Sleep Spindle Detection Using Deep Learning: a Validation Study Based on Crowdsourcing

Dakun Tan, Rui Zhao, Jinbo Sun and Wei Qin

Abstract— **Sleep spindles are significant transient oscillations observed on the electroencephalogram (EEG) in stage 2 of non-rapid eye movement sleep. Deep belief network (DBN) gaining great successes in images and speech is still a novel method to develop sleep spindle detection system. In this paper, crowdsourcing replacing gold standard was applied to generate three different labeled samples and constructed three classes of datasets with a combination of these samples. An F1-score measure was estimated to compare the performance of DBN to other three classifiers on classifying these samples, with the DBN obtaining an result of 92.78%. Then a comparison of two feature extraction methods based on power spectrum density was made on same dataset using DBN. In addition, the DBN trained in dataset was applied to detect sleep spindle from raw EEG recordings and performed a comparable capacity to expert group consensus.**

I. INTRODUCTION

The sleep spindle first described by Berger [1] and introduced by Loomis et al. [2] is a group of waveforms superimposed on low voltage background electroencephalogram (EEG), which has the characteristic of progressively increasing, then gradually decreasing, amplitude akin to the shape of spindle. As a brief distinct transient waveform, it mostly presents in stage 2 of non-rapid eye movement (NREM) sleep and becomes a important criterion for sleep stage scoring cooperating with muscle activity and eye movement [3]. Although there are no consensus on the frequency and the time duration of sleep spindle, all experts use the two features to identify sleep spindle by visual inspection and mark the appearance of spindle as one hallmark of human stage 2 sleep.

Compared to other types of EEG, sleep spindle received more favors from neurophysiologists than K-complex and vertex sharp wave for the reason that the neurophysiological mechanism of spindle generation is not always described clearly and repeatedly enough for the researchers to recognize and understand them [4]. On one hand, the relationship between variables at a neural level and sleep spindle density of EEG make a feasible and reliable method to investigate the disorders such as epilepsy [5], schizophrenia [6] and neurodegeneration problems [7]. on the other hand, affected by many factors, such as sleep deprivation [8], [9], circadian [10], [11] and ageing [12], [13], sleep spindles are heritable [14] and make an significant contribution to synaptic plasticity and memory consolidation [6], [15].

D. Tan and R. Zhao are students at the center of sleep and neural image, the school of life science and technology, Xidian University, shannxi 710126, China. tdkcs@hotmail.com

J. Sun and W. Qin are with the school of life science and technology, Xidian University, shannxi 710126, China.

Automatic spindle detection system has been a mature and active field of sleep [16], [17], [18]since schimicek et al. implemented the first automated spindle detector [19], but neurophysiologists still have an affinity to visual identification by hand which is a laborious and time consuming task. Strategies of many automatic detection algorithms can only rely on simple threshold of the time or frequency of EEG signal for developed to mimic and speed up the visual identification. Besides the fact that it is not sufficient to detect sleep spindle with those two features, an evidence shows that the least duration time of sleep spindle defined by America Academy of Sleep Medicine is meeting the challenge [20]. Further, the estimation of performance on spindle detection algorithm is also not strict due to the gold-standard dataset which is not rigorous enough in two aspects, one is that the dataset are derived from small subjects whose ages even span large years and the other is that gold-standard principle is different among experts. Therefore, many advanced spindle detection algorithms intentionally or unintentionally neglect this point since the process of collecting true 'gold' standard dataset is too difficult and complicated to execute[20].

In this paper, we use a crowdsourcing method to generate dataset, which alleviate the dilemma of gold-standard principle and is feasible and reliable. The power spectrum density (PSD) of 2 s EEG signal is calculated as input to different classifiers. Firstly, four features of PSD are selected to compare the performance of deep belief network (DBN) on sleep spindle classification to other traditional classifiers with low dimension feature. Secondly we use the raw PSD which have no other preprocess except PSD transformation to exploit the ability of DBN to separate spindle and non-spindle. Finally, under the best set of DBN generalization, we investigate the possibility of applying the DBN trained in the dataset to raw EEG activity.

