JYU DISSERTATIONS 298

Rui Yan

Automatic Sleep Scoring Based on Multi-Modality Polysomnography Data

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY

JYU DISSERTATIONS 298

Rui Yan

Automatic Sleep Scoring Based on Multi-Modality Polysomnography Data

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi lokakuun 28 päivänä 2020 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
on October 28, 2020 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2020

Editors
Timo Männikkö
Faculty of Information Technology, University of Jyväskylä
Ville Korkiakangas
Open Science Centre, University of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-8329-1

ISBN 978-951-39-8329-1 (PDF)
URN:ISBN:978-951-39-8329-1
ISSN 2489-9003

ABSTRACT

Yan, Rui
Automatic sleep scoring based on multi-modality polysomnography data
Jyväskylä: University of Jyväskylä, 2020, 60 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 298)
ISBN 978-951-39-8329-1 (PDF)

Over the past decades, probably due to our hectic lifestyle in modern society, complaints about sleep problems have increased dramatically, affecting a large part of the world's population. The polysomnography (PSG) test is a common tool for diagnosing sleep problems, but the scoring of PSG recordings is an essential but time-consuming process. Therefore, automatic sleep scoring becomes crucial and urgent to settle the growing unmet needs in sleep research.

This thesis extends the previous research on automatic sleep scoring from two aspects. One is to extensively explore signal modalities and feature types related to automatic sleep scoring. This exploratory work obtains the optimal signal fusion and feature set for automatic sleep scoring, and further clarifies the contribution of signals and features to the discrimination of sleep stages. Our results demonstrate that diverse features and signal modalities are coordinative and complementary, which benefits the improvement of classification accuracy. The other one is to develop automatic sleep scoring tools that can accommodate different datasets and sample populations without adjusting model structure and parameters across tasks. Experimental results show that the joint analysis of multiple signals can improve the stability, robustness and generalizability of the proposed models. Model performance has been verified on multiple public datasets, demonstrating good model transferability between different datasets and diverse disease populations.

In summary, this research finding will advance the understanding of underlying mechanism during automatic sleep scoring and clarify the association between manual scoring criteria and automatic scoring methods. The joint analysis of multiple signals enhances model versatility, which inspires the construction of cross-model in the field of automatic sleep scoring. Moreover, the proposed automatic sleep scoring methods can be integrated with diverse PSG systems, thereby facilitating sleep monitoring in clinical or routine care.

Keywords: automatic sleep scoring, polysomnography, multi-modality analysis, deep learning, machine learning

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Yan, Rui
Automaattinen unen pisteytys multimodaalisen polysomnografian tietojen avulla
Jyväskylä: University of Jyväskylä, 2020, 60 s. (+artikkelit)
(JYU Dissertations
ISSN 2489-9003; 298)
ISBN 978-951-39-8329-1 (PDF)

Viime vuosikymmenien aikana kiireisen elämäntavan takia modernissa yhteiskunnassa unihäiriöistä tehdyt ongelmat ovat lisääntyneet dramaattisesti ja vaikuttaneet suureen osaan maailman väestöstä. Polysomnografia (PSG) -testi on yleinen työkalu unihäiriöiden diagnosointiin ja PSG-tallenteiden pisteytys on välttämätöntä, mutta aikaa vievä prosessi. Siksi automaattisen unen pisteytyksen merkitys kasvaa ja siitä tulee tärkeä ja välttämätön menetelmä, jotta voidaan vastata kasvaviin tarpeisiin unentutkimuksessa.

Tämä väitöskirja laajentaa aiempaa automaattisen unen pisteytyksen tutkimusta kahdesta näkökulmasta. Yksi on tutkia laajasti automaattiseen unen pisteytykseen liittyviä signaalimuotoja ja toimintotyyppejä. Tämän tutkimustyön tuloksena saadaan aikaan optimaalisen signaalin fuusion ja ominaisuusjoukot automaattiselle unipisteytykselle ja lisäksi se selventää edelleen signaalien ja ominaisuuksien vaikutusta univaiheiden erottamiseen. Tulokset osoittavat, että erilaiset ominaisuudet ja signaalitavat ovat koordinoivia ja täydentäviä, mikä hyödyttää luokituksen tarkkuuden parantamista. Toinen tapa on kehittää automaattisia unen pisteytystyökaluja, joihin mahtuu erilaisia aineistoja ja otospopulaatioita säätämättä mallin rakennetta ja parametreja tehtävien välillä. Kokeelliset tulokset osoittavat, että useiden signaalien yhteinen analyysi parantaa ehdotettujen mallien vakautta, kestävyyttä ja yleistettävyyttä. Mallin suorituskyky on varmistettu useilla julkisilla aineistoilla, mikä osoittaa mallin hyvän siirrettävyyden eri aineistojen ja erilaisten tautipopulaatioiden välillä.

Yhteenvetona voidaan todeta, että tämä tutkimustulos edistää ymmärrystä taustamekanismista automaattisen unen pisteytyksen aikana ja selkeyttää manuaalisten pisteytyskriteerien ja automaattisten pisteytysmenetelmien välistä yhteyttä. Useiden signaalien yhteinen analyysi parantaa mallin monipuolisuutta, mikä inspiroi ristimallin rakentamista automaattisen unen pisteytyksen alalla. Lisäksi ehdotetut automaattiset unen pisteytysmenetelmät voidaan integroida erilaisiin PSG-järjestelmiin, mikä helpottaa unen seurantaa kliinisessä tai rutiinihoidossa.

Asiasanat: automaattinen unen pisteytys, polysomnografia, multimodaalisuus analyysi, syvä oppiminen, koneoppiminen

Author

Rui Yan
Faculty of Information Technology
University of Jyväskylä
Finland
Email: ruiyanmodel@foxmail.com

Supervisors

Professor Fengyu Cong
School of Biomedical Engineering
Dalian University of Technology
China
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Timo Hämäläinen
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers

Professor Zhiguo Zhang
School of Biomedical Engineering
Shenzhen University
China

Professor Xu Lei
Faculty of Psychology
Southwest University
China

Opponent

Professor Li Hu
Department of Psychology
University of Chinese Academy of Sciences
China

ACKNOWLEDGEMENTS

I am honored to be able to conduct my doctoral research at the University of Jyväskylä. I would never forget my supervisor, Professor Fengyu Cong, who offers me the precious opportunity. Under his help, I determined the research direction and initial research questions. Professor Fengyu Cong is a caring person with contagious curiosity and extraordinary enthusiasm for scientific research. His impressive scientific attitude has always inspired me to keep proactive in research work. I shall always thank him for his excellent mentoring, positive encouragement, continuous support and constructive criticism.

I extend my sincere gratitude to Professor Tapani Ristaniemi. As my supervisor, he gives me patient guidance and unstinting support during my PhD study. He encourages us to report regularly to summarize research progress and clarify research objectives. Meanwhile, he has always been ready to provide great ideas and practical suggestions, which directly or indirectly help me in addressing some of the challenges encountered in my projects.

Meanwhile, I also wish to express my thanks to Professor Timo Hämäläinen, who is my supervisor at the end of my PhD research. He provides enthusiastic support to my research, and always actively responds to my needs. He also kindly helps me translate the Finish text in this thesis. Without his indeed help, I could have not completed the thesis.

Special thanks to the external reviewers of the thesis, Professor Xu Lei and Professor Zhiguo Zhang, and the opponent Professor Li Hu. They must devote valuable time and effort to review this thesis in such a tight schedule. Their constructive comments and invaluable suggestions have greatly helped me improve the thesis.

I want to acknowledge Professor Karen Spruyt for her selfless help with my research, especially at the initial stage of my research. Professor Karen Spruyt has been actively concerned about my research progress and provides tailored advice in time. I would like to express my genuine appreciation and thanks to Doctor Jihui Zhang, who affords precious and positive clinical coaching on my sleep research, especially on the diagnosis of sleep disturbance. I would like to thank Dr. Piia Astikainen for giving me the opportunity to learn about their sleep studies in the Department of Psychology at the University of Jyväskylä. Moreover, she has organized many seminars, which offer me an opportunity to meet many scholars and learn cutting-edge technologies.

I would like to thank my friends and colleagues Fan Li, Dongdong Zhou, Deqing Wang, Yongjie Zhu, Lili Tian, Xueqiao Li, Jia Liu, Wenya Liu, Xiulin Wang, Guanghui Zhang, and many others. They are enthusiastic and supportive. There are many inspiring discussions, from which I absorb a lot of novel knowledge and creative ideas.

My sincere thanks to the National Natural Science Foundation of China (Grant No.91748105), National Foundation in China (No. JCKY2019110B009), the Fundamental Research Funds for the Central Universities [DUT2019] in Dalian

University of Technology in China, and the China Scholarship Council (Nos. 201606060227) for providing financial support for my research work.

Last but not least, I am very grateful to my parents, sisters and husband. They are always ready to support me and help me. Their care and love are my motivation to ride on this long and winding road.

Jyväskylä 10.8.2020
Rui Yan

LIST OF ACRONYMS

AASM	American Academy of Sleep Medicine
AdaBoost	Adaptive boosting
BDT	Binary decision tree
CNN	Convolutional neural network
DS	Deep sleep
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyogram
EOG	Electrooculogram
FFT	Fast Fourier transform
FIR	Finite impulse response
FSP	Forward selection process
GA	Genetic algorithms
GUI	Graphical user interface
HHT	Hilbert-Huang transform
HMM	Hidden Markov model
ICA	Independent component analysis
IDE	Improved distance-based evaluation method
IIR	Infinite impulse response
KNN	K-nearest neighbor
LS	Light sleep
LSTM	Long short-term memory unit
N1	Non-rapid eye movement stage 1
N2	Non-rapid eye movement stage 2
N3 (SWS)	Non-rapid eye movement stage 3; slow wave sleep; deep sleep
NB	Naive Bayes
NREM	Non-rapid eye movement
PSG	Polysomnography
REM (R)	Rapid eye movement
RF	Random forest
RNN	Recurrent neural networks
SBS	Sequential backward selection
SHHS	Sleep Heart Health Study
STFT	Short-time Fourier transform
SVM	Support vector machine
W	Wakefulness
WT	Wavelet transform

FIGURES

TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

PI. Rui Yan, Chi Zhang, Karen Spruyt, Lai Wei, Zhiqiang Wang, Lili Tian, Xueqiao Li, Tapani Ristaniemi, Jihui Zhang, and Fengyu Cong. 2019. "Multi-Modality of Polysomnography Signals' Fusion for Automatic Sleep Scoring." Biomedical Signal Processing and Control 49: 14–23. DOI: 10.1016/j.bspc.2018.10.001

PII. Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi and Fengyu Cong. 2019. "An Automatic Sleep Scoring Toolbox: Multi-modality of Polysomnography Signals' Processing." In Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019), pp. 301-309. DOI: 10.5220/0007925503010309

PIII. Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi and Fengyu Cong. 2020. "Automatic Sleep Scoring Toolbox and Its Application in Sleep Apnea." In: Obaidat M. (eds) E-Business and Telecommunications. ICETE 2019. Communications in Computer and Information Science, vol 1247, pp.256-275. Springer, Cham. DOI: 10.1007/978-3-030-52686-3_11

PIV. Rui Yan, Fan Li, DongDong Zhou, Tapani Ristaniemi and Fengyu Cong. 2020. " A Deep Learning Model for Automatic Sleep Scoring using Multimodality Time Series." In 28th European Signal Processing Conference (EUSIPCO 2020). 5 pages. IEEE, Amsterdam, Netherlands.

PV. Rui Yan, Fan Li, Dongdong Zhou, Tapani Ristaniemi, and Fengyu Cong. 2020. "Automatic Sleep Scoring: A Deep Learning Architecture for Multi-modality Time Series." Submitted to the Journal of Neuroscience Methods. (Under review)

AUTHOR'S CONTRIBUTION

Taking into account the suggestions and comments of coauthors, the author of this thesis made the following contributions to the attached five publications: she built the methodology, conducted all experiments and wrote the original manuscripts.

1 INTRODUCTION

This chapter firstly introduces research background and research motivation, then provides an overview of the whole research, and finally gives dissertation structure.

1.1 Research background

Sleep accounts for about one-third of human lifespan and is thus a vital physiological process (Pace-Schott and Hobson 2002). According to recommendations of the National Sleep Foundation (Hirshkowitz et al. 2015), the appropriate sleep duration is the longest for newborns, about 14 to 17 hours, and then gradually decreases to 9-11 hours for school-age children, 7-9 hours for young adults and adults, and the recommended sleep duration for the elderly is 7-8 hours. Besides sleep duration, sleep structure is also an important factor affecting sleep health. According to the latest standard developed by the American Academy of Sleep Medicine (AASM) (Berry et al. 2016), nocturnal sleep is split into five stages: wakefulness (W), non-rapid eye movement (NREM) stages 1, 2, 3 and rapid eye movement stage (R). FIGURE 1 displays an idealized sleep structure of healthy adults.

Adequate and high-quality sleep is essential to maintain of some basic physiological activities, such as fatigue recovery (Lim and Dinges 2008), memory consolidation (Stickgold and Walker 2007), immunity (Park et al. 2016), metabolism (Porkka-Heiskanen and Kalinchuk 2011), and endocrine (Cox and Olatunji 2016). Conversely, sleep loss or distorted sleep structure has negative impacts on our wellbeing, resulting in cognitive deficits, mood disturbances (Dinges et al. 1997), neuronal loss (Jan et al. 2010) and even brain atrophy (Van-Someren et al. 2018; Spira et al. 2016). Moreover, some studies have revealed that distorted sleep might be an early indicator of some neurodegenerative diseases (Iranzo et al. 2006), such as Alzheimer's disease and Parkinson's disease, and sleep disturbance may promote the progression of certain neurodegenerative

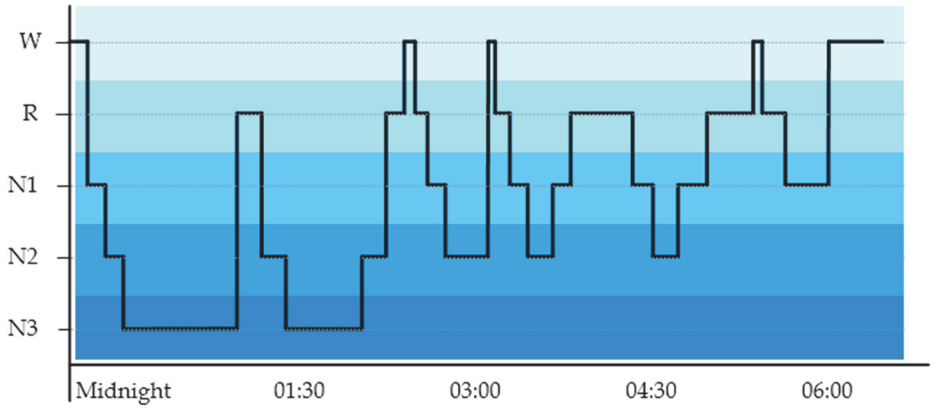


FIGURE 1 Idealized sleep structure of healthy adults.

diseases. Therefore, paying attention to sleep health is conducive to the early diagnosis of sleep-related diseases and might slow their deterioration.

Nowadays, probably due to our hectic lifestyle in the 21st century, complaints about sleep problems increase dramatically among people. According to the international classification of sleep disorders, there are about 80 types of sleep disorders, including insomnia, hypersomnia, bruxism, sleep apnea and sleep-related movement disorders (Thorpy 2017). Among them, insomnia is the most common, which affects approximately 30% of the world's population (Berry et al. 2016). The prevalence of sleep problems imposes a heavy burden on family well-being, public security and global finance. Worse of all, most individuals with sleep disorders cannot get timely diagnosis and treatment, and thus condition aggravation (Altevogt and Colten 2006), which may be attributed to the low awareness of sleep health and limited clinical resources.

The gold tool for monitoring sleep quality and diagnosing sleep problems is polysomnographic (PSG) test. FIGURE 2 gives an example of PSG recordings, which generally contains tens of sleep signals, such as electroencephalogram (EEG) to measure brain activity, electrooculogram (EOG) to measure eye movements, electromyogram (EMG) to measure chin muscle activity, electrocardiogram (ECG) to measure heart electrical activity, pulse oximetry, respiration, body movement, etc. Rich PSG signals can comprehensively record physiological activities during sleep, thus providing accurate diagnosis of sleep-related diseases. The PSG test is usually performed in a dedicated hospital or sleep laboratory for one to two nights, but there are also portable systems for PSG tests at home. After the PSG test, sleep technologists will first score sleep recordings by marking sleep stages and sleep events (such as abnormal breathing, leg movements) according to R&K rules (Rechtschaffen and Kales 1968) or the AASM standard (Berry et al. 2016). Then, sleep physicians will review the results to determine sleep problems the patient may have. The process of assigning a sleep stage to each sleep segment is called sleep scoring, which is the foundation of sleep research.

Due to the pivotal role of PSG test in sleep research, most sleep laboratories qualified for PSG tests are in an overloaded state. Tedious manual scoring is partially responsible for that busy state. In clinical, sleep scoring is performed

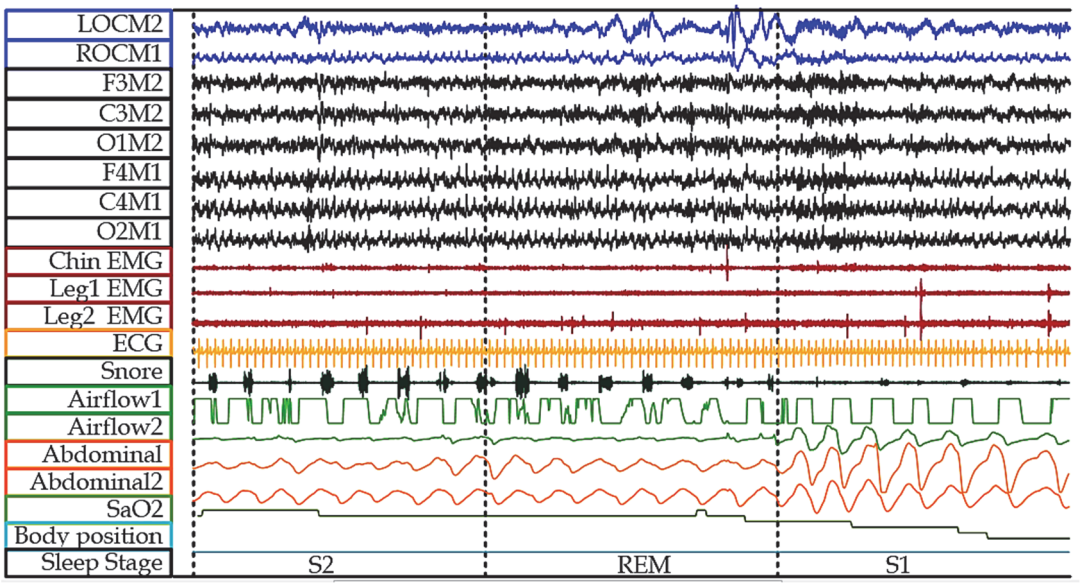


FIGURE 2 Screenshot of a PSG recording

manually, which is labor-intensive and subjective (Malafeev et al. 2018). Previous studies have reported that the annotation of an 8-h recording requires approximately 2-4 hours (Hassan and Bhuiyan 2016), and the inter-scorer reliability is about 80% (Danker-Hopfe et al. 2009). Therefore, with the development of computational technology, automatic sleep analysis is crucial and urgent to help address the growing unmet needs for sleep research.

1.2 Research motivation

Numerous attempts so far have been made in the field of automatic sleep scoring to extricate sleep technologists from heavy manual scoring. From the perspective of computational methods, these algorithms can be divided into conventional scoring methods and deep learning methods. Conventional scoring methods are mainly based on explicit feature extraction of sleep samples (Boostani, Karimzadeh, and Nami 2017; Hassan and Hassan Bhuiyan 2016; Seifpour et al. 2018). These conventional methods are easy to explain or present their model behaviors, but their classification accuracy highly relies on the effectiveness of hand-crafted features. Most recently, algorithms based on deep learning are relatively prevalent because they do not require explicit feature extraction and are particularly suitable for big data approach (Patanaik et al. 2018; Mousavi et al. 2019; Zhang et al. 2019). Moreover, studies based on deep learning have introduced some novel classification schemes to mimic the way sleep technologists perform manual sleep scoring, such as multiple-input-one-output scheme (Chambon et al. 2018), one-input-multiple-output scheme (Phan et al. 2019a) and sequence-to-sequence model (Phan et al. 2019b). These novel schemes utilize the dependence of consecutive epochs and achieve good classification performance.

