3	2	12	32.4	20.0
AD	26	9	11	4	39	65.0	22.0
	Run B						
Subjects	$N_{V/B}$	$N_{BFS/B}$	$N_{sFP/B}$	$N_{MAR/B}$	H_B	S_B	FP_B
Controls	18	11	18	6	35	81.4	34.0
MCI	10	7	3	6	23	62.2	11.5
AD	33	12	25	5	50	83.3	33.3

TABLE II. PERFORMANCE EVALUATION OF THE VISUAL SPINDLE DETECTION PROCESS

Subjects	V	V_{BFS}	$V_{MAR(1)}$	$V_{MAR(2)}$	S_{VIS}	FP_{VIS}
Controls	21	17	3	1	46,5	4,8
MCI	26	18	3	5	56,8	19,2
AD	38	29	7	2	60,0	5,3

the same subjects as those used for testing the MLP. On the other hand, although 10 and 2 of the 17 spindles used for training the MLP used in run B belonged to MCI and AD patients, respectively, sensitivity was much lower for the MCI group than for the AD and the control subjects group. Concerning the false positive rate, results for run B were improved compared to run A only for the MCI patients.

As can be seen from Table II, the visual detection process presents low sensitivity, in the range of 45% to 60%. This indicates that a lot of spindles that should have been detected by the scorers in their "1st pass" have been missed. This has serious repercussions, because the performance evaluation of an automatic detection system is absolutely dependent on the markings given by the human scorers, for providing a benchmark for computing its performance. The existence of a high number of spindles, indicated by the system but missed by the scorers, will lead to artificially high false positive rates. Therefore, the implementation of "feedback" processes, concerning the visually detected spindles, as the one used in the present study, might be beneficial for the reliability of the performance evaluation of automatic systems. Another aspect that emerges from the results of Table II is that the sensitivity of the visual detection process is not worsening for the patient classes, as could have been expected due to the relative deterioration of the spindle morphology in AD patients [6-8].

REFERENCES

[1] W. R. Jankel, and E. Niedermayer, "Sleep spindles," *J. Clin. Neurophysiol.,* vol. 2, pp. 1-35, 1985.

[2] L. Gennaro, L., and M. Ferrara, "Sleep spindles: an overview," *Sleep Med. Rev.,* vol. 7, no. 5, pp. 423-440, 2003.

[3] M. Steriade, and F. Amzica, "Coalescence of sleep rhythms and their chronology in corticothalamic networks," *Sleep Res. Online*, vol. 1, pp. 1-10,1998.

[4] G. Buzsaki, *Rhythms of the brain.* Oxford, UK: Oxford University Press, 2006.

[5] A. Rechtschaffen, and A. Kales, eds., *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects.* Washington, DC: Public Health Service, U.S. Government Printing Office, 1968.

[6] P. Y. Ktonas, S. Golemati, P. Xanthopoulos, V. Sakkalis, M. D. Ortigueira, H. Tsekou, et al., "Time-frequency analysis methods to quantify the time-varying microstructure of sleep EEG spindles: Possibility for dementia biomarkers?," *J. Neurosci. Meth.*, vol. 185, pp. 133-142, 2009.

[7] N. T. Economou, A. Kyrozis, I. Kritikou, P. Ktonas, A. Bonakis, D. Dikaios, et al., "Inter-hemispheric spectral coherence reduction in sleep spindle frequency activity in patients with cognitive decline associated with amnesic mild cognitive impairment and Alzheimer's disease," *J. Sleep Res.*, vol. 19 (Suppl.1), pp.1050, 2010.

[8] D. Petit, J.-F. Gagnon, M. L. Fantini, L. Ferini-Strambi, J. Montplaisir, "Sleep and quantitative EEG in neurodegenerative disorders," *J. Psychosom. Res.*, vol. 56, pp. 487-496, 2004.

[9] S. M. Fogel, R. Nader, K. A. Cote, C. T. Smith, "Sleep spindles and learning potential", *Behav. Neurosci.*, vol. 121, pp. 1-10, 2007.

[10] E. Ventouras, E. Monoyiou, P. Ktonas, T. Paparrigopoulos, D. Dikeos, N. Uzunoglu, C. Soldatos, "Sleep Spindle Detection Using Artificial Neural Networks Trained with Filtered Time-Domain EEG: A Feasibility Study," *Comput. Meth. Prog. Bio.,* vol. 78, pp. 191-207, 2005.

[11] L. Causa, C. Held, J. Causa, P. Estévez, C. Perez, R. Chamorro, M. Garrido, C. Algarín, P. Peirano, "Automated sleep-spindle detection in healthy children polysomnograms," *IEEE Trans. Biomed. Eng.,* vol. 57, pp. 2135-2146, 2010.

[12] P. Schimicek, J. Zeitlhofer, P. Anderer, B. Saletu, "Automatic sleep spindle detection procedure: aspects of reliability and validity," *Clin. Electroencephal.*, vol. 25, no. 1, pp. 26-29, 1994.

[13] E. Huupponen, G. Gomez-Herrero, A. Saastamoinen, A. Varri, J. Hasan, S.-L. Himanen, "Development and comparison of four sleep spindle detection methods," *Artif. Intell. Med.,* vol. 40, pp. 157-170, 2007.

[14] R. Bódizs, J. Körmendi, P. Rigó, A. Sándor Lázár, "The individual adjustment method of sleep spindle analysis: methodological improvements and roots in the fingerprint paradigm," *J. Neurosci. Meth.*, vol. 178, no. 1, pp. 205-213, 2009.

[15] N. Acır, and C. Gózelis, "Automatic recognition of sleep spindles in EEG by using artificial neural networks," *Expert Syst. Appl.*, vol. 27, no. 3, pp. 451–458, 2004.

[16] J. Zygierewicz, K. J. Blinowska, P. J. Durka, W. Szelenberger, S. Niemcewicz, W. Androsiuk, "High resolution study of sleep spindles," *Clin. Neurophysiol.*, vol. 110, pp. 2136-2147, 1999.

[17] S. V. Schonwald, E. L. de Santa-Helena, R. Rossatto, M. L. F. Chaves, G. J. L. Gerhardt, "Benchmarking matching pursuit to find sleep spindles," *J. Neurosci. Meth.*, vol. 156, no. 1–2, pp. 314–321, 2006.

[18] F. Duman, A. Erdamar, O. Erogul, Z. Telatar, S. Yetkin, "Efficient sleep spindle detection algorithm with decision tree," *Expert Syst. Appl.*, vol. 36, pp. 9980-9985, 2009.

[19] Gunes, S., Dursun, M., Polat, K., Yosunkaya, S., "Sleep spindles recognition system based on time and frequency domain features," *Expert Systems with Applications*, 38, 2455–2461, 2011.

[20] G. Katharakis, "Automatic investigation of sleep spindles with artificial neural networks and clinical applications," M.S. thesis, Dept. of Nursing, Athens Univ., Athens, Greece, 2004.

[21] G. Strakosha, "Automatic detection of sleep spindles with artificial neural networks: Optimization of detection thresholds," M.S. thesis, Dept. of Nursing, Athens Univ., Athens, Greece, 2008.

[22] I. Billas, "Development of an algorithm for the automatic evaluation of an electroencephalographic sleep spindle detection system," B.S. thesis, Dept. of Medical Instrumentation Technology, Athens Technol. Educ. Inst., Athens, Greece, 2009.