II. MATERIALS AND METHODS

A. Collecting EEG

EEG signals were derived from thirty young subjects (age 20-23) slept in the sleep laboratory. All subjects signed informed consents and were free of drug and sleep complaint, and the experimental procedure involving these human subjects were approved by Xidian University's Ethical Review Board. The EEG signals were recorded on electrodes according to the international 10-20 system sampled at 500Hz using Brain Products GmbH. Before scoring the sleep stages, EEG signals were low pass filtered at 50Hz. Sleep stages were scored based on the scoring system of the ASSM manual edited by Rechtschaffen and Kales [3], although

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

this guideline had small flaws on some respects [21]. 10 minutes EEG recording from the central channel (C3-A2) was selected from each subject. All the 10 minutes EEG signals were extracted from NREM 2 stage for the reason that the sleep spindle was most frequency occurred and the absence of delta waves during this stage.

B. Crowdsourcing

Inspired by the method first introduced by Simon C Warby st al. [20], we recruited 100 participates defined as non-experts who were not familiar with sleep spindle before visually identifying spindle from these thirty segment signals. Each non-expert was trained on how to use our visual-score spindle system and our guideline manual which introduced the difference between spindle and alike-spindle. On passing through the prepared test designed by some experts, each one was arranged with a task of visually identifying spindle on six segments selected from those thirty segments by a pseudo-random method. In other words, each of the thirty segments was scored exactly by twenty non-experts. Each non-expert had two choices about alike-spindle, possible (marked as 0.5) and true (marked as 1), and there was no limit on the time duration of sleep spindle. Average value of twenty non-experts' results was used to identify the spindle in segments.

C. Datasets

A section of EEG signal was defined as spindle when there were more than 50 percent non-experts scored it as spindle according to the principles of majority rule. Similarity, alike-spindles that less than 20 percent non-experts scored were defined as non-spindle-B; and the third class of alike-spindle was non-spindle-A which no non-experts consented (Table I). All the three different alike-spindles were extracted as 2 s EEG samples by expanding the center point of the sections to left and right 1 second respectively, but there were a limit on non-spindle-A samples that the duration time of non-spindle-A lasted at least 2 seconds. Since there were two different non-spindles and the performance of classifier were affected by the training samples, those three classes of alike-spindles made up the non-experts datasets by the different combination of those alike-spindles with identical quantity (Table II).

TABLE I

THE LABELED SAMPLE AGAINST NON-EXPERTS GROUP CONSENSUS

non-experts group consensus	labeled	number of samples
$T_{negc} = 0$	non-spindle-A	2698
$0 < T_{negc} \leq 20$	non-spindle-B	1338
$20 < T_{negc} \leq 50$	vague-spindle	645
$50 < T_{negc}$	spindle	753

Except for the non-experts group datasets, there were expert group dataset that consisted of ten segments selected from those thirty subjects using the same method as mentioned above with no interaction with or overlay on those thirty segments. All the ten segments then scored by ten experts who were professional to EEG signal and labeled by the threshold of expert group consensus (T_{egc}), which were used to test the generalization capacity and application ability of deep learning.

TABLE II

THE DIFFERENT DATASETS

dataset class	true sample	false sample(amount)
dataset1	spindle	non-spindle-A(753)
dataset2	spindle	non-spindle-B(753)
dataset3	spindle	non-spindle-A(377),non-spindle-B(376)

D. Feature extraction

Data dimension reduction was a general and plausible way in classification tasks to preprocess large dimension raw signal using features chosen by domain knowledge. Power spectrum density has widely used on signal processing taking the advantages of combining two features, frequency and time duration, which were the two distinctive characteristics of sleep spindle. On one hand, power spectrum density was a transformation from time signal that loss less information than feature extraction, on the other hand, the pattern of PSD was an fingerprint of human sleep [22]. Based on the raw PSD, we computed five sum of PSD on five frequency bands (delta-band: 1-4Hz, theta-band: 4-8, alpha-band: 8-13Hz, spindle-band: 10.5-16.1Hz, beta-band: 13-30Hz) and obtained four features through dividing the sum value of PSD on spindle-band by other four bands respectively, with 1024-point periodogram method using rectangle window was applied to calculate an power spectrum density for each 2 s sample. The time duration feature then could be represented by the PSD proportion of different band frequency. The choice of spindle band derived from the research that the oscillation frequency of spindles varies from 10.5 to 16.1 [20], and the other frequency bands meet the standards [3].