In terms of signal modality, the automatic sleep scoring algorithms proposed in previous studies have explored a variety of signals recorded by the PSG test, including EEG, EOG, EMG, ECG, airflow and respiratory effort (Šušmáková and Krakovská 2008; Supratak et al. 2017; P. Fonseca et al. 2015). Although there are numerous approaches to automate sleep scoring, there is no automatic sleep scoring algorithm that can establish itself as an industry standard neither in research nor in clinical applications. Most studies built their models for specific datasets and certain signal modalities, which limits their applications. Given different datasets and signal modalities, there are no unified comparison standards between studies, and therefore the specific contribution of signals and features to sleep scoring is still not univocal. There is no doubt that more elaborate exploration is necessary in order to promote the application of automatic sleep scoring methods. Here, we list four key issues below.

- The first one is which fusion of sleep signals can give the optimal classification results? How to choose signal modality according to the analysis target?
- The second issue is which features contribute to the discrimination of sleep stages? Are effective features in automatic sleep scoring methods consistent with stage characteristics specified in scoring criteria?
- The third is to explore the impact of temporal dependence on classification results? How to use context information to improve classification accuracy?
- The fourth is how does the model perform on other datasets? How to tackle the mismatched channels in different sleep studies?

Finding solutions to the above four issues is the motivation of this dissertation.

1.3 Introductory Overview

This thesis studies automatic sleep scoring methods by using PSG recordings. The whole thesis is divided into two parts.

The first part focuses on the exploration of underlying mechanisms of automatic sleep scoring. In order to investigate the effect of signal modalities on sleep scoring, we compare 12 different fusions of PSG signals and further demonstrate their performance on the classification of 2 to 5 classes of sleep stages. To elucidate features' contribution to automatic sleep scoring, we exploit a wide variety of features, covering statistical features, frequency features, time-frequency features, fractal, entropy and other nonlinear characteristics. The optimal feature fusion used to distinguish each pair of sleep stages is finally picked out. Studies of this part are indispensable for revealing interesting insights about how automatic scoring models associate specific features and signals with different sleep stages. Conclusions have been tested on different datasets, which is able to provide a direction for future research.

The second part of our research is centred on applications of automatic sleep scoring models. Most studies design their models for specific datasets and certain signal modalities, and thus task-specific modifications are required if the model is used in different tasks. That modification is difficult and even inefficient, especially for non-experts facing complex practical conditions. Besides, the recorded sleep signals may be different due to the diverse monitoring devices and specific experimental motivations. Therefore, we aim to develop a versatile and automatic scoring architecture that can handle various numbers of input channels and several signal modalities without model modifications across tasks. Based on our previous research on conventional automatic sleep scoring methods, we further propose two models to automate sleep scoring, one based on conventional machine learning methods and the other one based on deep learning networks. Model availability and versatility are demonstrated by public datasets with disparate attributes. The detailed model structure and experimental results will be further discussed in Chapter 3.

Collectively, articles PI-PIII contribute to the first research part. Article PI firstly assesses the performance of several feature selection methods and classifiers, and further develops an automatic method to split sleep into five stages using multi-modality PSG signals. Based on that, Paper PII constructs an automatic sleep scoring toolbox with the capability of multi-signal processing, and further investigates classification performance of 12 signal fusions in distinguishing 2-5 sleep stages. Article PIII modifies the toolbox, verifies its performance on healthy adults and patients with sleep disorders, and further clarifies features' contribution in distinguishing stage pairs.

Articles PIII-PV contribute to the second research part. Using random forest to classify the extracted features, article PIII develops an automatic sleep toolbox, and verifies its performance in different populations. Article PIV, departing from hand-crafted features, aims to construct an end-to-end CNN model to automate sleep scoring. Given the dependencies between sleep stages, article PV adds an LSTM unit to the CNN architecture to fully exploit the current and context information, thereby improving classification performance on diverse sample populations.

1.4 Structure of dissertation

The rest of this dissertation is organized as follows.

Chapter 2 reviews state-of-the-art methods in the field of automatic sleep scoring, together with their theoretical descriptions. Chapter 3 summarizes the five articles included in this thesis. Chapter 4 lists the findings and limitations of this study and further discusses the direction for future research. Chapter 5 draws the summary of the whole research work.

2 RESEARCH BACKGROUND

This chapter starts by introducing basic knowledge of sleep scoring (Section 2.1), which is the foundation of automatic sleep scoring. Then, we review recent studies related to conventional scoring methods (Section 2.2) and deep learning-based methods (Section 2.3). The main purpose of this chapter is to give readers an idea of various sleep scoring methods and to immerse readers into the topic before we review thesis contributions.

2.1 Manual sleep scoring

So far, manual sleep scoring is still the most acceptable method in clinical. The sleep technologist or physician scores sleep stages and sleep events according to R&K rules (Rechtschaffen and Kales 1968) or the AASM standard (Berry et al. 2016). The earlier R&K rules split sleep into five different stages: non-rapid eye movement (NREM) stages 1, 2, 3 and 4 and rapid eye movement stage (stage R). But the latest AASM standard merges NREM stages 3 and 4 into N3 due to their predominant low-frequency oscillations in brain activity. For the standard operation of sleep scoring, the latest AASM standard recommends three EEG derivations, namely F4-M1, C4-M1 and O2-M1, or the combination of Fz-Cz, Cz-Oz and C4-M1 is also acceptable. All electrodes are placed and denominated following the international 10-20 system, where M1 is the standard reference electrode referring to the left mastoid. In addition to brain activity recorded by EEG electrodes, the AASM standard also recommends the use of EOG electrodes to record eye movements and the use of EMG electrodes to measure chin muscle activity, thereby improving scoring performance. These three signal modalities constitute the fundamental part of a polysomnography (PSG) recording.

After capturing PSG recordings, sleep technologists or physicians firstly divide sleep recordings into 30-second intervals called one epoch, and then assign a stage to each epoch based on the amplitude and frequency characteristics of sleep signals, as described in TABLE 1. In addition to information from the

current stage, information from neighbor epochs is also worth considering to score the current epoch under certain conditions. We summarize the dependence relationship between sleep stages in TABLE 2. What stands out in TABLE 2 is that sleep events such as arousals or body movements can also result in a degradation of the sleep process, besides natural transitions of sleep stages. Therefore, information from neighbor epochs is particularly beneficial for the recognition of stage transitions. It should be mentioned that TABLE 1 and TABLE 2 introduce scoring criteria and stage transition rules according to the AASM standard (Berry et al. 2016), but this article does not intend to completely repeat the AASM standard. Therefore, for strict scoring criteria, the reader should refer to Berry et al.'s research. The main purpose of these two tables is to give the reader a general understanding of manual sleep scoring in order to better understand the content below.

TABLE 1 Sleep scoring criteria for healthy adults following the AASM standard, Chapter IV, pp. 16–31(Berry et al. 2016)

Sleep stage	Rule	Description
Stage W	E.2	a) EEG shows mixed beta and alpha activity or predominantly alpha activity (8–13 Hz) as the eyes remain closed. b) EOG shows eye blinking, reading eye movement or rapid eye movement. c) Submental EMG is relatively high.
Stage R	I.2; I.3	a) EEG is characterized by low-amplitude, mixed-frequency brain activity without K complexes or sleep spindles. Sawtooth waves (2-6Hz, sharply contoured triangular) are strongly supportive of the presence of stage R. b) Rapid eye movement is characteristic of stage R. c) EMG usually shows the lowest level of the entire recording.
Stage N1	F.2; F3	a) EEG is dominated by 4–7 Hz activity (theta wave). Vertex sharp wave with duration <0.5 seconds may occur. b) EOG often shows slow eye movements. c) EMG amplitude is variable but often lower than that in stage W.
Stage N2	G.2; G.note6-7	a) EEG is characterized by predominant theta activity and occasional quick bursts of faster activity. Sleep spindles (12–14 Hz) and K complexes may appear. b) EOG usually shows no eye movement activity. c) Chin EMG shows variable amplitude, but it is usually lower than that in stage W and maybe as low as that in stage R.
Stage N3	H.2; H.note5-6	a) EEG activity is marked by high-amplitude slow waves (amplitude>75μV; frequency<2Hz). b) Eye movements are not typically seen. c) Chin EMG is of variable amplitude, often lower than that in stage N2 and sometimes as low as that in stage R.

TABLE 2 Transition rules of sleep stages following the AASM standard, Chapter IV, pp. 16–31(Berry et al. 2016)

Transition pattern	Rule	Characteristics
W-N1	F2; F3; F4	Low-amplitude and mixed-frequency EEG, theta activity, vertex sharp wave, slow eye movement
W-R	I.2-I.4	Low-amplitude and mixed-frequency EEG, low chin EMG, eye movement
N1-N2	G.2	Sleep spindle, K-complex unassociated with arousal
N1-R	I.2-I.4	Low-amplitude and mixed-frequency EEG, low chin EMG, eye movement
N2-W	E.2	Alpha rhythm, eye blink, eye movement, high chin muscle tone
N2-R	I.2-I.4	Low-amplitude and mixed-frequency EEG, low chin EMG, eye movement
N2-N1	F2; F3; F4	Low-amplitude and mixed-frequency EEG, theta activity, vertex sharp wave, slow eye movement
	F.5; G.6.b	Arousal interrupt followed by low-amplitude and mixed-frequency EEG
	G.6.c	Major body movement, slow eye movement
N2-N3	H.2	≥20% of an epoch consists of slow-wave activity
N3-N2	G2	Sleep spindle, K-complex unassociated with arousal
	G5	No arousal, not meet criteria for stage N3
R-W	E.2	Alpha rhythm, eye blink, eye movement, high chin muscle tone
R-N1	F.6; I.6.b-d	Arousal interrupt, slow eye movement, and low-amplitude and mixed-frequency EEG
R-N2	I.6.e	non-arousal associated K complex, sleep spindle, no rapid eye movement
R-N3	H.2	≥20% of an epoch consists of slow-wave activity

2.2 Conventional scoring methods

2.2.1 System based on expert knowledge

The algorithm based on expert knowledge is a kind of machine learning method that is nearest to manual scoring criteria. The typical characteristic of these algorithms is the paradigm of "if-then". For example, if EEG amplitude>75μV, then the sleep stage is SWS. Liang et al. extracted fourteen rules from temporal and spectral measures of PSG signals in order to score sleep stages, and achieved an accuracy of 86.68% on 17 healthy participants (Liang, Kuo, Hu, and Cheng 2012). In their recent research, they constructed a genetic fuzzy inference system based on nine expert rules, reaching an accuracy of 86.44% on 24 participants with or without sleep loss (Liang et al. 2016). The most recent study combined

deep learning models with expert-defined rules, providing a good trade-off between model interpretability and classification accuracy (Al-Hussaini et al. 2019).

This type of methods is highly interpretable since each of their decisions can be traced back to an understandable rule. Therefore, these methods achieve complete control of their system. A disadvantage is that the selected rules may not be optimal, and the linear model composed of simple paradigms of "if-then" is not enough to describe the complex and time-varying sleep patterns. Consequently, there are not so many papers that build their models based on expert rules exclusively.

2.2.2 Machine learning based on hand-crafted features

In recent decades, there has sprung up numerous studies in the field of automatic sleep scoring, which indicates increasing attention of sleep practitioners to the interdisciplinary cooperation in order to automate sleep scoring. In general, scoring methods based on hand-crafted features comprise five common steps, including signal preprocessing, feature extraction, feature selection, classification and post-processing (Boostani, Karimzadeh, and Nami 2017). Among these steps, feature selection and post-processing are optional steps, which appear in some papers. The following briefly describes some commonly used techniques in each step according to the processing order of sleep signals.

2.2.2.1 Signal preprocessing

Signal preprocessing is requisite for the conventional methods based on hand-crafted features, because sleep signals (especially EEG signals) usually contain artefacts from many different sources. Common artefacts may come from power line interference (50Hz or 60Hz), electrode shedding, sweat or other electrophysiological signals, such as eye movement, muscle activity and cardiac signals (Radüntz et al. 2015). These artefacts severely affect the extraction of useful information, which may distort the quantitative signal analysis. Therefore, in order to eliminate the effect of artefacts and improve signal quality, several methods have been exploited in previous articles.

Digital filtering is a simple and commonly-used method, including the finite impulse response (FIR) filter and the infinite impulse response (IIR) filter. Digital filtering shows good performance on the elimination of technical artefacts, for example, a Butterworth notch filter to remove power line interference (50Hz or 60Hz) (Şen et al. 2014), or a bandpass Butterworth filter to screen out undesired parts (Khalighi et al. 2016).

In terms of electrophysiological artefacts caused by eye movement, muscle activity and cardiac signal, they are relatively difficult to remove since their frequency overlaps with target signal and their appearance is diverse. Therefore, more elaborate preprocessing is required, such as wavelet transform (WT) (Dimitriadis, Salis, and Linden 2018), short-time Fourier transform and independent component analysis (ICA) (Radüntz et al. 2015). It should be noted

that elaborate preprocessing is time-consuming and may suffer from performance degradation in practical applications. Therefore, in order to improve computational efficiency and model feasibility, we chose a coarse preprocessing in our studies, which applied only Butterworth filters. Readers can refer to the article [PI] for more details.

In the field of automatic sleep scoring, besides filtering, it is also worth to consider signal normalization and feature standardization to eliminate individual difference and accelerate algorithm convergence. According to the AASM standard, sleep signals also need to be split into 30 seconds per epoch, each epoch corresponding to one sleep stage.

2.2.2.2 Feature extraction

Feature extraction is essential for conventional scoring methods, which directly affects classification performance. Diverse features are responsible for describing sleep signals from multiple aspects, thereby providing an accurate analysis. There are a wide variety of techniques for feature extraction, including but not limited to statistic methods, Fourier transform, wavelet analysis and Hilbert transform. The extracted features can be broadly divided into four categories. In the following, we briefly present some popular features together with their extraction methods.

Time-domain features are simple but practical, which indicate signal changes in amplitude distribution and morphological characteristic. Some widespread features are mean, median, standard deviation, skewness, kurtosis, percentile and Hjorth parameters. A detailed description can be found in our articles [PI and PII] or Şen et al.'s research (Şen et al. 2014).

Frequency-domain features. Many electrophysiological activities exhibit constant variations and rhythmic dynamics. For example, sleep EEG contains slow waves (0.5–2.0Hz), delta waves (0–3.99Hz), theta waves (4–7.99Hz), alpha waves (8–13Hz) and beta waves (13-30Hz) (Berry et al. 2016). Frequency-domain features provide a pithy description for the variant and dynamic temporal structure embedded in sleep signals. Commonly-used features are power spectral density and power ratios of sub-frequency bands. Their calculations are mainly based on the fast Fourier transform (FFT). In order to meet the time-invariant assumption in the Fourier transform, sleep signals may be divided into mini-epoch by Hamming window or Hanning window (Šušmáková and Krakovská 2008).

Time-Frequency domain features are suitable to the analysis of non-stationary electrophysiological signals because they can provide descriptions with balanced resolution both in the time domain and in the frequency domain (Wacker and Witte 2013). The most commonly used time-frequency analysis techniques are short-time Fourier transform (STFT), wavelet transform (WT) and Hilbert-Huang transform (HHT) (Boostani, Karimzadeh, and Nami 2017). For sleep EEG, the time-frequency analysis is suitable for capturing sleep events (sleep spindles, K-complex and arousals) and analyzing the characteristics of rhythm waves (such as delta waves, theta waves, alpha wave and beta waves).

Nonlinear features provide a complex and dynamic description for nonstationary electrophysiological signals, which are widely used in automatic sleep scoring methods (Krakovská and Mezeiová 2011; Ebrahimi et al. 2013; Şen et al. 2014). Well-known nonlinear measures are energy, entropy and chaotic index. Energy is a measure that indicates instantaneous changes of signal amplitude. Commonly used energy features are mean energy, mean Teager energy, the second differences, the 4^{th} power, and so on (Şen et al. 2014). Entropy is a measure used to quantify the irregularity and unpredictability of signals (Molina-Picó et al. 2011). Signals with regular distribution have smaller entropy value than those with irregular distribution (Boostani, Karimzadeh, and Nami 2017). The famous entropy features are Shannon entropy, sample entropy, approximate entropy, Renyi's entropy, Tsallis entropy, permutation entropy and multi-scale entropy (Liang, Kuo, Hu, Pan, et al. 2012; Liu and Yue 2009; Acharya et al. 2005; Khalighi et al. 2013). Chaotic index is to measure the roughness or irregularity of signal morphology, for example, Hurst exponent, Katz fractal dimension, Petrosian fractal dimension and Higuchi fractal dimension. The calculation burden of chaos indices is relatively high, but their estimation value or fast calculation methods are widely used in automatic sleep scoring (Jiang et al. 2019; Acharya et al. 2005).

Mutual-based features are used to measure interactive information between multi-modality signals, which provide a new perspective for the same phenomenon beyond the perspective provided by exploiting each modality separately (Lahat, Adali, and Jutten 2015). For example, Gharbali et al. applied several distance-based features to measure the similarity of PSG signals and further investigated their contribution to automatic sleep scoring. In that article, they concluded that these measures, especially the similarity of EEG and EOG, were useful for automatic sleep scoring (Gharbali, Najdi, and Fonseca 2018). Other measures between two signals, such as correlation coefficient, coherence, phase angle and mutual information, were also explored in previous studies (Šušmáková and Krakovská 2008).

2.2.2.3 Feature selection

Feature selection is an optional step in conventional scoring methods. Some studies use feature selection methods to find effective subset from candidate features, thereby restraining overfitting, shortening training time and simplifying classification model. Some feature selection methods have been proposed in previous studies. For example, the forward selection process (FSP) and sequential backward selection (SBS) are two simple strategies, which check the features one by one to verify whether or not the newly added feature improves classification accuracy (Peng, Long, and Ding 2005). The feature that benefits the improvement of classification accuracy is selected into the optimal feature set. These two feature selection methods are simple but prone to fall into local minima due to the inability to re-evaluate feature effectiveness (Boostani, Karimzadeh, and Nami 2017). In addition, Genetic algorithms (GA) and ReliefF

are also commonly used in studies. The results of feature selection are affected by the distribution of feature values and the algorithm of feature selection.

In our study [PI], we investigated four widely used feature selection methods, namely ReliefF method, improved distance-based evaluation method (IDE), GA and FSP, and further compared their performance in automatic sleep scoring. Experiment results showed that different feature selection methods gave diverse features subsets, but they all comprised statistic features, frequency features, time-frequency features and nonlinear features. The result further demonstrated that the rich feature contributed to the enhancement of classification accuracy.

2.2.2.4 Stage classification

Various classifiers have been explored in studies of automatic sleep scoring, including unsupervised classifiers and supervised classifiers. The classifier learns classification strategies from extracted features to court the ability of assigning appropriate sleep stages to new samples.

Unsupervised classifiers, also known as clustering, learn classification strategies from unlabeled sleep data, thus greatly saving time cost in preparing labels for huge amounts of sleep data. The clustering algorithm groups a set of samples based on their similarity, so samples in the same group are more similar to each other than those in different groups (Alpaydin 2020). Due to the complexity of sleep phenomenon, the overlapped information between certain stages and the lack of prior knowledge, the applications of clustering methods are not optimistic in automatic sleep scoring. Several articles applied clustering methods to automate sleep scoring. El-Manzalawy et al. developed an unsupervised model using k-means clustering to distinguish sleep-wake states and reached an accuracy of 85% on 37 participants (El-Manzalawy, Buxton, and Honavar 2017). Rodríguez-Sotelo et al. used J-means clustering on entropy-based features, achieving an average accuracy of 80% in the classification of five sleep stages (Rodríguez-Sotelo et al. 2014).

In contrast, supervised classifiers are popular in studies of automatic sleep scoring. The widespread classifiers are support vector machine (SVM), random forest (RF), K-nearest neighbor (KNN), Naive Bayes (NB) and so on. The classifier description can be found in the following related studies. Huang et al. sent twelve selected features to the SVM classifier to assign sleep into six stages, obtaining an accuracy of 92.34% on thirty recordings (Huang et al. 2020). After extracting spectral features using wavelet transform, Hassan and Bhuiyan constructed a classification scheme based on RF classifier that yielded an accuracy of 90.38% (Hassan and Bhuiyan 2016). They further compared the RF classifier with ten other widely used classifiers, including NB, KNN, least-square support vector machine and adaptive boosting (AdaBoost), and concluded that RF outperformed others on their experiment data.

In our study [PI], we compared five widely used classifiers, namely KNN, NB, RF, SVM and binary decision tree (BDT). Experiment results showed that the RF classifier was not very sensitive to outliers of features set, indicating strong

robustness of the classifier. Therefore, in our subsequent research, we were prone to choose the RF classifier.

2.2.2.5 Post-processing

Post-processing of classification results has been explored in some studies to improve classification accuracy using context information or expert knowledge. The simplest post-processing method was to average probabilities across neighbor epochs (Klok et al. 2018). Similarly, a multi-layer perceptron (Patanaik et al. 2018) or auto-encoder (Golmohammadi et al. 2019) was constructed to reweight probabilities across neighbor epochs. Some papers built smoothing rules for sleep stages based on expert knowledge, such as hard smoothing rules (Liang, Kuo, Hu, and Cheng 2012) and fuzzy smoothing rules (Tian and Liu 2005). The hidden Markov model (HMM) and its variants were common in the post-processing step, in which HMM was used to capture stage transition rules from training data (Li et al. 2018; Jiang et al. 2019). Then, these learned transition rules were applied to smooth classification results. These post-processing methods were mainly to make up for the insufficiency of conventional classifiers by exploiting stage transition rules, and thus corrected some wrong classification results. One possible outcome of post-processing was to capture the slow-changing characteristics of most sleep stages, but to penalize short, abrupt changes such as transitory awakening. Therefore, the improved accuracy of some stages might come at the expense of others.

2.3 Deep learning based methods

Deep learning networks are a branch of machine learning methods based on artificial neural networks. In general, as shown in FIGURE 3, deep learning architecture consists of three parts: input layer, hidden layer and output layer (LeCun, Bengio, and Hinton 2015). The input layer is where the data enters the network. A deep learning network usually contains multiple hidden layers, and each layer contains multiple processing units called neurons. Each neuron in the network owns a weight and a bias, which determine the neuron's contribution to final decisions. Each neuron receives its input, further calculates the input by a complicated function determined by weights, biases and activation functions, and then passes its result to the next layer. In that way, the input data is calculated and refined hierarchically until reaching the final output layer. For a classification problem, the final layer is usually a decision layer, which outputs a probability matrix where the prediction for each sample corresponds to the class with the maximum probability. The network applies a cost function, such as cross-entropy, to measure the total error between network's outputs and expected outputs, and then uses gradient descent and backpropagation to iteratively update weights and biases of neurons in each layer, thereby minimizing the output error.

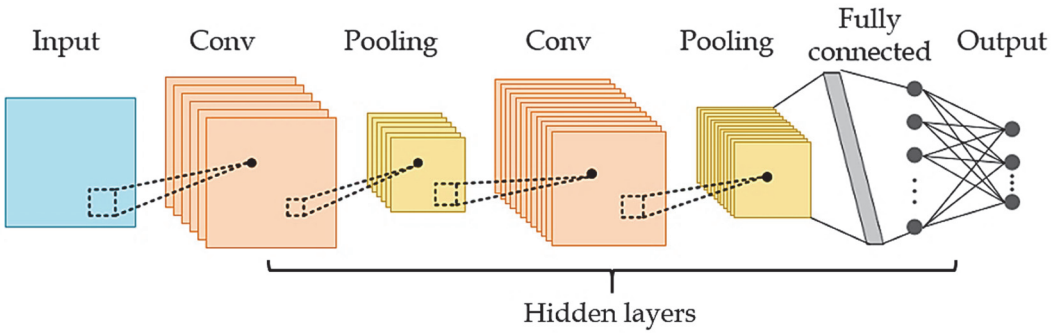


FIGURE 3 An illustration of deep learning architecture.

Deep learning algorithms are very popular in recent years because they avoid the problem of vanishing gradient or exploding gradient that occurs during backpropagation in traditional neural networks (Balas, Kumar, and Srivastava 2020). That character increases the depth of neural networks to hundreds of layers, so the network can extract high-level features and thus can handle more complicated problems. So far, deep learning networks have made dramatic progress in diverse fields including computer vision, natural language processing, self-driving car and healthcare. In the field of automatic sleep scoring, widely used networks are the convolutional neural network and the long-short time memory network. In the following, we will introduce the methodology of these two networks, and review their applications in automatic sleep scoring.

2.3.1 Convolutional neural network

Convolutional neural network (CNN) also includes an input layer, an output layer and multiple hidden layers, but its hidden layers are typically composed of a few convolution layers, as shown in FIGURE 3. The convolution layer performs convolution operations on its input to extract features (LeCun, Bengio, and Hinton 2015). A few parameters of the convolution layer are set by the user, including the number of filters, filter shape, stride and the activation function. We present a schematic diagram of calculations in a CNN neuron in FIGURE 4, where the filter shape is set to 3 × 3, the stride is 2, and the activation function is Relu. From FIGURE 4, it is easy to get that the number of filters directly controls the capacity of layer output, thereby determining network complexity. The filter shape determines the size of convolution kernels, and thus affects the receptive field of convolution operation. The stride controls the overlap of windows, and thus determines the dimension of outputs. Therefore, the output of convolution layer can be described by Eq.(1).

$$X^{(l)} = f^{l}\left(X^{(l-1)} * W^{(l)} + b^{(l)}\right) \tag{1}$$

where $X^{(l-1)}$ is the input of layer l and also the output of the previous layer, $X^{(l)}$ the output of layer l, $W^{(l)}$ the weight matrix, $*$ the convolution operation, $b^{(l)}$ the bias, and $f^{(l)}$ is the activation function.

The activation function attaches itself to each neuron in the network, and it determines whether the neuron's output should be activated or inhibited. The

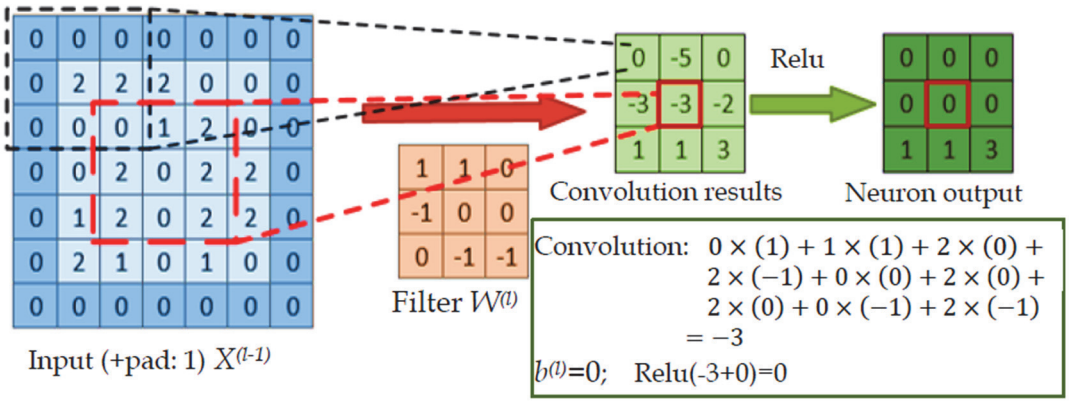


FIGURE 4 An illustration of the calculation of CNN neurons.

use of activation functions increases network nonlinearity, thereby facilitating the extraction of high-level features. There are several commonly used activation functions. The most popular one in CNN is the rectified linear unit (relu) function (Nair and Hinton 2010), which is defined as,

$$f(x) = max(0, x) \tag{2}$$

The second one is the tanh function, which is defined as,

$$f(x) = tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \tag{3}$$

The function softmax is usually used in the output layer, which can be described by,

$$f(x) = e^{x_i}/\sum_{j=1}^{J} e^{x_j}, i = 1, \ldots, J \tag{4}$$

In general, a convolution layer is followed by a pooling layer to compress features and reduce the dimension of outputs. The pooling layer helps to improve the network's nonlinearity and reduce calculation parameters. Two common pooling methods are average-pooling and max-pooling, which calculate the average value and the maximum value of elements in the pooling window, respectively. The user needs to set two pooling parameters: pooling size and stride, which respectively determines the size of pooling windows and the step length of window shift.

2.3.2 Recurrent neural network

Recurrent neural networks (RNN) are a type of artificial neural network, in which neurons consider not only the current input but also the previous state (Hochreiter and Schmidhuber 1997). Therefore, the RNN can use previous information for the current decision. This advantage has brought it a great success in sequence data processing, such as natural language processing and speech recognition. The most famous RNN variant is the long short-term memory (LSTM) unit, which is noted for avoiding gradient explosion that occurs in traditional RNNs. A general architecture of LSTM unit is illustrated in FIGURE 5.

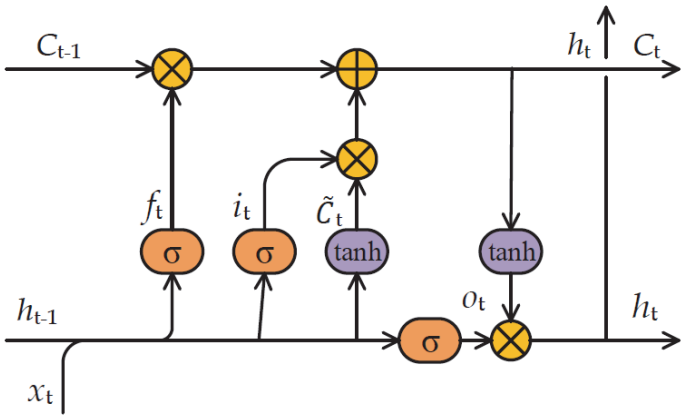


FIGURE 5 An illustration of the LSTM unit.

The LSTM unit modulates the flow of information through three gates (an input gate, an output gate and a forget gate) and remembers the valuable information through a cell (Dong et al. 2018). The three gates are controlled by sigmoid functions, see Eq (5). The sigmoid function outputs a number between 0 and 1, where 0 means completely discarded and 1 means completely accepted.

$$f(x) = \sigma(x) = 1/(1 + e^{-x}) \tag{5}$$

The forget gate controls the cell's capacity for the current input x_t and the previous output h_{t-1}. Its output can be calculated by,

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{6}$$

Similar to the forget gate, the outputs of input gate and output gate can be described as,

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{8}$$

Then, the function tanh creates a new candidate current state for the cell,

$$\tilde{C}_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{9}$$

Update cell state by adding new information from the current input,

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \tag{10}$$

Finally, the output of the LSTM unit can be obtained by,

$$h_t = o_t \circ tanh(C_t) \tag{11}$$

where W and b represent the weight and the bias, x_t the input to the LSTM unit, and the operator $\circ$ denotes the Hadamard product.

2.3.3 Hyperparameter optimization

As previously mentioned, deep learning networks contain some parameters that need to be set by the user during model constructing, such as the number of filters and kernel size in CNN. These parameters are called hyperparameters, which has

to be tuned by users so that the model can achieve optimal performance. Hyperparameter optimization is to find a set of hyperparameters that produces the optimal model performance. The most widely used strategies for hyper-parameter optimization are grid search and manual search, which exhaustively search a hyperparameter subset that is manual specified based on the user's judgment or experience. These strategies suffer from the curse of dimensionality (Yu and Zhu 2020). Random search is a variant of the grid search, which performs a random search in hyperparameter subset instead of the exhaustive search. Random search has proven its superiority, especially in the cases that only a small number of hyperparameters affect model performance (Bergstra and Yoshua Bengio 2012). This algorithm also suffers from the curse of dimensionality since it is still a computationally intensive method (Yu and Zhu 2020). Some automatic optimization strategies attempt to model hyperparameters and network performance by using the previous outcomes, and thus evaluate a promising hyperparameter set, such as Bayesian optimization and Tree Parzen Estimator. Experimental results have shown these strategies are prone to obtain better results in fewer searches compared to grid search and random search (Bergstra, Yamins, and Cox 2013; Fernández-Varela, Hernández-Pereira, and Moret-Bonillo 2018). In addition, some new approaches have also been developed, for example, evolutionary optimization (Miikkulainen et al. 2019) and population-based training (Jaderberg et al. 2017).

Although there are many available hyperparameter optimization methods and available toolboxes, such as scikit-learn and hyperopt, the selection of hyperparameters is still a tricky problem for engineers in the field of machine learning. High computational cost and specialized knowledge should be partly responsible for that issue. In view of the difficulty of tuning hyperparameters in practical applications, the model transferability has become an important indicator for the model quality.

2.3.4 Review of related articles

Numerous studies based on deep learning have sprung up in the field of automatic sleep scoring. Some studies used hand-crafted features as input and constructed deep learning networks as classifiers to classify sleep samples (Chriskos et al. 2020; P. Fonseca et al. 2020). However, most studies took raw PSG signals or spectrograms as input, aiming to build an end-to-end deep learning architecture to avoid the defects of hand-crafted features. For example, Chambon et al. constructed a two-dimensional convolution neural network to perform sleep stage classification from multiple PSG signals (Chambon et al. 2018). A combination of CNN and RNN was employed to automate sleep scoring and to detect sleep apnea and limb movements, reaching accuracy of 87.6%, 88.2%and 84.7%, respectively (Biswal et al. 2018). By applying the combination of CNN and LSTM to long-term scalp EEG recordings (>12 hours), their algorithm achieved a Cohen's kappa of 0.74 on 650 patients with obstructive sleep apnea (Jaoude et al. 2020).

Moreover, studies based on deep learning had introduced some novel classification schemes to exploit the long-term dependency between sleep stages. The "multi-input-one-output" classification scheme was common in mining the dependence of sleep stages (Chambon et al. 2018; Sors et al. 2018). Phan et al. used the CNN architecture to build a "one-input-multi-output" classification scheme, which yielded an accuracy of 83.6% on 200 subjects from Montreal Archive of Sleep Study (Phan et al. 2019a). In order to explore a good tradeoff between information utilization and computational efficiency, a "multi-input-multi-output" classification scheme was proposed, which utilized attention-based bidirectional RNN to balance the weight of epochs (Phan et al. 2019b).

Even though many studies built their models using deep learning network in the field of automatic sleep scoring, few studies tested their model on other unrelated datasets. Zhang et al. did so, where they trained a model on 5213 recordings from the SHHS dataset and then tested model performance on the patients from the study of osteoporotic fractures (SOF), achieving a kappa value of 0.68. And the input channels of these two datasets were matched. In practical applications, the recorded sleep signals might be different due to the specific monitoring devices and experimental motivations. When the model was used in disparate tasks, a task-specific modification was required. That modification was difficult and even inefficient, especially for non-experts facing complex practical conditions. Therefore, it was necessary to explore a model that can accommodate multiple input signals and sample populations.

3 OVERVIEW OF INCLUDED ARTICLES

The research works included in this thesis are published in the attached five articles. The employed PSG recordings are from public sleep datasets, including the CAP dataset provided by PhysioBank (Terzano Giovanni et al. 2002; Goldberger et al. 2000), the Sleep Heart Health Study (SHHS) (Dean et al. 2016), the Sleep-EDF dataset (Kemp et al. 2000; Goldberger et al. 2000) and the ISRUC-Sleep dataset (Khalighi et al. 2016). The four datasets have disparate sample attributes, in which the CAP dataset and the Sleep-EDF dataset consist of healthy adults, the SHHS dataset containing old adults with suspicious heart disease and the ISRUC dataset covers healthy participants and patients with sleep disorders.

In this thesis, model performance is evaluated by the following indices.

Accuracy indicates the fraction of the total number of correct detections in the classification of sleep stages. It is defined as,

$$Acc = \frac{\sum_{i=1}^{c} TP_i}{N} \tag{12}$$

where TP_i denotes true positives in class i, c the total number of sleep stage, and N is the total number of samples.

Cohen's kappa (K) is an agreement measure between the proposed model and a human expert, which takes into account the chances of random agreement.

$$K = \frac{P_o - P_e}{1 - P_e} \tag{13}$$

where $P_o = Acc$, $P_e = \sum_{i=1}^{c}(TP_i + FP_i) \times (TP_i + FN_i)/(N \times N)$, and TP_i, FP_i and FN_i respectively denote true positives, false positives and false negatives in class i.

Before introducing the included articles, we provide an article overview in TABLE 3 to facilitate readers to quickly grasp and understand this dissertation as a whole work. From the TABLE 3, readers can easily capture the connections and differences of five articles in terms of sample information, model input, model architecture, classification accuracy and main contributions. Then, the following sections describe research methods, main results and the author's contribution to the research.

TABLE 3 An overview of included articles

Paper	Dataset	Subject		Input	Model	Acc.	Contributions
		Number	Attribute				
PI	CAP	6	Healthy	Features	ReliefF+RF	86.2%	• Develop an automatic scoring method by fusing multiple PSG signals • Assess the performance of four feature selectors and five classifiers • Evaluate classification performance of eight signal fusions • Pick out the most discriminative features in sleep scoring
PII	SHHS	100	Near-healthy	Features	ReliefF+RF+ HMM	85.8%	✧ Build an automatic sleep scoring toolbox ✧ Evaluate classification performance of 12 signal fusions in distinguishing 2-5 sleep stages ✧ Analyze signals' contribution in automatic sleep scoring
PIII	SHHS	190	Near-healthy Apnea	Features	ReliefF+RF	86.7%	• Improve the proposed toolbox • Analyze the influence of sleep apnea severity on classification accuracy • Assess features' contribution to distinguishing stage pairs
PIV	SHHS; Sleep-EDF	100 19	Near-healthy Healthy	Time series	CNN	85.2% 85%	✧ Develop an end-to-end CNN architecture to automate sleep scoring ✧ Evaluate model performance on two disparate datasets with different input channels
PV	SHHS Sleep-EDF ISRUC	100 19 99	Near-healthy Healthy Patients	Time series	CNN+LSTM	87% 86% 86%	• Construct an end-to-end model using CNN and LSTM • Verify the impact of LSTM units by comparing classification results and visualizing layer outputs • Analyze the effects of electrode positions and signal modality on classification accuracy • Evaluate model performance on different datasets and disease populations • Demonstrate model transferability with or without channel mismatch

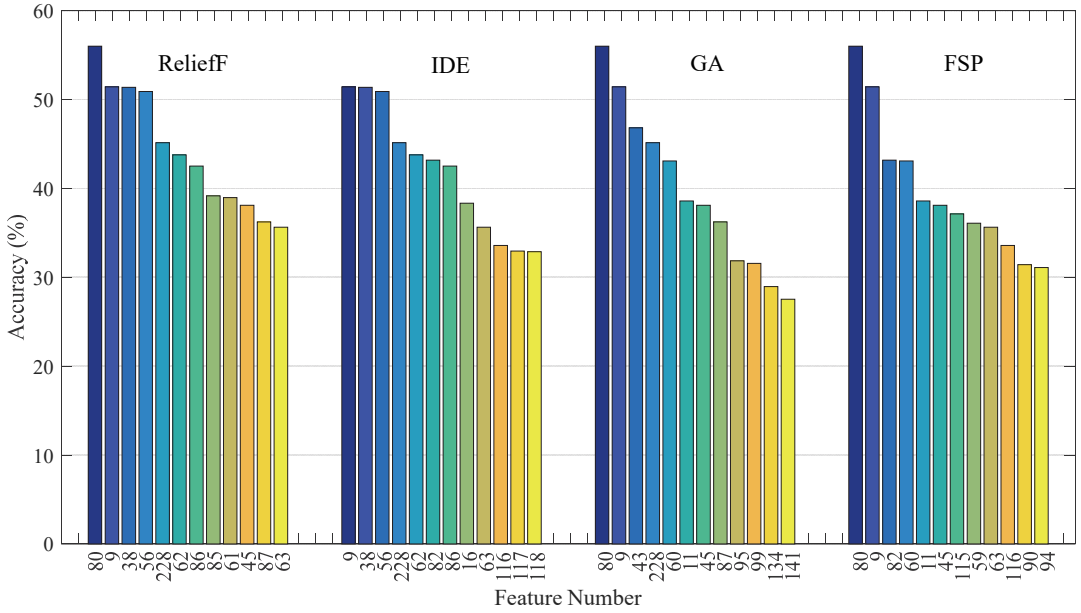
3.1 Study I

Rui Yan, Chi Zhang, Karen Spruyt, Lai Wei, Zhiqiang Wang, Lili Tian, Xueqiao Li, Tapani Ristaniemi, Jihui Zhang, and Fengyu Cong. 2019. "Multi-Modality of Polysomnography Signals' Fusion for Automatic Sleep Scoring." Biomedical Signal Processing and Control 49: 14–23. DOI: 10.1016/j.bspc.2018.10.001

Objective: Adequate and high-quality sleep is vital to our physical and mental health. Conversely, sleep loss or sleep distortion can result in catastrophic consequences, such as cognitive decline, emotional disturbance, metabolic disorders. Therefore, the monitoring and analysis of sleep are crucial in clinical applications. In order to ease tedious manual scoring, automatic sleep scoring becomes a research hotspot due to its advantage of cost-effectiveness and stable scoring performance. Although there are many automatic sleep scoring methods, most studies build their model on certain sleep signals. The specific effects of multi-modality PSG signals on sleep scoring are not univocal. Therefore, this study aims to develop an automatic sleep scoring method by applying multiple PSG signals and further to explore signals' contribution in sleep scoring.