E. Deep belief network

As a new brilliant deep architecture, deep belief network was stacked by a defined number of restricted Boltzmann machines (RBMs),where the output from a lower-level RBM was the input to a high-level RBM [23]. Hinton et al. used an efficient greedy layer-wise algorithm to MNIST digital classification with a three-layer (500-500-2000) DBN [24]. We used two classes of DBNs in this study, one two-layer DBN was used on P-PSD with four feature as input, the other three-layer DBN to raw 5-20Hz PSD. Based on the proportion of power spectrum density (P-PSD), we compared DBN, as a classifier, to other three traditional classifiers on F1-score measure,

$$F1 = 2\frac{sensitivity \cdot precision}{sensitivity + precision}. \tag{1}$$

where

$$sensitivity = \frac{TP}{TP + FN}, \tag{2}$$

$$precision = \frac{TP}{TP + FP}. \tag{3}$$

The parameters estimated were TP (True Positive), FN (False Negative), TN (True Negative), FP (False Positive). F1-score measure consiting of both sensitivity and precision was common used on measuring detection performance. We used four different classifiers, Decision Tree, K Nearest Neighbors, Support Vector Machine, and DBN, to classify train samples and evaluated the performance of DBN using raw PSD in three different datasets. Finally, in order to research DBN's generalization capability, a DBN classifier was trained to compare the results to expert group consensus against T_{negc}.

III. RESULTS

Table III shows the amount of sleep spindle identified by non-expert scorers was determined by the 'threshold for non-expert group consensus' (T_{negc}), as descending with the T_{negc}. At $T_{negc} = 0.5$, it identified 753 sleep spindles which were critical positive samples on training classifiers. In a certain case, except the range from 0 to 0.2 where assembled a much more alike-spindles than others, the number of sleep spindle had uniform distribution between 0.2 to 1. At the same time, the frequency of alike-spindle was closely related to T_{negc} with a significant positive linear correlation. In other words, the frequency of EEG signal played an important role when these non-expert scorers annotated sleep spindle, and most of them were confused by the similarity of alpha EEG to sleep spindle.

Table IV shows the average F1-score measures on classification for four different classifiers applied with P-PSD feature in three datasets. It is clearly obvious that the performance of DBN on classification was about 3 percent better than traditional classifiers, DT and KNN. Dataset1 had the best performance of all the classifiers, and dataset2 was the most difficult part to classify for the reason that non-spindle-B labeled when T_{negc} was less than 0.2 was harder than obvious non-spindle-A. All classifiers obtained distinct results in the datasets of different classes, but the classifier of deep belief network performed the best performance of F1-score measure in all datasets, especially in dataset2. Table IV also shows the performance of DBN trained with raw PSD feature on F1-score measures, in three datasets. For the comparison of the DBN trained on P-PSD to raw PSD , the raw PSD absolutely performed better than the other in all the three datasets; and dataset2 was still the most difficult part to the raw PSD trained DBN.

TABLE IV
THE COMPARISON BETWEEN FOUR CLASSIFIERS ON F1-SCORE MEASURE

		P-PSD				PSD
		DT	KNN	SVM	DBN	DBN
dataset1	mean	0.8284	0.8407	0.8525	0.8573	**0.9278**
	std	0.0436	0.0310	0.0323	0.0036	0.0039
dataset2	mean	0.6561	0.6832	0.6783	0.7106	0.8769
	std	0.0527	0.0484	0.0726	0.0043	0.0029
dataset3	mean	0.7444	0.7530	0.7704	0.7909	0.8945
	std	0.0392	0.0447	0.0467	0.0021	0.0036

Fig. 1 shows the ability of two classes of DBNs to detect spindle. The DBN trained in dataset1 was more sensitivity than the one trained in dataset2, and both of them had detected the spindle scored by experts. Table V is the number of scored spindles by expert group consensus (T_{egc}) and the number of spindles detected by DBN trained in dataset2. The performance of DBN was comparable to results from expert group consensus at $T_{egc} = 0.2$.