Method: PSG recordings for sleep analysis were provided by the CAP dataset. A total of 6047 samples from 6 healthy subjects were analyzed. For each sample, four modalities of PSG signals, EEG, EOG, EMG and ECG, were employed to automate sleep scoring. To provide a comprehensive description of sleep signals, we extracted 232 features, covering statistical features, frequency features, time-frequency features, nonlinear features and mutual information between two signals. Then, we adopted four feature selection methods and further compared their performance on feature selection. The adopted feature selection methods were ReliefF algorithm, improved distance-based evaluation methods (IDE), genetic algorithms (GA) and forward selection process (FSP). Five commonly used classifiers were applied to distinguish sleep samples, namely k-nearest neighbor classifier (KNN), binary decision tree (BDT), Naïve Bayes (NB), random forest (RF) and support vector machine (SVM). By comparing the performance of different feature selection methods and classifiers on the same dataset, we can find out the most suitable method for automatic sleep scoring, thereby providing a direction for the future research. After determining the suitable classification method for automatic sleep scoring, we further compared eight fusions of four signal modalities to elucidate signals' contribution to sleep scoring. In addition, the most discriminative features of each signal modality were emphasized in the article to clarify the contribution of features.

Main results: In terms of the employed CAP dataset, the best classifier, random forest, achieved the optimal consistency of 86.24% with manual scoring results. The optimal accuracy was reached by fusing multiple features from four signal modalities.



* 1-63: EEG features; 64-87: EOG features; 88-111: EMG features; 112-148: ECG features; 149-187: Coherence; 188-226: Phase angles; 227-232: Mutual information.

FIGURE 6 Selected features of different feature selection methods

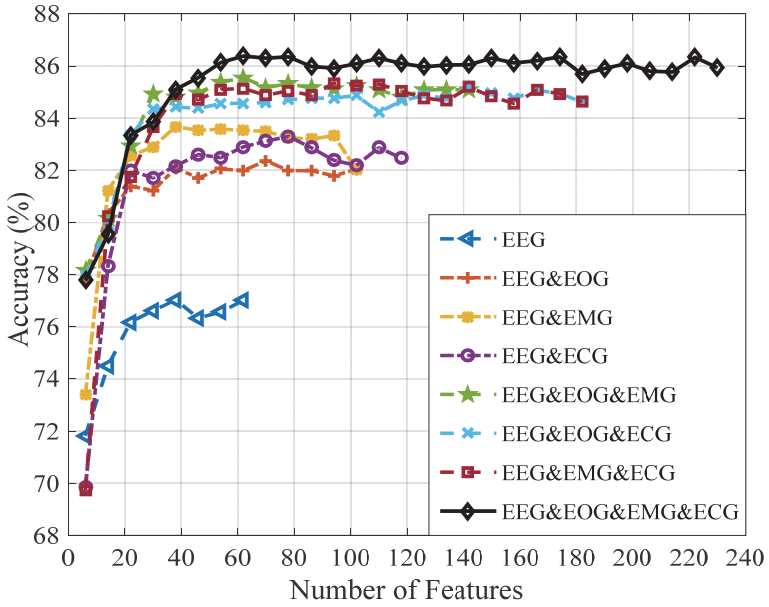


FIGURE 7 Mean accuracy of 10-fold cross-validation from different signals' fusions

From the perspective of feature selection, four feature selection methods (ReliefF, IDE, GA, and FSP) gave different results, which might be attributed to their different evaluation criteria. For example, ReliefF used the distance to measure the similarity of features (Kononenko, Šimec, and Robnik-Šikonja 1997), while GA was a population-based technique (C. M. Fonseca and Fleming 1993). In order to evaluate the effectiveness of features, the selected features were fed into a random forest classifier one by one to evaluate its capability to distinguish sleep stages. FIGURE 6 presented the top twelve outcomes and their

corresponding classification accuracy. As can be seen from FIGURE 6, the EOG features (P: 80, spectral edge) and the EEG features (P: 9, zero crossings; P: 45, relative power spectral of theta wave; P: 63, Spectral entropy) appeared most frequently in feature selection results, indicating their stable performance in distinguishing sleep stages. The comparison results shown in FIGURE 6 demonstrated that ReliefF obtained the best feature set since its elements showed the highest discrimination accuracy than other selectors. The top twelve features in the optimal feature set were respectively EEG features named zero-crossings, spectral edge, relative power spectral of theta, Petrosian fractal dimension, approximate entropy, permutation entropy and spectral entropy, and EOG features named spectral edge, approximate entropy, permutation entropy and spectral entropy, and the mutual information between EEG and submental EMG.

In terms of signal modality, FIGURE 7 displayed the accuracy of different signal fusions. The single EEG signal only achieved a classification accuracy of 77%. Nevertheless, the addition of some features from other signal modalities contributed to the enhancement of classification accuracy. As shown in FIGURE 7, the fusion of two signals modalities significantly improved classification accuracy. Adding the third and the fourth signal types promoted classification accuracy to some extent. The slight improvement, produced by the addition of the fourth signal type, might be attributed to information saturation. However, the definitive conclusion still required further investigation.

Contributions in the article

Rui Yan built the methodology, conducted all experiments, and wrote the original draft.

Chi Zhang, Karen Spruyt, Lili Tian and Xueqiao Li revised the manuscript.

Lai Wei, Zhiqiang Wang, and Jihui Zhang provided clinical suggestions.

Tapani Ristaniemi and Fengyu Cong supervised the whole research work.

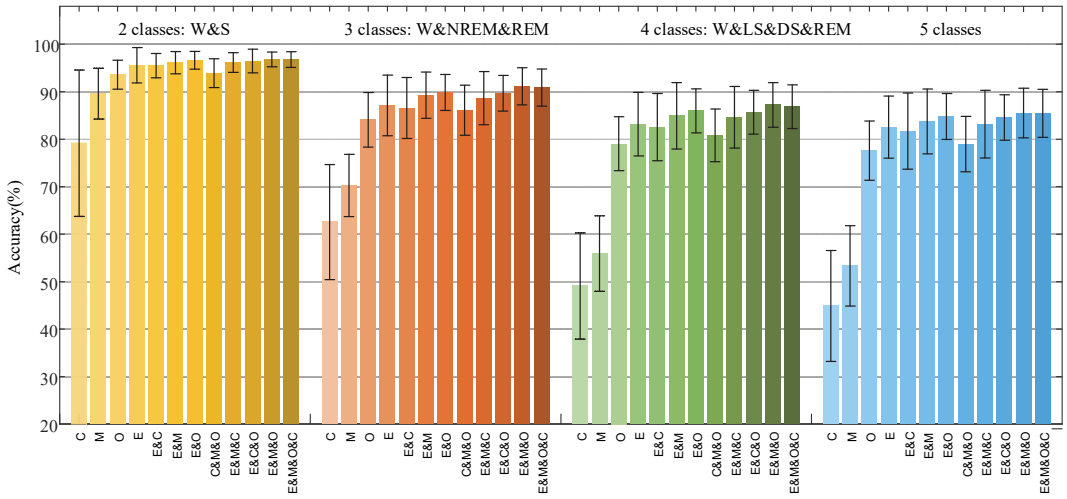
3.2 Study II

Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi and Fengyu Cong. 2019. "An Automatic Sleep Scoring Toolbox: Multi-modality of Polysomnography Signals' Processing." In Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019), pp. 301-309. DOI: 10.5220/0007925503010309

Objective: Sleep scoring is an essential but time-consuming process in any sleep laboratory. Even though many studies so far have been published in the field of automatic sleep scoring, the available software and toolbox are still limited. To speed up the process of sleep scoring without compromising classification accuracy, this paper aims to develop an automatic sleep scoring toolbox with the capability of multi-signal processing.

Method: Applying the graphical user interface (GUI) in MATLAB, we created an interactive interface to automate sleep scoring and to visualize sleep structure and sleep parameters. The proposed interface consisted of a training module to allow users to train their own models, an offline prediction module to predict sleep structure using a predefined model or a user-specified model, an online prediction module to monitor sleep state online, and several parameter panels. The interactive interface allowed users to select available signals and the number of sleep stages. Once the required parameters were set, the toolbox automatically performed the following processes: signal pre-processing, feature extraction, classifier training (or prediction) and result correction. After automatic analysis, the user interface displayed the predicted sleep structure, related sleep parameters and a sleep quality index for reference. To improve classification accuracy, a layer-wise classification strategy was proposed according to signals' characteristics in sleep stages. In the correction step, context information was taken into consideration by applying a hidden Markov model to study transition rules of sleep stages. We employed 100 subjects from the SHHS dataset and applied 10-fold cross-validation to evaluate model performance on 100 subjects.

Main results: By using the fusion of four modalities of PSG signals, the proposed toolbox achieved an average accuracy of 85.76% on 100 subjects, which exceeded the accepted benchmark $Acc = 80\%$ among trained human scorers (Danker-Hopfe et al. 2009). In order to fully explore the performance of the proposed toolbox, we conducted a greedy search for several signal fusions referring to different sleep stages. The results were shown in FIGURE 8, where the columns denoted the mean accuracy of 100 subjects, and the bars represented the standard deviation. FIGURE 8 displayed four scoring configurations in different colours. For each scoring configurations, twelve signal fusions were analysed. The names



*C: single-modality ECG; M: single-modality EMG; O: single-modality EOG; E: single-modality EEG; &: combination signals; 2 classes: wakefulness and sleep (W&S); 3 classes: wakefulness, non-rapid eye movement sleep and rapid eye movement sleep (W&NREM&REM); 4 classes: wakefulness, light sleep (containing N1 and N2), deep sleep (SWS) and rapid eye movement sleep (W&LS&DS &R); 5 classes: W, N1, N2, SWS and R.

FIGURE 8 Classification accuracy for different signal fusions and target classes.

of fused signals were listed along X-axis in FIGURE 8 where signals' names were shortened to their middle letter.

FIGURE 8 depicted the uncertainty or variation of classification accuracy under each condition. Altogether, the richness of signal types contributed to the improvement of accuracy and the reduction of uncertainty, and the required signal types varied with the target number of sleep stages. More specifically, If sleep recordings were classified into two stages, namely wakefulness (W) and sleep (S), all considered signal fusions gave satisfactory results. As the number of sleep stages increased, there was a corresponding increase in required signals. For the common five-stage classification in sleep scoring, at least two signals from the same modality or different modalities could achieve an average accuracy that exceeded the accepted benchmark $Acc = 80\%$.

In terms of signal modality, the signal fusions containing EEG signals produced higher classification accuracy, indicating a crucial role of EEG signals in sleep scoring. Moreover, the discriminative information provided by ECG and EMG signals was inferior to that from EEG and EOG signals. Nevertheless, that conclusion might suffer from suspicion about the number of features. In our experiments, there were two EOG channels, namely EOGR and EOGL, but only one EMG channel. That resulted in 102 EOG features, but only 41 EMG features. Insufficient description of the EMG signal might be partly responsible for the poor classification performance.

Contributions in the article

Rui Yan built the methodology, conducted all experiments, and wrote the original draft.

Fan Li and Xiaoyu Wang provided suggestions for the construction of the proposed toolbox, and tested model performance.

Tapani Ristaniemi and Fengyu Cong supervised the whole research work.

3.3 Study III

Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi and Fengyu Cong. 2020. "Automatic Sleep Scoring Toolbox and Its Application in Sleep Apnea." In: Obaidat M. (eds) E-Business and Telecommunications. ICETE 2019. Communications in Computer and Information Science, vol 1247, pp.256-275. Springer, Cham. DOI: 10.1007/978-3-030-52686-3_11

Objective: In our previous research, we proposed a sleep scoring toolbox that can handle multiple signal modalities and automatically analyze sleep structure. Using 10-fold cross validation, the proposed toolbox achieved an average accuracy of 85.76% on 100 healthy subjects. However, in clinical applications, most sleep studies are conducted on patients with sleep problems. One of the most common sleep disorders in sleep medicine centers is sleep apnea, which is characterized by repetitive cessations of respiratory flow during sleep (Pagel and

Pandi-Perumal 2014). Therefore, the present article aims to extend the toolbox to patients with sleep apnea, and to further analyze the impact of sleep apnea severity on classification accuracy.

Method: The research method was similar to our previous study, except that more non-statistical features were used to effectively track the morphological changes of signals. In addition, in order to speed up model training and reduce memory requirements, we abandoned the previous layer-wise classification strategy, but chose a single random forest classifier to conduct sleep scoring. We further evaluated the effectiveness of features extracted by the proposed toolbox. The random forest classifier can export feature importance, which was achieved by permuting the value of each feature to measure how much the permutation decreased model accuracy. In view of that random forest randomly selected feature subsets for each decision split, so each experiment was repeated 100 times to obtain statistical conclusions. The final importance of each feature was the average value of 100 tests. We further tested model performance on patients with sleep apnea to investigate the influence of sleep-disordered breathing on classification accuracy.

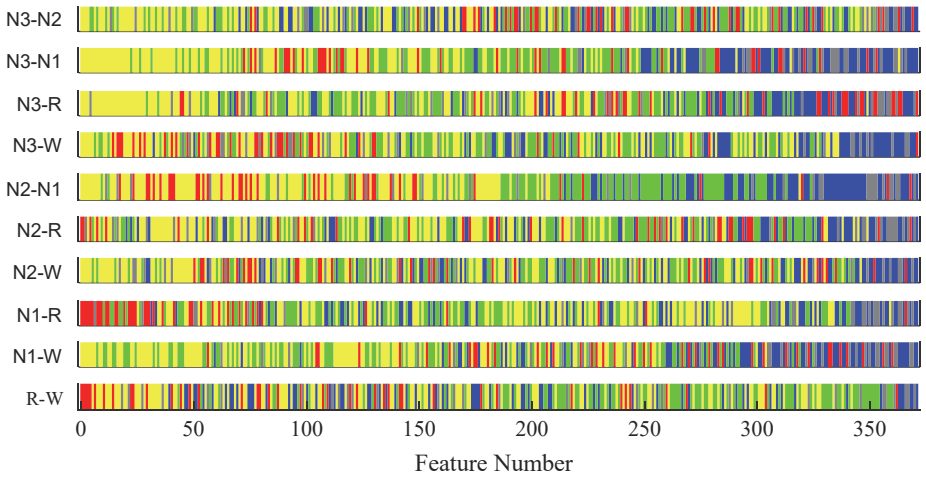
Main results: Our model achieved classification accuracy of 86.7%, 86.1%, 82.1%, and 81.9% on healthy subjects and patients with mild, moderate, and severe sleep-disordered breathing, respectively. It should be noted that the model was trained on 100 subjects with normal sleep breathing, and thus did not contain any

	Sleep stages	N3-N2	N3-N1	N3-R	N3-W	N2-N1	N2-R	N2-W	N1-R	N1-W	R-W
Top 15 features	Top1	E.K	E.lgE	E.lgE	E.ZC	E.Sf	M.lgE	E.P-St	M.lgE	E.P-St	M.lgE
	Top2	E.lgE	E.Sf	E.per75	E.PFD	E.PFD	M.per25	E.spE-St	M.per25	E.spE-St	M.K
	Top3	E.per75	E.per75	E.K	E.ifq	E.P-K	Cf	E.PFD	M.per75	E. θ/β	M.per25
	Top4	O.ifq	E.K	Cf	O.P4	E.ryE	E.spE-sp	E.ZC	M.per5	E. θ (α+β)	M.per75
	Top5	O.mdF	E.ZC	E.PFD	E.Sf	E.spE-K	M.per5	E.ifq	M.SecD	E.ifq	M.per5
	Top6	E.Sf	E.spE-K	E.Sf	O.ryE	E.P4	O.spE-α	O.P4	M.K	O.P4	E.P-St
	Top7	O.K	E.ifq	E.per25	E.θ/β	O.PFD	E.P-sp	E.θ/(α+β)	Ch	E. θ/α	M.CL
	Top8	E.PFD	E.P-K	E.spE-K	E.spE-β	Ch	M.per75	O.ryE	M.CL	O.spE-δ	E.spE-St
	Top9	E.Sw	E.spE-β	E.Rms	M.per25	E.δ/θ	O.P-α	E.P-δ	M.mdPSD	E.P-δ	E. θ/(α+β)
	Top10	Cf	E.PFD	E.per95	E.Sw	E.Rms	M.K	E. δ/α	M.per95	O.SecD	M.SecD
	Top11	O.ZC	E.per25	E.P-β	M.P4	Cf	E.spE-K	E. δ/(α+β)	O.spE-α	O.ryE	E. θ/α
	Top12	E.per25	E.HM	E.spE-β	M.per5	M.P4	E.ryE	Ch	M.ryE	E.ZC	M.per95
	Top13	O.edge50	E.per95	E.spE-St	M.per95	E.per95	Cf	E.spE-δ	M.Sw	E. δ/(α+β)	E.PFD
	Top14	E.SecD	E.per5	E.ryE	E.K	E.lgE	Ch	O.S	O.P-α	E.P-K	E. θ/β
	Top15	E.spE-θ	E.Sw	E.P-δ	E.HM	E.spE-θ	E.PFD	Ch	M.Sk	O.S	Ch

*EEG features (colour: yellow); EOG features (colour: green); EMG features ((colour: red)); ECG features (colour: blue); Mutual information (color: gray).

*-K: K-complex; -sp: sleep spindle; -St: Sawthooth;

FIGURE 9 Top 15 features in distinguishing specific pair of sleep stages



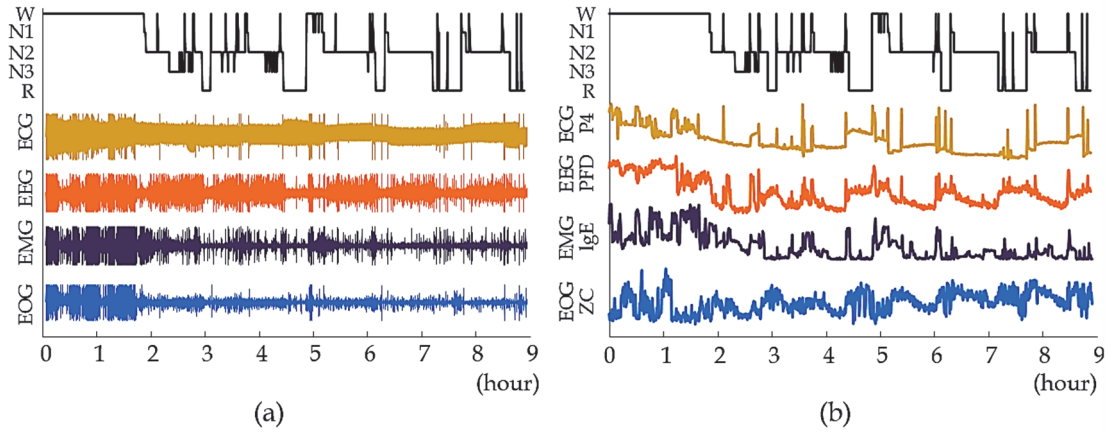
*EEG features (color: yellow); EOG features (color: green); EMG features (color: red); ECG features (color: blue); Mutual information (color: gray). Decreasing importance from left to right.

FIGURE 10 Feature distributions for distinguishing each pair of sleep stages

information of patients with abnormal sleep breathing. Comparing the accuracy of different sample populations, we can get that the model performed slightly better on cases with normal breathing and mild apnea than on those with moderate and severe apnea. Nevertheless, even for patients with severe sleep-disordered breathing, the proposed model provided acceptable per-epoch sensitivity and precision. The stable performance on all cases indicated good model transferability between healthy people and patients with sleep apnea.

FIGURE 9 showed the top 15 features in the discrimination of each pair of sleep stages, where features were arranged in descending order of discriminative ability. FIGURE 9 indicated that the optimal feature subset was a fusion of statistical measures (e.g. Percentiles, Hjorth parameters, Zero-Crossing), spectral measures (e.g. spectral edge, power spectral density), entropy measures (e.g. spectral entropy), fractal measures (e.g. Petrosian fractal dimension) and nonlinear measures (e.g. mean curve length, The 4th Power). Moreover, FIGURE 10 listed the distribution of feature importance when distinguishing each pair of sleep stages. What stood out in FIGURE 10 was that EEG features contributed to the discrimination of most stages, but they were weak in distinguishing stage pairs of N1-R and R-W which might be attributed to morphological similarity of EEG activity in these three sleep stages. EMG features showed good performance on the discrimination of R sleep and wakefulness, partly due to the conspicuous changes of EMG amplitude in stage R and stage W. ECG features exhibited their contribution in distinguishing stage pairs of N3-W, N2-W and R-W, indicating their good performance on the discrimination of wakefulness and sleep.

To verify the validity of discriminative features showed in FIGURE 9, we visualized several features in FIGURE 11 (b). Comparing the feature's value with raw PSG signals, it was easily to get that there was higher correlation between feature values and sleep stages. Moreover, the differences between sleep stages were highlighted at the feature level. This comparison in FIGURE 11 not only



*P4: the 4th power; PFD: Petrosian fractal dimension; lgE: Log energy entropy; ZC: Zero-Crossings.

FIGURE 11 Comparison of raw signals (a) and extracted features (b).

illustrated the effectiveness of discriminative features, but also explained why machine learning methods could achieve accurate and stable results.

Contributions in the article

Rui Yan built the methodology, conducted all experiments, and wrote the original draft.

Fan Li and Xiaoyu Wang provided suggestions for the construction of the proposed toolbox, and tested model performance.

Tapani Ristaniemi and Fengyu Cong supervised the whole research work.

3.4 Study IV

Rui Yan, Fan Li, DongDong Zhou, Tapani Ristaniemi and Fengyu Cong. 2020. " A Deep Learning Model for Automatic Sleep Scoring using Multimodality Time Series." In 28th European Signal Processing Conference (EUSIPCO 2020). 5 pages. IEEE, Amsterdam, Netherlands.

Objective: Automatic sleep scoring is crucial and urgent to help address the increasing unmet need for sleep research. In practical, due to different monitor devices and various experimental motivations, the recorded sleep signals are diverse, which result in channel mismatch in automatic sleep scoring. However, most of the automatic scoring methods published so far are based on human-crafted features or designed for specific datasets. When these models are used in other datasets, task-specific modifications are required, which are difficult and even inefficient in practical applications. Therefore, to facilitate the application of automatic sleep scoring, this paper aims to develop an end-to-end CNN architecture that can handle various numbers of input channels and several signal modalities at the same time without changing any layers or hyper-parameters across tasks.

Method: The proposed architecture took multi-modality PSG signals as input without the calculation of spectrograms or hand-crafted features. In order to accommodate various numbers of input channels, the first convolution layer mapped the input into a virtual space with fixed dimensions, which was achieved by setting its activation function to a time-independent linear operation. It was followed by two integration blocks that included three components: a "squeeze and excitation" block to recalibrate channel-wise feature response, a convolution layer with a smaller filter size to capture local features and a convolution layer with the larger filter size to capture the big contextual features. The learnt representations were finally sent to the decision layer that was a fully-connected layer with 5 units and was activated by a softmax function. The decision layer was responsible for providing evaluations of sleep stages. The network was trained by minimizing categorical cross-entropy. Model performance was evaluated using different signal fusions from two public sleep datasets.

There were 100 recordings from the SHHS dataset, and each recording had 6 available channels for analysis, namely C3-M2, C4-M1, EOGR, EOGL, EMG and ECG. We used 80 recordings for model training and the remaining 20 recordings for testing. For PSG data from the Sleep-EDF dataset, there were only three available channels Fpz-Cz, Pz-Oz and horizontal EOG. A total of 19 recordings were employed, and leave-one-out cross-validation was conducted to evaluate model performance on the Sleep-EDF dataset.

Main results: Experimental results showed that our model achieved the overall accuracy of 85.2% and 85% on the datasets of SHHS and Sleep-EDF, respectively. Even though available channels, amplitude distributions and acquisition environments were significantly different between these two datasets, the proposed architecture obtained comparable results. This indicated that the proposed model could handle various numbers of input channels and several signal modalities from different datasets.

By comparing the proposed model with state-of-the-art methods, we could conclude that the proposed approach not only achieved outstanding classification performance on both datasets, but also exhibited short runtime and low computational cost.

Contributions in the article

Rui Yan built the methodology, conducted all experiments, and wrote the original draft.

Fan Li and Dongdong Zhou provided suggestions for the model structure and revised the manuscript.

Tapani Ristaniemi and Fengyu Cong supervised the whole research work.

3.5 Study V

Rui Yan, Fan Li, DongDong Zhou, Tapani Ristaniemi and Fengyu Cong. 2020. "Automatic Sleep Scoring: A Deep Learning Architecture for Multi-modality Time Series." Manuscript submitted to the Journal of Neuroscience Methods.

Objective: Studies have found that the transitions among sleep stages are not a random process. This finding is consistent with the manual sleep scoring criteria. When sleep experts assign a sleep stage to sleep segments, they also check previous information and posterior information, as shown in TABLE 2. However, conventional classifiers and CNN architectures can only give their decisions based on information from the current stage, but can't remember the context. Recurrent neural networks (RNN), especially the long short-term memory network (LSTM) units, have demonstrated their good performance in capturing temporal correlations of inputs. Therefore, we aim to develop a combination architecture of CNN and LSTM to fully exploit the current and context information from raw PSG recordings, thereby improving the performance of automatic sleep scoring.

Method: The proposed model adopted two-dimensional convolution neural networks to automatically learn features from multi-modality PSG signals, a "squeeze and excitation" block to recalibrate channel-wise features, together with a long short-term memory unit to exploit long-range contextual relation. The learnt features were finally fed to a fully-connected layer activated by a softmax function to generate discrete predictions for each stage. The network was trained by minimizing categorical cross-entropy.

The model performance was tested on three public datasets: the ISRUC dataset, the SHHS dataset and the Sleep-EDF dataset, of which the Sleep-EDF dataset consisted of healthy participants, the SHHS dataset containing near-healthy subjects, and the ISRUC dataset covered healthy participants, patients with sleep disorders and patients under the effect of sleep medication. In addition, the age distributions, available channels, acquisition environments and scoring criteria of the three datasets were also different. For detailed information, please refer to Table 1 in the attached article [PV]. In the process of model evaluation, 5-fold cross validation was conducted on the SHHS dataset and the ISRUC dataset, while a leave-one-out classification strategy was applied on the Sleep-EDF dataset.

Main results: For all tasks with different available channels, our model achieved outstanding performance on diverse participants, even those with complex sleep disturbances (SHHS: Accuracy-0.87, Kappa-0.81; ISRUC: Accuracy-0.86, Kappa-0.82; Sleep-EDF: Accuracy-0.86, Kappa-0.81). The highest classification accuracy was achieved by a fusion of multiple PSG signals.

Comparing the results in article PIV and article PV, we can find that the addition of LSTM units enhanced classification accuracy and kappa value, but did not significantly increase model parameters. Moreover, the proposed model exhibited better performance and low computational cost compared to state-of-the-art methods that used the same dataset.

Given the rich available channels of the ISRUC dataset, we further studied how the electrode location and signal modality affected classification accuracy. FIGURE 12 displayed the mean value and the standard deviation of classification accuracy for each input configuration. In line with our conclusions in previous studies, FIGURE 12 showed that the addition of EEG signals or other PSG modalities were conducive to improving accuracy and reducing uncertainty. In terms of electrode locations, time series from the C4-M1 channel (shorted to C4 in FIGURE 12) produced the best performance with the mean accuracy of 0.78 and the standard deviation of 0.004. Time series from the O2-M1 channel (indicating by O2 in FIGURE 12) performed the worst, which might be attributed to the poor signal quality caused by uncomfortable electrode location.

Model transferability was crucial for practical applications since a model with good transferability would save considerable training costs, both in training time and in training data. Therefore, we tested model transferability among three datasets. In terms of channel-matched cases, after the model was trained on the SHHS dataset, the trained model was directly tested on the ISRUC dataset. Experimental results showed that the direct prediction achieved moderate classification accuracy on the test dataset (Accuracy-0.73, Kappa-0.64), while fine-tuning the trained model with 20% of test data could significantly improve classification accuracy (Accuracy-0.84, Kappa-0.79). In the case with channel mismatch, the fine-tuning strategy was necessary and resulted in a faster and smoother convergence curve compared to the model trained from scratch. In addition, classification performance improved by 1.6% on accuracy and 2.7% on kappa using the fine-tuning strategy.

To summarize, the proposed architecture could handle various numbers of input channels and several signal modalities without changing model architecture or hyperparameters across tasks. Model generalization and model

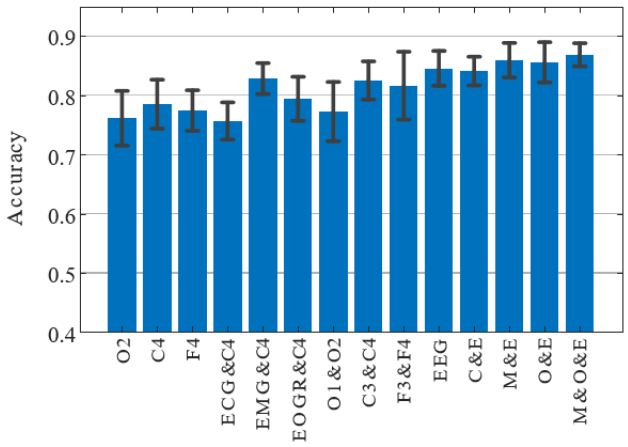


FIGURE 12 Classification accuracy for different signal fusions

transferability were evaluated among different datasets and disease populations. Due to the demonstrated versatility, the proposed method could be integrated with diverse polysomnography systems, thereby facilitating sleep monitoring in clinical or routine care.

Contributions in the article

Rui Yan built the methodology, conducted all experiments, and wrote the original draft.

Fan Li and Dongdong Zhou provided suggestions for the model structure and revised the manuscript.

Tapani Ristaniemi and Fengyu Cong supervised the whole research work.

4 DISCUSSION

This research work contributes to the development of automatic sleep scoring. Sleep is a complex physiological process affected by biological and environmental factors, which makes automatic sleep scoring a challenge. This thesis approaches the challenge from three aspects. The first is to explore an optimal signal fusion by investigating how PSG signals affect the performance of automatic sleep scoring. The second is to explore the most effective features in distinguishing sleep stages, thereby revealing the underlying mechanism of automatic sleep scoring. The third is to develop automatic sleep scoring models with good transferability among different datasets and sample populations, thereby facilitating clinical applications. Thesis contributions are summarized below, followed by research limitations and future work prospects.

4.1 Contributions

The thesis provides insights into the reasoning behind automatic sleep scoring and further develops automatic sleep scoring models to facilitate practical applications. The detailed contributions are listed as follows.

To reveal the underlying mechanism of automatic sleep scoring, first of all, we systematically explore the effect of PSG signals on scoring performance. Previous studies have shown that abundant signal modalities are conducive to the improvement of scoring accuracy, but up to a certain point (Chambon et al. 2018; Krakovská and Mezeiová 2011; Šušmáková and Krakovská 2008). We share this view and add that the electrode position also impacts scoring accuracy. In terms of EEG electrodes, the central electrodes (C3 and C4) perform best in sleep scoring, followed by frontal electrodes (F3 and F4), and occipital electrodes (O1 and O2) are the worst. The divergent performance may be associated with the excitation position of EEG patterns and the comfort of electrode locations. For signal modality, EEG signals help to recognize most stages. The EMG signals contribute to the discrimination of stage R. The addition of EOG and ECG signals

show an auxiliary effect on scoring, especially for distinguishing N1 and R stages that are difficult to classify using only EEG signals. Moreover, ECG signals show good performance to distinguish wakefulness from sleep (especially, stage N2, N3 and R). In summary, diverse signal modalities complement and reinforce each other, and thus jointly promote classification performance.

Evidence from sleep physiology further supports the above conclusion. Regulation of the sleep-waking cycle is complex and involves diverse brain circuits (Pace-Schott and Hobson 2002). The collective triggering of brain circuits or regions produces rhythm waves (EEG activity). Due to functional diversity of brain area, the rhythm waves during sleep appear in different brain regions, such as alpha rhythm mostly appears in the occipital region and theta activity is often maximal over the central and temporal areas (Berry et al. 2016). Therefore, the electrode position is associated with the capture of rhythm waves, which in turn affects the scoring accuracy. In addition, the release of molecules is accompanied by the regulation of the sleep-waking cycle, leading to a series of changes in the muscles, eyes and cardiovascular system. Specifically, with the deepening of sleep, eye movements become gradually rare, muscles relax, and cardiovascular and respiratory behaviors slow down (Trinder et al. 2001). These provide theoretical foundations for multimodality signals to act together in sleep scoring.

The exploration of a mass of features further demonstrates the above conclusion, since the elements in the optimal feature set come from multiple signal modalities. In terms of feature type, the top 15 discriminative features are a fusion of statistical measures (e.g. Percentiles, Hjorth parameters, Zero-Crossing), spectral measures (e.g. spectral edge, power spectral density) and nonlinear measures (e.g. spectral entropy, Petrosian fractal dimension, mean curve length). By comparing the feature values in different sleep stages, we found that the most discriminative features have the following quality: capturing sudden changes of sleep stages, being robust to noise, and being sensitive to a certain rhythm wave. Nevertheless, a single feature is hard to achieve good classification accuracy, as shown in FIGURE 9 that diverse features exhibit their contribution in distinguishing sleep stages. Therefore, a compromise solution is to build a collection that covers multiple feature types. To the best of our knowledge, this thesis analyzes for the first time the features' contribution in distinguishing stage pairs. This exploratory work helps us understand which signal and which feature contribute to the decision, thereby revealing the veil of automatic sleep scoring. This experimental conclusion helps us understand automatic scoring models and control them.

This thesis also evaluates the influence of temporal context on classification performance by applying HMM for post-classification processing or using LSTM units in classification. We found that the addition of temporal context helps to the improvement of model performance, but the accuracy improvement of some stage comes at the expense of other stages. This conclusion can be drawn by comparing Table 4 in Paper IV and Paper V. The comparison shows that the addition of LSTM unit improves the accuracy of N1, N2 and R stages, while slightly decreases the accuracy of W and N3 stages. Some studies have claimed

that contextual input does not always lead to performance improvement, but it can improve the recognition of transitional stages (Chambon et al. 2018; Phan et al. 2019b). In addition, the length of temporal context is also an issue that needs to be quantified because too long temporal context might result in modelling ambiguity and a penalization of short stage changes (Chambon et al. 2018). Therefore, when HMM is used as a post-classification processing step, it is necessary to carefully select the number of hidden states, which is set to three in the attached Paper [III] (Esmael et al. 2012; Li et al. 2018). Besides, the HMM is a supplement to the classification model, but it cannot change model property. Instead, the LSTM unit is integrated into the classifier and affects model performance by adjusting weights and biases. Moreover, the LSTM unit implicitly remembers useful context information without hand-crafted adjustment. However, the visualization of LSTM units needs to be further explored.

Based on the above research, we have developed two sets of automatic sleep scoring tools, one based on conventional classification methods and the other based on deep learning networks. These two models have the capability of multiple signal processing and can adapt to different input channels without changing model parameters. The advantage makes the model avoid cumbersome task-oriented adjustments to model architecture and parameters, thereby facilitating practical applications. In addition, the linked analysis of multiple signals is conducive to the exploitation of cross-modality information, so it has the potential to detect information that is not discoverable by any single modality. The model has achieved promising performance on subjects with different characteristics, not only on the healthy participants but even subjects with complex sleep disturbances. Due to the demonstrated availability and versatility, the proposed method can be integrated with diverse polysomnography systems, thereby facilitating sleep monitoring in clinical or routine care.

4.2 Limitations and future research

Even though our results are encouraging, our study still suffers from several limitations. One of them is that model performance may be affected by data attributes, but that is a common defect of machine learning. Since the proposed models learn from the training data, they might not perform well when severe difference exist in the test data and the training data. For example, a scoring model trained on healthy participants may not perform well in the prediction of patients' sleep stages. In that case, fine-tuning the model with a small amount of test data can help to improve model performance. Besides, applying a huge training dataset can also compensate for this defect since a vast training data theoretically benefits to the improvement of model generalization. Therefore, a large and high-quality dataset is needed to further improve the generalizability of proposed models.

Secondly, the work involved in this thesis mainly focuses on healthy adults, although model performance has been verified on healthy participants and patients with sleep disorders. Given the complex and diverse clinical symptoms of suspected patients, it is necessary to test model generalizability on a large population with diverse sleep problems before taking models into clinical applications.

Third, this research only contains sleep scoring. However, when sleep experts perform manual sleep scoring, they will score sleep stages and sleep events together, such as marking abnormal breath, abnormal muscle activity, and body movements. The scoring of sleep events plays an important role in the diagnosis of sleep diseases. For example, the number of apnea and hypopnea during sleep is required to diagnose sleep apnea syndrome, the number of periodic leg movements during sleep to diagnose periodic leg movement syndrome, and abnormal muscle activity in REM sleep is typical symptoms in the diagnosis of rapid eye movement behavior disorder. Therefore, the scoring of sleep events needs to be addressed if it is expected that automatic sleep scoring completely replaces manual scoring. Fortunately, deep learning networks based on big data provide an opportunity to score sleep stages and sleep events together. That may be an interesting topic for future research.

The fourth is the visualization of the deep learning model. We have explored discriminative features for conventional classification methods and revealed underlying reasoning behind automatic sleep scoring. However, the deep learning model proposed in this thesis is still a black box for us, even though its good performance has been demonstrated and its layer outputs have been visualized. The visualization of deep learning model needs further exploration to verify which part of the input contributes to the final decisions and whether the features extracted automatically by deep learning model are consistent with hand-crafted features or manual scoring standards. This research is important to reveal the physiological mechanism of automatic scoring model based on deep learning networks, and it is also beneficial to understand and control deep learning models.

5 CONCLUSION

This thesis has explored the reasoning behind automatic sleep scoring and further developed two sets of automatic sleep scoring tools with the capability of multiple signal processing. By exploring diverse fusions of PSG signals, in line with other scholars, I advocate the benefits of joint analysis of multiple PSG signals to improve classification accuracy, especially in complex practical applications. Different signal modalities complement each other, which is beneficial for mining interactive information that cannot be discovered in any single signal modality.

Another part of the thesis, departing from a direct application of feature extraction for sleep scoring, attempts to find an end-to-end solution for automatic sleep scoring. Our research indicates that the input of raw PSG signals results in a robust and versatile model, thereby facilitating model applications under complex clinical conditions. The versatile models proposed in this thesis are capable to accommodate multi-channel and multi-modal signals, which might inspire the construction of cross-models in the field of automatic sleep scoring.

This work contributes to addressing the growing unmet needs for sleep research by supplementing and extending previous studies on automatic sleep scoring. From a broader perspective, the proposed automatic sleep scoring methods facilitate sleep monitoring in clinical or routine care. The increasing monitoring of sleep health would promote personal wellbeing and even create economic and social benefits.

YHTEENVETO (SUMMARY IN FINNISH)

Tässä väitöskirjassa on tutkittu automaattisten unen pisteytysmenetelmien taustalla olevaa mekanismia ja kehitetty edelleen kaksi automaattisten unen pisteytystyökalujen sarjaa, jotka kykenevät moninkertaiseen signaalinkäsittelyyn. Tutkimalla PSG-signaalien erilaisia fuusioita muiden tutkijoiden kanssa kannatan useiden PSG-signaalien yhteisen analyysin etuja luokitustarkkuuden parantamiseksi, erityisesti monimutkaisissa käytännön sovelluksissa. Eri signaalimoodit täydentävät toisiaan, mikä on hyödyllistä vuorovaikutteisen tiedon louhimiselle, jota ei voida löytää missään yksittäisessä signaalimoodaalissa.

Opinnäytetyön toinen osa, joka poikkeaa ominaisuuksien poiminnan suorasta soveltamisesta unen pisteytykseen, yrittää löytää päästä-päähän ratkaisun automaattiseen unen pisteytykseen. Tutkimus osoittaa, että raakojen PSG-signaalien syöttö johtaa vankkaan ja monipuoliseen malliin, mikä helpottaa mallisovelluksia monimutkaisissa kliinisissä olosuhteissa. Väitöskirjassa ehdotetut monipuoliset mallit pystyvät vastaanottamaan monikanava- ja multimodaaliset signaalit, mikä saattaa inspiroida ristimallien rakentamista automaattisen unen pisteytyksen alalla.