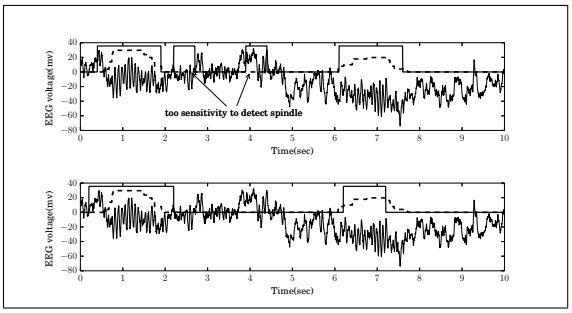

Fig. 1. the application of DBN to detect sleep spindle, two different DBNs trained in two different datasets, dataset1(upper) and dataset2(down)

TABLE V
THE COMPARISON BETWEEN DBN CLASSIFIER AND CROWDSOURCING

		crowdscourcing(T_{egc})					
subject	DBN	0.10	0.15	0.20	0.25	0.30	0.35
1	31	40	33	**29**	29	26	24
2	49	50	**44**	37	37	33	32
3	54	54	**49**	45	40	35	31
4	27	36	32	**27**	20	16	14
5	42	47	**41**	38	34	30	25
6	46	61	58	52	**40**	37	30
7	53	57	53	**52**	47	45	39
8	50	56	53	52	**47**	43	41

IV. DISCUSSION AND CONCLUSIONS

The number of sleep spindle identified through crowdsourcing present a uniform distribution. As the T_{negc} increases from 0.2 to 0.8, the amount of spindle in 0.1 bin tends to stable. A large amount of alike-spindle are concentrated in the bin where T_{negc} below 0.1 for the reason that non-experts are confused by the small gap between sigma band and delta band.

While P-PSD as an extracted feature is feasible in detecting sleep spindle, SVM and DBN classifiers perform better than DT and KNN classifiers using the P-PSD feature in three datasets. Dataset2 consisting of spindle and non-spindle-B is most difficult to classify to all classifiers. Comparing the raw PSD to P-PSD, DBN classifier gain a superiority of more than 10 percent since raw PSD loss less information of raw EEG than P-PSD. In addition, the non-spindle-B play a critical role in our experiment, other automatic methods threw out the non-spindle-B for they missed them, but DBN do that by rejecting it. In the application of DBN, the DBN trained in dataset2 becomes

TABLE III

THE RESULTS OBTAINED BY CROWDSOURCING

	T_{negc}	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
amount of spindle	sum	1742	1366	1106	906	**753**	606	447	290	127
	bin	418	284	212	149	195	167	164	165	108
max frequency(Hz)	sum	12.64	12.83	13.10	13.15	13.27	13.38	13.48	13.60	13.60
	bin	11.99	11.72	12.58	12.24	12.80	12.95	13.31	13.54	13.60

susceptible to non-spindle-B and make a more reasonable judgement on alike-spindle. Further, the results obtained by DBN trained in dataset2 were comparable to expert group consensus. Although the potential capacity of DBN is very exciting in the future, how much the value of T_{ngec} should hold is still a challenge.

The main motivation for investigating the deep belief network performance on detecting sleep spindle in this work is the observations that the deficiency of dataset used by various automatic sleep spindle detection systems and that DBN gain an enormous success on machining learning,especially in images and speech. Taking into consideration the interesting findings presented in this work, we can conclude that deep neural network, a new deep learning architecture, has an ability to detect sleep spindles with comparable performance on F1-score measures and has potential ability to learn the raw inner characteristic of sleep spindle. Nevertheless, crowdsourcing is a new method and the datasets derived from that have a significant effect on using DBN, further research on these matters is needed. In addition, the DBN should be applied to a larger number of subjects and recordings before it may be appropriate for clinical monitoring and data mining task.

REFERENCES

[1] H. Berger, Über das Elektroenzephalogramm des Menschen ICXIV[J]. Arch. Psychiat. Nervenkr.(1929C1938) (87-108).

[2] A. L. Loomis, E. N. Harvey, G Hobart, Potential rhythms of the cerebral cortex during sleep[J]. Science, 1935.