Tämä väitöskirja auttaa vastaamaan kasvaviin unentutkimustarpeisiin täydentämällä ja laajentamalla aiempia automaattisia unipisteytyksiä koskevia tutkimuksia. Laajemmasta näkökulmasta ehdotetut automaattiset unen pisteytysmenetelmät helpottavat unen seurantaa kliinisessä tai rutiinihoidossa. Unen terveyden lisääntyvä seuranta edistäisi henkilökohtaista hyvinvointia ja jopa luo taloudellisia ja sosiaalisia etuja.

REFERENCES

Acharya, Rajendra U., Oliver Faust, N. Kannathal, Tjileng Chua, and Swamy Laxminarayan. 2005. "Non-Linear Analysis of EEG Signals at Various Sleep Stages." *Computer Methods and Programs in Biomedicine* 80 (1): 37–45. https://doi.org/10.1016/j.cmpb.2005.06.011.

Al-Hussaini, Irfan, Cao Xiao, M. Brandon Westover, and Jimeng Sun. 2019. "SLEEPER: Interpretable Sleep Staging via Prototypes from Expert Rules." *arXiv Preprint arXiv:1910.06100*, 1–18.

Alpaydin, Ethem. 2020. *Introduction to Machine Learning*. MIT press.

Altevogt, Bruce M, and Harvey R Colten. 2006. *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. National Academies Press.

Balas, Valentina Emilia, Raghvendra Kumar, and Rajshree Srivastava. 2020. *Recent Trends and Advances in Artificial Intelligence and Internet of Things*. Springer.

Bergstra, James, Daniel Yamins, and David Cox. 2013. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures." In *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP*, 28:115–23.

Bergstra, James, and Yoshua Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research* 13 (1): 281–305. https://doi.org/10.5555/2188385.2188395.

Berry, Richard B., Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Robin M. Lloyd, Carole L. Marcus, and Bradley V. Vaughn. 2016. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. *Darien, Illinois:American Academy of Sleep Medicine*.

Biswal, Siddharth, Haoqi Sun, Balaji Goparaju, M. Brandon Westover, Jimeng Sun, and Matt T. Bianchi. 2018. "Expert-Level Sleep Scoring with Deep Neural Networks." *Journal of the American Medical Informatics Association* 25 (12): 1643–50. https://doi.org/10.1093/jamia/ocy131.

Boostani, Reza, Foroozan Karimzadeh, and Mohammad Nami. 2017. "A Comparative Review on Sleep Stage Classification Methods in Patients and Healthy Individuals." *Computer Methods and Programs in Biomedicine* 140: 77–91. https://doi.org/10.1016/j.cmpb.2016.12.004.

Chambon, Stanislas, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. 2018. "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26 (4): 758–69. https://doi.org/10.1109/TNSRE.2018.2813138.

Chriskos, Panteleimon, Christos A. Frantzidis, Polyxeni T. Gkivogkli, Panagiotis D. Bamidis, and Chrysoula Kourtidou-Papadeli. 2020. "Automatic Sleep Staging Employing Convolutional Neural Networks and Cortical Connectivity Images." *IEEE Transactions on Neural Networks and Learning Systems* 31 (1): 113–23. https://doi.org/10.1109/TNNLS.2019.2899781.

Cox, Rebecca C., and Bunmi O. Olatunji. 2016. "A Systematic Review of Sleep Disturbance in Anxiety and Related Disorders." *Journal of Anxiety Disorders* 37: 104–29. https://doi.org/10.1016/j.janxdis.2015.12.001.

Danker-Hopfe, Heidi, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, et al. 2009. "Interrater Reliability for Sleep Scoring according to the Rechtschaffen & Kales and the New AASM Standard." *Journal of Sleep Research* 18 (1): 74–84. https://doi.org/10.1111/j.1365-2869.2008.00700.x.

Dean, Dennis A., Ary L. Goldberger, Remo Mueller, Matthew Kim, Michael Rueschman, Daniel Mobley, Satya S. Sahoo, et al. 2016. "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource." *Sleep* 39 (5): 1151–64. https://doi.org/10.5665/sleep.5774.

Dimitriadis, Stavros I, Christos Salis, and David Linden. 2018. "A Novel, Fast and Efficient Single-Sensor Automatic Sleep-Stage Classification Based on Complementary Cross-Frequency Coupling Estimates." *Clinical Neurophysiology* 129 (4): 815–28. https://doi.org/10.1101/160655.

Dinges, David F., Frances Pack, Katherine Williams, Kelly A. Gillen, John W. Powell, Geoffrey E. Ott, Caitlin Aptowicz, and Allan I. Pack. 1997. "Cumulative Sleepiness, Mood Disturbance, and Psychomotor Vigilance Performance Decrements during a Week of Sleep Restricted to 4-5 Hours per Night." *Sleep* 20 (4): 267–77. https://doi.org/10.1093/sleep/20.4.267.

Dong, Hao, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. 2018. "Mixed Neural Network Approach for Temporal Sleep Stage Classification." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26 (2): 324–33. https://doi.org/10.1109/TNSRE.2017.2733220.

Ebrahimi, Farideh, Seyed Kamaledin Setarehdan, Jose Ayala-Moyeda, and Homer Nazeran. 2013. "Automatic Sleep Staging Using Empirical Mode Decomposition, Discrete Wavelet Transform, Time-Domain, and Nonlinear Dynamics Features of Heart Rate Variability Signals." *Computer Methods and Programs in Biomedicine* 112 (1): 47–57. https://doi.org/10.1016/j.cmpb.2013.06.007.

El-Manzalawy, Yasser, Orfeu Buxton, and Vasant Honavar. 2017. "Sleep/wake State Prediction and Sleep Parameter Estimation Using Unsupervised Classification via Clustering." In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 718–23. IEEE. https://doi.org/10.1109/BIBM.2017.8217742.

Esmael, Bilal, Arghad Arnaout, Rudolf K. Fruhwirth, and Gerhard Thonhauser. 2012. "Improving Time Series Classification Using Hidden Markov Models." In *2012 12th International Conference on Hybrid Intelligent Systems (HIS)*, 502–7. https://doi.org/10.1109/HIS.2012.6421385.

Fernández-Varela, Isaac, Elena Hernández-Pereira, and Vicente Moret-Bonillo. 2018. "A Convolutional Network for the Classification of Sleep Stages." *Multidisciplinary Digital Publishing Institute Proceedings* 2 (18): 1174. https://doi.org/10.3390/proceedings2181174.

Fonseca, Carlos M, and Peter J Fleming. 1993. "Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization." In *Proceedings of the Fifth International Conference (S.Forrest,ed.),San Mateo, CA: Morgan Kaufmann*, 93:416–23.

Fonseca, Pedro, Merel M van Gilst, Mustafa Radha, Marco Ross, Arnaud Moreau, Andreas Cerny, Peter Anderer, Xi Long, Johannes P van Dijk, and Sebastiaan Overeem. 2020. "Automatic Sleep Staging Using Heart Rate Variability, Body Movements, and Recurrent Neural Networks in a Sleep Disordered Population." *Sleep*, 1–10. https://doi.org/10.1093/sleep/zsaa048.

Fonseca, Pedro, Xi Long, Mustafa Radha, Reinder Haakma, Ronald M Aarts, and Jérôme Rolink. 2015. "Sleep Stage Classification with ECG and Respiratory Effort." *Physiological Measurement* 36: 2027–40. https://doi.org/10.1088/0967-3334/36/10/2027.

Gharbali, Ali Abdollahi, Shirin Najdi, and José Manuel Fonseca. 2018. "Investigating the Contribution of Distance-Based Features to Automatic Sleep Stage Classification." *Computers in Biology and Medicine* 96: 8–23. https://doi.org/10.1016/j.compbiomed.2018.03.001.

Goldberger, Ary L., Luis AN Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Golmohammadi, Meysam, Amir Hossein Harati Nejad Torbati, Silvia Lopez De Diego, Iyad Obeid, and Joseph Picone. 2019. "Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures." *Frontiers in Human Neuroscience* 13: 76. https://doi.org/10.3389/fnhum.2019.00076.

Hassan, Ahnaf Rashik, and Mohammed Imamul Hassan Bhuiyan. 2016. "A Decision Support System for Automatic Sleep Staging from EEG Signals Using Tunable Q-Factor Wavelet Transform and Spectral Features." *Journal of Neuroscience Methods* 271: 107–18. https://doi.org/10.1016/j.jneumeth.2016.07.012.

Hassan, Ahnaf Rashik, and Mohammed Imamul Hassan Bhuiyan. 2016. "Automatic Sleep Scoring Using Statistical Features in the EMD Domain and Ensemble Methods." *Biocybernetics and Biomedical Engineering* 36 (1): 248–55. https://doi.org/10.1016/j.bbe.2015.11.001.

Hirshkowitz, Max, Kaitlyn Whiton, Steven M. Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, et al. 2015. "National Sleep Foundation's Sleep Time Duration Recommendations: Methodology and Results Summary." *Sleep Health* 1 (1): 40–43. https://doi.org/10.1016/j.sleh.2014.12.010.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

Huang, Wu, Bing Guo, Yan Shen, Xiangdong Tang, Tao Zhang, Dan Li, and Zhonghui Jiang. 2020. "Sleep Staging Algorithm Based on Multichannel Data Adding and Multifeature Screening." *Computer Methods and Programs in Biomedicine* 187: 105253. https://doi.org/10.1016/j.cmpb.2019.105253.

Iranzo, Alex, José Luis Molinuevo, Joan Santamaría, Mónica Serradell, María José Martí, Francesc Valldeoriola, and Eduard Tolosa. 2006. "Rapid-Eye-Movement Sleep Behaviour Disorder as an Early Marker for a Neurodegenerative Disorder: A Descriptive Study." *The Lancet Neurology* 5 (7): 572–77. https://doi.org/10.1016/S1474-4422(06)70476-8.

Jaderberg, Max, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, and Karen Simonyan. 2017. "Population Based Training of Neural Networks." *arXiv Preprint arXiv:1711.09846*.

Jan, James E., Russ J. Reiter, Martin C.O. Bax, Urs Ribary, Roger D. Freeman, and Michael B. Wasdell. 2010. "Long-Term Sleep Disturbances in Children: A Cause of Neuronal Loss." *European Journal of Paediatric Neurology* 14 (5): 380–90. https://doi.org/10.1016/j.ejpn.2010.05.001.

Jaoude, Maurice Abou, Haoqi Sun, Kyle R. Pellerin, Milena Pavlova, Rani A. Sarkis, Sydney S. Cash, M. Brandon Westover, and Alice D. Lam. 2020. "Expert-Level Automated Sleep Staging of Long-Term Scalp EEG Recordings Using Deep Learning." *Sleep*. https://doi.org/10.1093/sleep/zsaa112.

Jiang, Dihong, Ya-nan Lu, Yu Ma, and Yuanyuan Wang. 2019. "Robust Sleep Stage Classification with Single-Channel EEG Signals Using Multimodal Decomposition and HMM-Based Refinement." *Expert Systems with Applications* 121: 188–203. https://doi.org/10.1016/j.eswa.2018.12.023.

Kemp, B., Aeilko H. Zwinderman, Bert Tuk, Hilbert A.C. Kamphuisen, and Josefien J.L. Oberye. 2000. "Analysis of a Sleep-Dependent Neuronal Feedback Loop: The Slow-Wave Microcontinuity of the EEG." *IEEE Transactions on Biomedical Engineering* 47 (9): 1185–94. https://doi.org/10.1109/10.867928.

Khalighi, Sirvan, Teresa Sousa, Gabriel Pires, and Urbano Nunes. 2013. "Automatic Sleep Staging: A Computer Assisted Approach for Optimal Combination of Features and Polysomnographic Channels." *Expert Systems with Applications* 40 (17): 7046–59. https://doi.org/10.1016/j.eswa.2013.06.023.

Khalighi, Sirvan, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. 2016. "ISRUC-Sleep: A Comprehensive Public Dataset for Sleep Researchers." *Computer Methods and Programs in Biomedicine* 124: 180–92. https://doi.org/10.1016/j.cmpb.2015.10.013.

Klok, Aske B, Joakim Edin, Matteo Cesari, Alexander Neergaard Olesen, Poul Jennum, and Helge B D Sorensen. 2018. "A New Fully Automated Random-Forest Algorithm for Sleep Staging." In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4920–23. https://doi.org/10.1109/EMBC.2018.8513413.

Kononenko, Igor, Edvard Šimec, and Marko Robnik-Šikonja. 1997. "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF." *Applied Intelligence* 7 (1): 39–55. https://doi.org/10.1023/A:1008280620621.

Krakovská, Anna, and Kristína Mezeiová. 2011. "Automatic Sleep Scoring: A Search for an Optimal Combination of Measures." *Artificial Intelligence in Medicine* 53 (1): 25–33. https://doi.org/10.1016/j.artmed.2011.06.004.

Lahat, Dana, Tülay Adali, and Christian Jutten. 2015. "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects." *Proceedings of the IEEE* 103 (9): 1449–77. https://doi.org/10.1109/JPROC.2015.2460697.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Li, Xiaojin, Licong Cui, Shiqiang Tao, Jing Chen, Xiang Zhang, and Guoqiang Zhang. 2018. "Hyclasss: A Hybrid Classifier for Automatic Sleep Stage Scoring." *IEEE Journal of Biomedical and Health Informatics* 22 (2): 375–85. https://doi.org/10.1109/JBHI.2017.2668993.

Liang, Sheng-Fu, Chih-En Kuo, Fu-Zen Shaw, Ying-Huang Chen, Chia-Hu Hsu, and Jyun-Yu Chen. 2016. "Combination of Expert Knowledge and a Genetic Fuzzy Inference System for Automatic Sleep Staging." *IEEE Transactions on Biomedical Engineering* 63 (10): 2108–18. https://doi.org/10.1109/TBME.2015.2510365.

Liang, Sheng-Fu, Chin-En Kuo, Yu-Han Hu, and Yu-Shian Cheng. 2012. "A Rule-Based Automatic Sleep Staging Method." *Journal of Neuroscience Methods* 205 (1): 169–76. https://doi.org/10.1016/j.jneumeth.2011.12.022.

Liang, Sheng-Fu, Chin-En Kuo, Yu-Han Hu, Yu-Hsiang Pan, and Yung-Hung Wang. 2012. "Automatic Stage Scoring of Single-Channel Sleep EEG by Using Multiscale Entropy and Autoregressive Models." *IEEE Transactions on Instrumentation and Measurement* 61 (6): 1649–57. https://doi.org/10.1109/TIM.2012.2187242.

Lim, Julian, and David F. Dinges. 2008. "Sleep Deprivation and Vigilant Attention." *Annals of the New York Academy of Sciences* 1129 (1): 305–22. https://doi.org/10.1196/annals.1417.002.

Liu, Xiao-Feng, and Wang Yue. 2009. "Fine-Grained Permutation Entropy as a Measure of Natural Complexity for Time Series." *Chinese Physics B* 18 (7): 2690–95. https://doi.org/10.1088/1674-1056/18/7/011.

Malafeev, Alexander, Dmitry Laptev, Stefan Bauer, Ximena Omlin, Aleksandra Wierzbicka, Adam Wichniak, Wojciech Jernajczyk, Robert Riener, Joachim Buhmann, and Peter Achermann. 2018. "Automatic Human Sleep Stage Scoring Using Deep Neural Networks." *Frontiers in Neuroscience* 12: 781. https://doi.org/10.3389/fnins.2018.00781.

Miikkulainen, Risto, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, et al. 2019. "Evolving Deep Neural Networks." In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, 293–312. Elsevier. https://doi.org/10.1016/B978-0-12-815480-9.00015-3.

Molina-Picó, Antonio, David Cuesta-Frau, Mateo Aboy, Cristina Crespo, Pau Miró-Martínez, and Sandra Oltra-Crespo. 2011. "Comparative Study of Approximate Entropy and Sample Entropy Robustness to Spikes." *Artificial Intelligence in Medicine* 53 (2): 97–106. https://doi.org/10.1016/j.artmed.2011.06.007.

Mousavi, Z., T. Yousefi Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi. 2019. "Deep Convolutional Neural Network for Classification of Sleep Stages from Single-Channel EEG Signals." *Journal of Neuroscience Methods* 324: 108312. https://doi.org/10.1016/j.jneumeth.2019.108312.

Nair, Vinod, and Geoffrey E. Hinton. 2010. "Rectified Linear Units Improve Restricted Boltzmann Machines." In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, 807–14. https://doi.org/10.5555/3104322.3104425.

Pace-Schott, Edward F, and J Allan Hobson. 2002. "The Neurobiology of Sleep: Genetics, Cellular Physiology and Subcortical Networks." *Nature Reviews Neuroscience* 3: 591–605. https://doi.org/10.1038/nrn895.

Pagel, James F, and Seithikurippu R Pandi-Perumal. 2014. *Primary Care Sleep Medicine: A Practical Guide*. Springer.

Park, Heejung, Kim M. Tsai, Ronald E. Dahl, Michael R. Irwin, Heather McCreath, Teresa E. Seeman, and Andrew J. Fuligni. 2016. "Sleep and Inflammation during Adolescence." *Psychosomatic Medicine* 78 (6): 677–85. https://doi.org/10.1097/PSY.0000000000000340.

Patanaik, Amiya, Ju Lynn Ong, Joshua J Gooley, Sonia Ancoli-Israel, and Michael W L Chee. 2018. "An End-to-End Framework for Real-Time Automatic Sleep Stage Classification." *Sleep* 41 (5): zsy041. https://doi.org/10.1093/sleep/zsy041.

Peng, Hanchuan, Fuhui Long, and Chris Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1226–38. https://doi.org/10.1109/TPAMI.2005.159.

Phan, Huy, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. 2019a. "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification." *IEEE Transactions on Biomedical Engineering* 66 (5): 1285–96. https://doi.org/10.1109/TBME.2018.2872652.

———. 2019b. "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27 (3): 400–410. https://doi.org/10.1109/TNSRE.2019.2896659.

Porkka-Heiskanen, Tarja, and Anna V. Kalinchuk. 2011. "Adenosine, Energy Metabolism and Sleep Homeostasis." *Sleep Medicine Reviews* 15 (2): 123–35. https://doi.org/10.1016/j.smrv.2010.06.005.

Radüntz, Thea, Jon Scouten, Olaf Hochmuth, and Beate Meffert. 2015. "EEG Artifact Elimination by Extraction of ICA-Component Features Using Image Processing Algorithms." *Journal of Neuroscience Methods* 243: 84–93. https://doi.org/10.1016/j.jneumeth.2015.01.030.

Rechtschaffen, Allan, and Anthony Kales. 1968. "A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects." *Washington DC: US National Institute of Health Publication*.

Rodríguez-Sotelo, Jose Luis, Alejandro Osorio-Forero, Alejandro Jiménez-Rodríguez, David Cuesta-Frau, Eva Cirugeda-Roldán, and Diego Peluffo. 2014. "Automatic Sleep Stages Classification Using EEG Entropy Features and Unsupervised Pattern Analysis Techniques." *Entropy* 16 (12): 6573–89. https://doi.org/10.3390/e16126573.

Seifpour, Saman, Hamid Niknazar, Mohammad Mikaeili, and Ali Motie Nasrabadi. 2018. "A New Automatic Sleep Staging System Based on Statistical Behavior of Local Extrema Using Single Channel EEG Signal." *Expert Systems with Applications* 104: 277–93. https://doi.org/10.1016/j.eswa.2018.03.020.

Şen, Baha, Musa Peker, Abdullah Çavuşoğlu, and Fatih V. Çelebi. 2014. "A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms." *Journal of Medical Systems* 38: 18. https://doi.org/10.1007/s10916-014-0018-0.

Sors, Arnaud, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean François Payen. 2018. "A Convolutional Neural Network for Sleep Stage Scoring from Raw Single-Channel EEG." *Biomedical Signal Processing and Control* 42: 107–14. https://doi.org/10.1016/j.bspc.2017.12.001.

Spira, Adam P., Christopher E. Gonzalez, Vijay K. Venkatraman, Mark N. Wu, Jennifer Pacheco, Eleanor M. Simonsick, Luigi Ferrucci, and Susan M. Resnick. 2016. "Sleep Duration and Subsequent Cortical Thinning in Cognitively Normal Older Adults." *Sleep* 39 (5): 1121–28. https://doi.org/10.5665/sleep.5768.

Stickgold, Robert, and Matthew P. Walker. 2007. "Sleep-Dependent Memory Consolidation and Reconsolidation." *Sleep Medicine* 8 (4): 331–43. https://doi.org/10.1016/j.sleep.2007.03.011.

Supratak, Akara, Hao Dong, Chao Wu, and Yike Guo. 2017. "DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25 (11): 1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116.

Šušmáková, Kristína, and Anna Krakovská. 2008. "Discrimination Ability of Individual Measures Used in Sleep Stages Classification." *Artificial Intelligence in Medicine* 44 (3): 261–77. https://doi.org/10.1016/j.artmed.2008.07.005.

Terzano Giovanni, Mario, Liborio Parrino, Arianna Smerieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, et al. 2002. "Atlas, Rules, and Recording Techniques for the Scoring of Cyclic Alternating Pattern (CAP) in Human Sleep." *Sleep Medicine* 3 (2): 187–99. https://doi.org/10.1016/S1389-9457(02)00004-7.

Thorpy, Michael. 2017. "International Classification of Sleep Disorders." In *Chokroverty S. (Eds) Sleep Disorders Medicine. Springer, New York, NY.*, 475–84. Springer. https://doi.org/10.1007/978-1-4939-6578-6_27.

Tian, J. Y., and J. Q. Liu. 2005. "Automated Sleep Staging by a Hybrid System Comprising Neural Network and Fuzzy Rule-Based Reasoning." In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 4115–18. https://doi.org/10.1109/IEMBS.2005.1615368.

Trinder, J., J. Kleiman, M. Carrington, S. Smith, S. Breen, N. Tan, and Y. Kim. 2001. "Autonomic Activity during Human Sleep as a Function of Time and Sleep Stage." *Journal of Sleep Research* 10 (4): 253–64. https://doi.org/10.1046/j.1365-2869.2001.00263.x.

Van-Someren, Eus J.W., J. M. Oosterman, B. Van Harten, R. L. Vogels, A. A. Gouw, H. C. Weinstein, A. Poggesi, Ph. Scheltens, and E. J.A. Scherder. 2018. "Medial Temporal Lobe Atrophy Relates More Strongly to Sleep-Wake Rhythm Fragmentation than to Age or Any Other Known Risk." *Neurobiology of Learning and Memory* 160: 132–38. https://doi.org/10.1016/j.nlm.2018.05.017.

Wacker, M., and H. Witte. 2013. "Time-Frequency Techniques in Biomedical Signal Analysis." *Methods of Information in Medicine* 52 (4): 279–96. https://doi.org/10.3414/ME12-01-0083.

Yu, Tong, and Hong Zhu. 2020. "Hyper-Parameter Optimization: A Review of Algorithms and Applications." *arXiv Preprint arXiv:2003.05689*, 1–56.

Zhang, Linda, Daniel Fabbri, Raghu Upender, and David Kent. 2019. "Automated Sleep Stage Scoring of the Sleep Heart Health Study Using Deep Neural Networks." *Sleep* 42 (11): zsz159. https://doi.org/10.1093/sleep/zsz159.

ORIGINAL PAPERS

I

MULTI-MODALITY OF POLYSOMNOGRAPHY SIGNALS' FUSION FOR AUTOMATIC SLEEP SCORING

by

Rui Yan, Chi Zhang, Karen Spruyt, Lai Wei, Zhiqiang Wang, Lili Tian, Xueqiao Li, Tapani Ristaniemi, Jihui Zhang & Fengyu Cong 2019

Journal of Biomedical Signal Processing and Control, Volume 49: 14-23.

DOI: 10.1016/j.bspc.2018.10.001

Biomedical Signal Processing and Control 49 (2019) 14–23

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Multi-modality of polysomnography signals' fusion for automatic sleep scoring

Rui Yan[a,b], Chi Zhang[a], Karen Spruyt[c], Lai Wei[d], Zhiqiang Wang[d], Lili Tian[e], Xueqiao Li[e], Tapani Ristaniemi[b], Jihui Zhang[f,*], Fengyu Cong[a,b,**]

[a] School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China
[b] Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland
[c] Lyon Neuroscience Research Center, INSERM U1028-CNRS UMR 5292-Waking Team, University Claude Bernard, School of Medicine, Lyon, France
[d] Otolaryngology Department, Affiliated Zhongshan Hospital of Dalian University, China
[e] Department of Psychology, University of Jyväskylä, 40014, Jyväskylä, Finland
[f] Department of Psychiatry, The Chinese University of Hong Kong, Shatin, Hong Kong, China

ARTICLE INFO

Article history:
Received 28 March 2018
Received in revised form
11 September 2018
Accepted 6 October 2018
Available online 23 November 2018

Keywords:
Polysomnography
Multi-modality analysis
Rules of R&K
Automatic sleep scoring

ABSTRACT

Objective: The study aims to develop an automatic sleep scoring method by fusing different polysomnography (PSG) signals and further to investigate PSG signals' contribution to the scoring result.
Methods: Eight combinations of four modalities of PSG signals, namely electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and electrocardiogram (ECG) were considered to find the optimal fusion of PSG signals. A total of 232 features, covering statistical characters, frequency characters, time-frequency characters, fractal characters, entropy characters and nonlinear characters, were derived from these PSG signals. To select the optimal features for each signal fusion, four widely used feature selection methods were compared. At the classification stage, five different classifiers were employed to evaluate the validity of the features and to classify sleep stages.
Results: For the database in the present study, the best classifier, random forest, realized the optimal consistency of 86.24% with the sleep macrostructures scored by the technologists trained at the Sleep Center. The optimal accuracy was achieved by fusing four modalities of PSG signals. Specifically, the top twelve features in the optimal feature set were respectively EEG features named zero-crossings, spectral edge, relative power spectral of theta, Petrosian fractal dimension, approximate entropy, permutation entropy and spectral entropy, and EOG features named spectral edge, approximate entropy, permutation entropy and spectral entropy, and the mutual information between EEG and submental EMG. In addition, ECG features (e.g. Petrosian fractal dimension, zero-crossings, mean value of R amplitude and permutation entropy) were useful for the discrimination among W, S1 and R.
Conclusions: Through exploring the different fusions of multi-modality signals, the present study concluded that the multi-modality of PSG signals' fusion contributed to higher accuracy, and the optimal feature set was a fusion of multiple types of features. Besides, compared with manual scoring, the proposed automatic scoring methods were cost-effective, which would alleviate the burden of the physicians, speed up sleep scoring, and expedite sleep research.

1. Introduction

Sleep covers almost one-third of the human lifespan [1]. Adequate and high-quality sleep is vital to our physical and mental well-being [2]. To study sleep dynamic, Rechtschaffen and Kales [3] introduced rules for labelling each sleep segment of 30 s as wakefulness (W), NREM stage (S1, S2, S3, or S4) or REM stage (R). Each sleep stage has a certain proportion and plays a vital role in the recuperation of living organisms. Studies have found that the distortion of sleep structure could lead to catastrophic outcomes. For example, REM disturbance slows down the perceptual skill improvement [4], deprivation of slow wave sleep is associated with Alzheimer's disease [5], insufficient sleep duration has detrimental effects on metabolic health [6], etc. Therefore, assessing sleep behavior and analyzing sleep structure are crucial in clinical applications.

* Corresponding author.
** Corresponding author at: School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China.
E-mail addresses: jihui.zhang@cuhk.edu.hk (J. Zhang), cong@dlut.edu.cn (F. Cong).

https://doi.org/10.1016/j.bspc.2018.10.001

A golden tool for quantitatively assessing sleep is PSG test that records simultaneously tens of physiological signals containing EEG, EOG, EMG, ECG, etc. Generally, according to the rules of Rechtschaffen & Kales (R&K) [3] and the recently updated American Academy of Sleep Medicine rules (AASM) [7], these PSG recordings are scored mutually by at least one registered sleep technologist (RST) to get the sequence of sleep stages. However, the manual scoring process is rather time-consuming [8] and subjective to some extent [9,10]. By contrast, automatic sleep scoring has shown advantages of cost-effective and preferable scoring performance. According to signal types employed in previous researches, these scoring methods can be divided into two categories, single-modality processing and multi-modality processing.

Single-modality scoring methods mainly based on EEG since EEG signals contained valuable and interpretable information resembling brain activities. Şen et al. [11] extracted 41 parameters from channel C3-A2 achieving a classification accuracy of 97.03%, which was fairly high among other related studies. However, as Diykh et al. [12] claimed, it was challenging to achieve such high accuracy. Instead of traditional linear features, Liang et al. [13] employed multiscale entropy and autoregressive models for single-channel EEG and obtained a good performance. Another new attempt appeared in Dimitriadis and his colleagues' study [14], in which sleep stages were evaluated by calculating cross-frequency coupling of predefined frequency pairs. In addition, to improve the performance of automatic sleep scoring, multiple EEG channels were investigated in existing studies. By fusing joint collaborative representation and joint sparse representation algorithms [15], a two-stage multi-view learning algorithm was constructed achieving a mean scoring accuracy of $81.10 \pm 0.15\%$.

According to the R&K rules, multi-modality of signals (such as EEG, EOG and EMG) were required for technologists to score sleep stages. Therefore, numerous studies considered multiple PSG signals. In Estrada et al.'s research [16], it concluded that EOG and EMG served as an important switching index of different sleep stages. In addition, using EEG, EMG and EOG recordings of five healthy subjects, Özşen [17] developed five different artificial neural network architectures to train each sleep stage separately. This separation of training procedure exhibited its superiority as Özşen claimed. Similarly, the multiple classifiers were used in Zhang and his colleagues' study [18] which achieved a high accuracy of 91.31%.

In previous studies, ECG signals, as vital physiological measures, were mainly used in home sleep monitoring systems [19–21]. To our knowledge, very limited articles explored it together with EEG, EOG and EMG in automatic sleep scoring algorithms. In Šušmáková and Krakovská's study [22], ECG was considered to be negligible compared with EEG, EOG and EMG. By extracting 74 measures from the signals of EEG, EMG, EOG and ECG, Krakovská and Mezeiová [23] found the ECG feature named zero-crossing rate performed well in automatic sleep scoring, but was still inferior to those from the other three signals. Whereas, four distance-based ECG features, related to the similarity of a baseline ECG epoch to the rest of epochs, were considered important in Gharbali et al.'s study [24]. Therefore, new evidence is needed to clear the PSG signals' contribution to automatic sleep scoring.

Based on the above issues, the present study aims to develop a multi-modality sleep scoring method and to detect PSG signals' contribution to sleep scoring. Besides, different feature selectors and classifiers are compared to provide a reference for future studies in this area. The main contributions of this work are presented as following:

Table 1
Distribution of sleep stages on dataset.

Sleep stages	W	R	S1	S2	S3	S4	Total
Number of epochs	449	1405	280	2162	570	1181	6047

1 Developing an automatic sleep scoring method by fusing four modalities of PSG signals,
2 Analyzing four types of PSG signals' contribution to distinguishing sleep stages,
3 Comparing the performance of different signals' fusions in sleep scoring,
4 Evaluating the discrimination ability of the optimal features selected from different signals' fusions,
5 Assessing the effect of different feature selectors and classifiers.

The article is organized as follows. Section 2 explains the details of experimental data and methodology of this study, together with a brief description of feature selection methods and classifiers employed in this study. Section 3 demonstrates the performance of proposed method, different feature selectors and classifiers. Section 4 provides discussions of results and limitations of this study. Finally, Section 5 gives conclusions of this paper.

2. Methodology

2.1. Data description

All-night PSG sleep recordings were provided by PhysioBank [25] (The CAP Sleep Database [26]). The CAP Sleep Database was a collection of 108 polysomnographic recordings registered at the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy, which included EEG channels, EOG channels, submental EMG channel and other electrophysiological signals. A detailed description and definition of CAP Sleep Database was given in Terzano et al.' s study [27]. In the present study, four types of PSG signals (EEG, EOG, EMG and ECG) were required. However, these measurements were not contained in each recording of the database. Given that, a total number of 6 healthy subjects that contained requisite four types of PSG signals were selected. The total duration of all recordings combined was 50 h, 33 min and 30 s with 8.5 h' average sleep time for each subject. The age of subjects ranged from 23 to 37 years, with a mean of 32 years and a standard deviation of 5.4 years.

For each overnight sleep recording, a traditional hypnogram followed the R&K rules was available, which represented the manual classification of sleep stages by the experts on 30 s non-overlapping segments. The hypnograms were used as a reference to evaluate automatic classification results. For each subject, the following four modalities of signals were analyzed: one EEG channel (C4, referred to the left mastoids A1 following the 10–20 international electrode placement system), one relative EOG channel (ROC, referred to LOC), one submental EMG channel and one ECG channel.

All signals were sampled or resampled to 512 Hz. In order to remove noise and artifacts, a notch filter at 50 Hz, a high-pass filter with a cut-off frequency of 0.3 Hz and a low-pass filter with a cut-off frequency of 30 Hz were applied to the signals of EEG, EOG and ECG [26]. In terms of EMG, a notch filter at 50 Hz, a high-pass filter with a cut-off frequency of 10 Hz and a low-pass filter with a cut-off frequency of 75 Hz were performed [22]. Afterwards, all the signals were divided into 30-second epochs, each epoch corresponding to a single sleep stage in hypnogram. Table 1 presented the distribution of sleep stages in the present data set.

Table 2
List of features from EEG signal.

P	Feature	P	Feature	P	Feature	P	Feature
1	minV	11	The 25_{th} percentile	37	Power spectral density	61	Approximate entropy
2	maxV	12	The 50_{th} percentile	38	Spectral edge	62	Permutation entropy
3	SD	13	The 75_{th} percentile	39-42	Absolute power spectral	63	Spectral entropy
4	Mean	14	The 90_{th} percentile	43-46	Relative power spectral		
5	Variance	15	HA	47-55	Power ratios		
6	Skewness	16	HM	56	PFD		
7	Kurtosis	17	HC	57	Mean teager energy		
8	Median	18-20	AR coefficients for EEG epoch	58	Energy		
9	ZCs	21-32	AR coefficients for rhythm waves	59	Mean curve length		
10	The 5_{th} percentile	33-36	Energy for rhythm waves	60	Hurst exponent		

Table 3
List of features from EOG signal.

P	Features	P	Features	P	Features
64	minV	74	HM	86	Spectral entropy
65	maxV	75	HC	87	Permutation entropy
66	Mean	76-78	AR coefficients		
67	SD	79	Power spectral density		
68	Variance	80	Spectral edge		
69	Skewness	81	PFD		
70	Kurtosis	82	Hurst exponent		
71	Median	83	Energy		
72	ZCs	84	Mean teager energy		
73	HA	85	Approximate entropy		

Table 4
List of features from EMG signal.

P	Features	P	Features	P	Features
88	minV	98	HM	110	Spectral entropy
89	maxV	99	HC	111	Permutation entropy
90	Mean	100-102	AR coefficients		
91	SD	103	Power spectral density		
92	Variance	104	Spectral edge		
93	Skewness	105	PFD		
94	Kurtosis	106	Hurst exponent		
95	Median	107	Energy		
96	ZCs	108	Mean teager energy		
97	HA	109	Approximate entropy		

2.2. Methodology and algorithm description

A total of 232 features from 7 different categories (time, frequency, time-frequency, fractal, entropy, nonlinearity and mutual-based features) are extracted in the feature extraction phase. The following section introduces the details of feature extraction methods. All the features and their corresponding origins are listed in Tables 2–5.

2.2.1. Time domain features

Statistical parameters, containing minimum value (minV), maximum value (maxV), standard deviation (SD), arithmetic mean (Mean), variance, skewness, kurtosis and median are derived from the segments of EEG, EOG, EMG and ECG. These statistical parameters are good indicators of the amplitude and distribution of time series. Details of these computations can be found in Şen et al.'s research [11].

Hjorth parameters (i.e., activity, mobility and complexity) are often used in the analysis of EEG signals [28]. In this article, they are derived from the segments of EEG, EOG, EMG and ECG.Hjorth activity:

$$HA = \sigma_0^2 \tag{1}$$

Hjorth mobility:

$$HM = \sigma_1/\sigma_0 \tag{2}$$

Hjorth complexity:

$$HC = \frac{\sigma_2/\sigma_1}{\sigma_1/\sigma_0} \tag{3}$$

where σ_0, σ_1 and σ_2 respectively denote the standard value of time series x_n, its first derivative $\dot{x}_n$, and its second derivative $\ddot{x}_n$.

AR coefficients, encoding a signal into several coefficients, are capable of undermining time-domain dynamics of signals which cannot be revealed by other features [29]. The following equation illustrates an AR model.

$$x(n) = -\sum_{i=1}^{p} a_i x(n-i) + e(n) \tag{4}$$

where $x(n)$ is a time series, a_i denoting AR coefficients, and $e(n)$ indicates prediction error. In the present study, only the first three coefficients were considered.

Zero-crossings is a time-based feature widely used in electronics, mathematics, image processing and signal processing [11]. Zero-crossings is an indicator of signal's noise ratio.

$$(x_{n-1} < 0 \,\&\, x_n > 0) \,\|\, (x_{n-1} > 0 \,\&\, x_n < 0) \,\|\, (x_{n-1} \neq 0 \,\&\, x_n == 0) \tag{5}$$

Percentile analysis provides information about the distribution of signal's amplitude, which is helpful to the discernment of sleep stages. The 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles of signal's amplitude are calculated in this study.

Table 5
List of features from ECG signal.

P	Feature	P	Feature	P	Feature	P	Feature
112	Energy	124	Petrosian Fractal dimension	134	Mean RRIs	144	MAD of detrended RRIs changes
113	The 4^{th} order power	125	ZCs	135	Median RRIs	145	SD of RRIs difference
114	Curve length	126	HA	136	Mean detrended RRIs	146	Approximate entropy
115	Mean teager energy	127	HM	137	Standard deviation of RRIs	147	Permutation entropy
116-118	AR coefficients	128	HC	138	Difference between Max and Min RRIs	148	Spectral entropy of RRIs
119	Power spectral density	129	Mean of R_{amp}	139	Inter-quartile range		
120	Spectral edge	130	SD of R_{amp}	140	Mean absolute deviation (MAD)		
121	Mean-PSD	131	Mean HR	141	RMS of RRIs changes		
122	Median-PSD	132	Mean detrend HR changes	142	Mean of RRIs changes		
123	Spectral entropy	133	Mean nondetrend HR changes	143	MAD of nondetrended RRIs changes		

2.2.2. Frequency and time-frequency features

Power spectral density: Each epoch is segmented into fifteen non-overlapping parts (1024 points each) by hamming window. Then, the fifteen vectors are zero-padded to the length of 2048 points, respectively. The final spectral density is achieved by averaging spectral densities of the fifteen segments [22].

Spectral edge is defined as the frequency corresponding to 90% of the total spectral power [30]:

$$\sum_{f=f_{min}}^{edge} P(f) = 0.9 \sum_{f=f_{min}}^{30Hz} P(f) \tag{6}$$

where f_{min} is 0.3 Hz in terms of EEG, EOG and ECG, and 10 Hz in EMG.

Absolute and relative spectral power are obtained from four frequency bands of EEG, namely, 0.3–4 Hz (delta), 4–8 Hz (theta), 8–16 Hz (alpha) and 16–30 Hz (beta). Here, maximum overlap discrete wavelet is employed to decompose EEG signals. Afterwards, power spectral density is calculated within each frequency band. Then the relative spectral power is defined as the ratio of spectral power within the specific frequency band to the total spectral power. The total spectral powers of EEG, EOG and ECG signals are computed within the range of 0.3–30 Hz. Whereas, the total spectral power of EMG is calculated within the range of 10–30 Hz.

Power ratios are computed based on relative spectral powers in aforementioned frequency bands. The following power ratios are computed: delta/theta, delta/alpha, delta/beta, theta/alpha, theta/beta, alpha/beta, alpha/(theta + delta), delta/(theta + alpha) and theta/(beta + delta).

2.2.3. Nonlinear-based features

Energy is deemed as a reliable indicator to discern different activities of sleep stages [11]. It is defined as,

$$E = \sum_{n=1}^{N} x(n)^2 \tag{7}$$

where $x(n)$ is time series, and N denotes the length of time series.

Mean teager energy (MTE) is a non-linear operator that can effectively track the energy of signals. It can be derived by the following formula [31],

$$MTE = \frac{1}{N}\sum_{n=3}^{N} \left(x(n-1)^2 - x(n)x(n-2)\right) \tag{8}$$

where $x(n)$ is time series, and N denotes the length of time series.