[3] A. Rechtschaffen, A. Kales, A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects[J]. 1968.

[4] L. D. Gennaro, M. Ferrara, Sleep spindles: an overview[J]. Sleep medicine reviews, 2003, 7(5): 423-440.

[5] E. H. Miller, A note on reflector arrays (Periodical style Accepted for publication),IEEE Trans. Antennas Propagat., to be publised.

[6] E. J. Wamsley, M. A. Tucker, A. K. Shinn, et al., Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation[J]. Biological psychiatry, 2012, 71(2): 154-161.

[7] D. Petit, J. F. Gagnon, M. L. Fantini, et al., Sleep and quantitative EEG in neurodegenerative disorders[J]. Journal of psychosomatic research, 2004, 56(5): 487-496.

[8] D. J. Dijk, B. Hayes, C. A. Czeisler, Dynamics of electroencephalographic sleep spindles and slow wave activity in men: effect of sleep deprivation[J]. Brain research, 1993, 626(1): 190-199.

[9] A. A. Borbély, F. Baumann, D. Brandeis, et al., Sleep deprivation: effect on sleep stages and EEG power density in man[J]. Electroencephalography and clinical neurophysiology, 1981, 51(5): 483-493.

[10] D. J. Dijk, C. A. Czeisler, Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans[J]. The Journal of neuroscience, 1995, 15(5): 3526-3538.

[11] D. J. Dijk, T. L. Shanahan, J. F. Duffy, et al., Variation of electroencephalographic activity during non-rapid eye movement and rapid eye movement sleep with phase of circadian melatonin rhythm in humans[J]. The Journal of physiology, 1997, 505(3): 851-858.

[12] A. Nicolas, D. Petit, S. Rompre, et al., Sleep spindle characteristics in healthy subjects of different age groups[J]. Clinical Neurophysiology, 2001, 112(3): 521-527.

[13] K. R. Peters, L. B. Ray, S. Fogel, et al., Age Differences in the Variability and Distribution of Sleep Spindle and Rapid Eye Movement Densities[J]. PloS one, 2014, 9(3): e91047.

[14] U. Ambrosius, S. Lietzenmaier, R. Wehrle, et al., Heritability of sleep electroencephalogram[J]. Biological psychiatry, 2008, 64(4): 344-348.

[15] M. Barakat, J. Doyon, K. Debas, et al., Fast and slow spindle involvement in the consolidation of a new motor sequence[J]. Behavioural brain research, 2011, 217(1): 117-121.

[16] A. Nonclercq, C. Urbain, D. Verheulpen, et al., Sleep spindle detection through amplitude-frequency normal modelling[J]. Journal of neuroscience methods, 2013, 214(2): 192-203.

[17] S. A. Imtiaz, S. Saremi-Yarahmadi, E. Rodriguez-Villegas. Automatic detection of sleep spindles using Teager energy and spectral edge frequency[C]//Biomedical Circuits and Systems Conference (BioCAS), 2013 IEEE. IEEE, 2013: 262-265.

[18] S. Devuyst, T. Dutoit, P. Stenuit, et al., Automatic sleep spindles detection Overview and development of a standard proposal assessment method[C]//Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. IEEE, 2011: 1713-1716.

[19] P. Schimicek, J. Zeitlhofer, P. Anderer, et al., Automatic sleep-spindle detection procedure: aspects of reliability and validity[J]. Clinical EEG and neuroscience, 1994, 25(1): 26-29.

[20] S. C. Warby, S. L. Wendt, P. Welinder, et al., Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods[J]. Nature methods, 2014, 11(4): 385-392.

[21] S. L. Himanen, J. Hasan. Limitations of rechtschaffen and kales[J]. Sleep medicine reviews, 2000, 4(2): 149-167.

[22] L. Gennaro, M. Ferrara M, F. Vecchio, et al., An electroencephalographic fingerprint of human sleep[J]. Neuroimage, 2005, 26(1): 114-122.

[23] Y. Bengio. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning, 2009, 2(1): 1-127.

[24] G. Hinton, S. Osindero, Y. W. Teh. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.