Mean curve length (MCL) was proposed by Esteller et al. [32] to provide an estimation of Katz fractal dimension. It is widely used in the identification of EEG signals' activities [11]. MCL is defined as

$$MCL = \frac{1}{N}\sum_{n=2}^{k} \left|x(n) - x(n-1)\right| \tag{9}$$

where $x(n)$ is time series, N denoting the length of time series, and k is the last sample in the epoch.

Hurst exponent is used in time series analysis to present non-stationary or antistatic signal states observed in sleep [33]. It is defined as

$$H = \log\left(R/S\right)/log(T) \tag{10}$$

where T is the duration of the time series, R/S the value of rescaled range, R the difference between maximal and minimal "accumulated" values, and S is the standard deviation of observed series $x(n)$ [34].

2.2.4. Fractal feature

Fractal dimension is a chaotic parameter elucidating the complexity of signals. Petrosian fractal dimension (PFD) facilitates the rapid calculation of fractal dimension [11]. The rapid calculation is achieved through transforming the signal into a binary sequence. It can be estimated by the following formula,

$$PFD = log_{10}N/\left(log_{10}N + log_{10}\left(N/(N + 0.4N_\sigma)\right)\right) \tag{11}$$

where N is the length of time series, and N_σ is the number of sign changes in the signal derivative.

2.2.5. Entropy-based features

Approximate entropy (ApEn) is a measure used to quantify the unpredictability or randomness of signals. It has been reported that the mean value of approximate entropy changed significantly with different sleep stages[33].

Permutation entropy (Pen) is a complexity measure of time series based on comparing neighboring values [11]. More details of this measure can be found in Liu and Wang's study [35].

Spectral entropy is computed based on the relative power spectral density P_{ref} [22]. It is defined as

$$SEN = \frac{1}{ln(N)}\sum_{f=f_{min}}^{30Hz} P_{ref}(f)\, ln\left(P_{ref}(f)\right) \tag{12}$$

where N is the length of time series, and f_{min} is set as 0.3 Hz in terms of EEG, EOG and ECG signals, and 10 Hz of EMG.

2.2.6. Mutual-based features

To evaluate the relationship between two signals, the spectral coherence and phase angle are computed between every two epochs from four modalities of signals. To elaborate frequency bands, aforementioned four frequency bands (0.3–4 Hz, 4–8 Hz, 8–16 Hz and 16–30 Hz) are further divided as following: 0.3–2 Hz, 2–4 Hz, 4–6 Hz, 6–8 Hz, 8–10 Hz, 10–12 Hz, 12–14 Hz, 14–16 Hz and 16–30 Hz. For EMG signals, spectral measures are only computed in frequency bands higher than 10 Hz because a high-pass filter is used in the pre-processing stage. More details about coherence and phase angle can be found in Šušmákovác and Krakovská's study [22].

Mutual information that measures the mutual dependence of two variables is derived from the signals of EEG, EOG, EMG and ECG. It is calculated based on marginal entropy and joint entropy of two variables. The definition of mutual information can be found in Šušmákovác and Krakovská's study [22].

2.2.7. Heart-beat related features

To extract heartbeat-related features, RR intervals (RRI, the time interval between consecutive heartbeats or R peaks) are required. In order to highlight R wave, the median filters of 200 ms and 600 ms are used to ECG epochs successively. In the present study, Pan and Tompkins' QRS detector [36] is employed to locate R peaks. Then, the amplitude of R waves, heart rates, RR intervals and the change of RR intervals are calculated. A total of 37 features are extracted from ECG segments, a detailed description of these measures which can be found in Noviyanto et al.'s study [37].

2.3. Feature normalization and selection

2.3.1. Feature normalization

After all features extracted, a feature set, with the dimension of 6047 × 232 for four modalities of PSG signals' fusion, is achieved. In order to balance numerical ranges and to avoid numerical difficul-

ties in classification, each feature is separately normalized to [0, 1] by the following formula,

$$\bar{p}_{i,j} = \left[p_{i,j} - min\left(p_{:,j}\right)\right] / \left[max\left(p_{:,j}\right) - min\left(p_{:,j}\right)\right] \quad (13)$$

where p_j denotes a vector of each independent feature, and $p_{i,j}$ is an element in the j_{th} feature vector.

2.3.2. Feature selection methods

Feature selection is a process of selecting an effective subset from canonical features to reduce dimension, shorten training time and simplify learning model. Feature selection methods are highly dependent on their defined objective function that heavily influences selection results. Therefore, to find the most discriminative features, four different feature selectors are considered in the present article. The employed feature selectors are ReliefF algorithm, improved distance-based evaluation methods (IDE), genetic algorithms (GA) and forward selection process (FSP). A brief description of these methods is provided below.

ReliefF is a supervised feature-weighting algorithm of the filter model that searches for the nearest neighbors of instances from different classes. *ReliefF* weights features according to how well they differentiate instances of different classes [38]. ReliefF is robust and also able to deal with incomplete and noisy data [39]. Therefore, ReliefF, as a widely used feature selector in the multi-modality analysis, is employed as a prime selector in the present study.

Improved distance-based evaluation methods (IDE) was developed by Lei et al. [40]. The method is especially useful in feature selection for the purpose of classification. It grades features between [0, 1] in where a higher value indicates a higher discriminative capability. The discriminative feature set can be selected by setting a threshold for graded results. IDE method has the advantage of simplicity and reliability as other distance-based feature selectors.

Genetic algorithms (GA) is a population-based technique. Instead of single potential solutions, it uses a population of potential solutions. That strategy is particularly suitable for multi-objective optimization. GA has motivated an increasing number of applications in engineering and related fields due to its capability of finding global optima and solving discontinuity and noise problems [41].

Forward selection process (FSP) is the simplest method among sequential strategies. It is a greedy search algorithm that determines iteratively the effective feature subset by adding one feature per iteration, on the condition that the newly added feature increases the value of objective function [42]. Once the termination condition is satisfied, the selected number of features, which reaches the lowest error at the first time, will be chosen as the dimension of optimal features set [43]. The main drawback of the sequential approach is that it gravitates toward local minima due to the inability to re-evaluate the effectiveness of features. Once a feature is added or discarded from the final set of features, the results would be irreversible.

2.4. Classification

In order to capture characters of sleep stages and to predict new instances, five different classifiers are employed, namely k-nearest neighbor classifier (KNN), binary decision tree (BDT), naive Bayes (NB), random forest (RF) and support vector machine (SVM). Their performances are compared to ascertain the optimal classifier for automatic sleep scoring. Introduction of these classifiers is given in the following.

K-nearest neighbor classifier (KNN) is a nonparametric classification approach. It has been pervasively used in the fields of science and engineering as a benchmark classifier due to its robust performance [18]. An instance is classified based on the majority votes of its closest training neighbors. The number of closest training neighbors k is crucial for classification results. Generally, the increasing of k value would attenuate the noise effect in the classification, whereas it would also obscure the boundaries between classes. In the present study, different k values are examined. It turns out $k = 5$ is the best choice. A peculiarity of the KNN algorithm is its sensitivity to the local structure of data, especially when the class distribution is skewed.

Binary decision tree (BDT) is a decision-support tool that used a tree-like graph or model of decisions. It takes booleans as inputs and produces booleans as output. Due to its simplicity and ease of understanding, BDT is widely used in classification, data mining and machine learning [44]. It adopts multi-stages or consecutive approaches in the classification procedure. Trees are generated at the first stage of classification. At the second stage, the test sample is discriminated from the root node to the child node with higher probability. The process is recursively executed until the sample is assigned to the leaf node corresponding to a specific category. The defect of the BDT algorithm is that it is sensitive to noise.

Naive Bayes (NB) is a probabilistic classifier based on Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions between features. It assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [45]. Naive Bayes classifier is highly scalable, which requires a linear relationship between the number of instances and the number features. Maximum-likelihood training, one of the probability model of Naive Bayes classifiers, evaluates a closed-form expression that requires linear time, rather than expensive iterative approximation used in other classifiers. An advantage of NB is that it requires only a small number of training data to weight the importance of parameters in classification.

Random forest (RF) is an ensemble learning method for classification. It constructs a multitude of trees at the training period, and the final classification results achieved by the most votes in the forest [1]. The training algorithm of random forest applies the general technique of bootstrap aggregating which selects random instances with replacement from the training set. The bootstrapping procedure ensures that even though a single tree's decision is highly sensitive to noise, the average decision of multiple trees would not be influenced, as long as these trees are not correlated.

Support vector machine (SVM) is a supervised method with associated learning algorithms [46]. Generally, SVM implicitly constructs a hyperplane or a set of hyperplanes in a high- or infinite-dimensional space for its inputs in terms of kernel function. A satisfactory separation is often achieved in the hyperplane with the largest distance to the nearest training data point (so-called functional margin). The distance of functional margin negatively correlates with classifier's error rate. Compared with other types of tree algorithms, SVM is capable of classifying complicated problems via different kernels. The main advantage of SVM is its predominant generalization capability in statistical learning.

3. Performance assessment

For a given set of features, the following 10-fold cross-validation procedure was performed:

1 All samples were randomly divided into ten equal sized subsets.
2 In ten subsets, a single subset was retained as test data, and the remaining nine subsets were used as training data.
3 The cross-validation process was then repeated ten times, with each of ten subsets used exactly once as the test data.

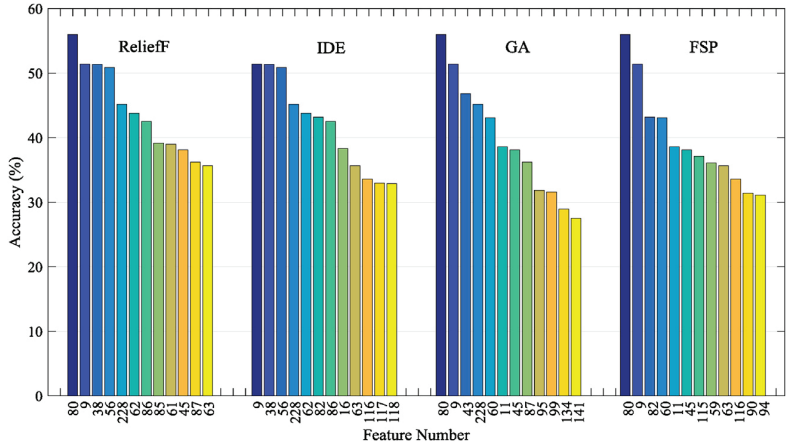


Fig. 1. Selected features of different feature selection methods.
* 1-63: EEG features; 64-87: EOG features; 88-111: EMG features; 112-148: ECG features; 149-187: Coherence; 188-226: Phase angles; 227-232: Mutual information.

All assessment indexes were calculated based on the total results of 10-fold cross-validation. The following indexes used to evaluate the performance of the proposed methods.

Accuracy which indicates the fraction of the total number of correct detections in the sleep scoring. It is defined as,

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}(\%) \tag{14}$$

where, *TP*, *TN*, *FP* and *FN* respectively denote true positives, true negatives, false positives and false negatives [11].

Sensitivity which represents the fraction of positive epochs that are correctly identified by the algorithm [11].

$$Sen = \frac{TP}{TP + FN}(\%) \tag{15}$$

Specificity which denotes the fraction of corresponding negative epochs being correctly rejected [11].

$$Spe = \frac{TN}{FP + TN}(\%) \tag{16}$$

Positive predictive value which is the fraction of correct detections of positive epochs with respect to the total number of positive epochs [11].

$$Ppv = \frac{TP}{TP + FP}(\%) \tag{17}$$

As described above, a total of 232 features were extracted from the signals of EEG, EOG, EMG and ECG. After normalization, these features were fed into four different feature selectors in order to pick out discriminative features. Section 3.1 compared the performance of different feature selectors. In Section 3.2, five different classifiers were compared to ascertain the optimal one for automatic sleep scoring. In Section 3.3, different signals' fusions were compared to explore PSG signals' contribution and to highlight discriminative features of different signals' fusions. Detailed comparative analysis was provided below.

3.1. Performance evaluation of different feature selectors

In order to determine the effective feature selector, a grid search was carried out in terms of the results obtained from four feature selectors (ReliefF, IDE, GA and FSP). The features, selected by feature selectors, were fed into RF classifier one by one to evaluate its capability of distinguishing sleep stages. The top twelve outcomes and its corresponding classification accuracy were displayed in Fig. 1. As Fig. 1 described, the EOG feature named spectral edge (P: 80; Acc.: 55.88%) achieved the highest accuracy that indicated its outstanding capability of distinguishing sleep stages. The comparison results (showed in Fig. 1) demonstrated that the optimal feature set was obtained by ReliefF, since its elements showed higher discriminative accuracy than those from other selectors. The top twelve features in the optimal feature set were respectively EEG features named zero-crossings, spectral edge, relative power spectral of theta, Petrosian fractal dimension, approximate entropy, permutation entropy and spectral entropy, and EOG features named spectral edge, approximate entropy, permutation entropy and spectral entropy, and the mutual information between EEG and submental EMG. The fact that most of discriminative features were generated from EEG and EOG signals revealed its indispensable role in automatic sleep scoring.

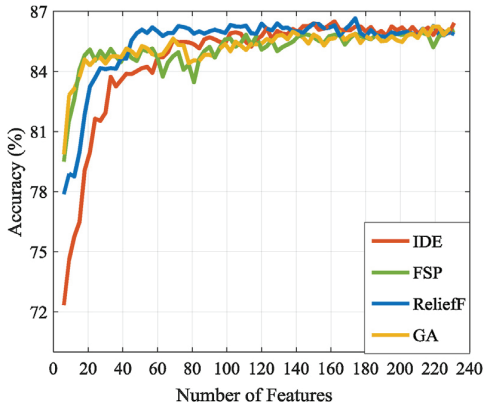


Fig. 2. Mean value of 10-fold cross-validation for specific feature selector.

Fig. 2 showed the mean classification accuracy of 10-fold cross-validation corresponding to the numbers of features from four feature selectors. As can be seen from Fig. 2, the automatic sleep analysis required at least four or five features. Meanwhile, an increasing number of features would contribute to classification accuracy. Fig. 2 also revealed that GA, SFP and ReliefF achieved a higher classification accuracy at the initial phase that meant they could select the most discriminative features in the first several steps. However, GA and SFP showed a clumsy increasing of accuracy when the number of features exceeded 20. On the contrary, ReliefF showed sustained growth, and its performance was more stable after the number of features exceeded 60. Based on the results

Table 6
Performance of the proposed method for various feature selection methods and classifier.

Classifiers	Feature selection methods			
	ReliefF	IDE	GA	FSP
KNN	85.26%	82.69%	81.97%	81.41%
Tree	77.08%	76.25%	75.67%	74.93%
Bayes	64.51%	56.95%	70.49%	74.48%
RF	86.24%	85.29%	84.74%	85.17%
SVM	83.00%	82.37%	78.39%	80.73%
Num.Feature	56	70	20	20

of Fig. 1 and Fig. 2, ReliefF was chosen as a predominant feature selector in the following experiments.

3.2. Evaluation of different classifiers

After the optimal feature selector determined, the performance of different classifiers was evaluated in this section. Table 6 depicted the mean classification accuracy of 10-fold cross-validation obtained from five classifiers and their corresponding feature selectors. Here, the number of features was determined based on Fig. 2. In order to decrease the consumption of CPU and memory, we chose as few features as possible in the condition of no significant improvement of accuracy. Table 6 revealed that the optimal performance with the highest accuracy was achieved by random forest classifier (RF) with 56 features selected by ReliefF. The superior performance of RF has been verified in many studies [20,29].

Table 7 presented the confusion matrix of the optimal classification result. NREM-stage S4 was identified with the highest classification sensitivity of 93.3%. Wakefulness, REM and NREM-stage S2 were correctly classified with a sensitivity above 80%, whereas, stage S3 was correctly classified with a sensitivity of 77.6%. Stage S3 was misclassified as its neighbor NREM stage S2 in 27.9% of the instances and misclassified as stage S4 in 11.6% instances. The least recognizable was stage S1 with a sensitivity of 66.3%. Stage S1 was considered as a transient state between wakefulness and "real" sleep. Stage S1 commonly accounted for about 5% of the total sleep duration of a healthy subject. Therefore, the consensus in scoring S1 among experts was quite obscure, compared with other stages [23]. The confusion matrix showed that 8.5% instances of stage S1 were misclassified as REM sleep, 28.6% as S2, and 17.8% as wakefulness. According to our interpretation, the low classification accuracy of stage S1 and stage S3 may also attribute to the imbalanced instances which consequently made the result inclined to the labels with greater numbers.

To demonstrate the performance of proposed methods, we compared the proposed algorithms with those from the other three studies. These algorithms were applied to the present dataset, as described in Section 2.1. The comparison results were shown in Table 8, wherein the classification accuracy of 10-fold cross-validation was shown by the mean value along with the standard deviation in parentheses. As shown in Table 8, the proposed methods, with ReliefF as a feature selector and RF as a classifier, generated the optimal performance in each signal fusion. However, the results generated by referred algorithms were less satisfactory than those reported in the cited studies. The divergence of results might be influenced by different datasets and limited instances. The number of instances in the present article is the least compared to those in the three references.

3.3. Comparison of different signals' fusions

Different fusions of EEG, EOG, EMG and ECG signals were performed to clarify signals' contribution. Table 9 and Fig. 3 summarized the results of different signals' fusions. Table 9 presented the top twelve features from six different signals' fusions. These features were selected by ReliefF. Closer inspection of Table 9 showed that features extracted from EEG and EOG achieved a higher classification accuracy when compared with those from EMG and ECG signals. Meanwhile, the features, namely entropy, spectral edge, Petrosian fractal dimension and zero-crossings showed a crucial role in automatic sleep scoring. In addition, some traditional features named arithmetic mean, standard deviation and Hjorth parameters also showed good discrimination in single modality process.

Fig. 3 showed the accuracy of different signals' fusions. EEG showed the outstanding contribution to sleep scoring. Furthermore, the addition of a few more characteristics from other modalities of signals would still be worth considering. The richness of signal modalities contributed to the increasing accuracy of sleep stages classification. According to the results shown in Fig. 3, the fusion of two modalities PSG signals showed significant improvement in the accuracy. The addition of the third and the fourth modality of signal improved classification accuracy to some extent. No significant difference was found among signal types. The slight improvement, generated by the addition of the fourth signal, might attribute to information saturation. However, the definitive conclusion would require further investigations. Besides, the addition of signals and features was a challenge for feature selector so that a preeminent feature selector was requisite.

To further elucidate multi-modality signals' contributions to automatic sleep scoring, a feature selection process, for four modalities of signals (EEG, EOG, EMG and ECG), was performed to distinguish every pair of sleep stages. The top 15 features selected by ReliefF were shown in Table 10, wherein features sorted in ascending order of discriminative capability. As can be seen from Table 10, the features from EEG accounts for a large proportion. Meanwhile, the features from ECG signal showed its contribution to the distinguishing among W, S1 and R. For mutual-based features, coherence and mutual information between two signals performed well in the discrimination of S1-S2, S1-R and S3-S2. In term of EMG

Table 7
Scoring agreement between manual scoring and the proposed method.

	Technologists' score stage							Sen.
	Stage	W	R	S1	S2	S3	S4	
Proposed	W	394	8	50	35	2	2	80.2%
	R	6	1299	24	99	0	0	90.9%
	S1	26	8	126	30	0	0	66.3%
	S2	21	88	80	1931	159	15	84.2%
	S3	0	0	0	57	343	42	77.6%
	S4	2	2	0	10	66	1122	93.3%
Ppv		87.8%	92.4%	45.0%	89.3%	60.2%	95.0%	
Spe.		99.0%	97.7%	97.4%	93.8%	95.9%	98.8%	
Acc.								86.24%

Table 8
Algorithm performance with different signals' fusions and its comparison with three references.

	EEG	EEG, EOG&EMG	EEG, EOG, EMG&ECG
Reference	59.2715% (±1.4580%) [10]	82.9519% (±1.4870%) [2]	76.1726% (±3.3857%) [23]
Proposed	76.0531% (±0.9056%)	85.3068% (±0.8244%)	86.244% (±1.0725%)

Table 9
Selected features of different signals' fusions.

EEG		EOG		EMG		ECG		EEG & EOG & EMG		EEG & EOG & EMG & ECG	
P	Acc.	P	Acc.	P	Acc.	P	Acc.	P	Acc.	P	Acc.
9	51.43%	80	56.01%	104	40.57%	120	38.88%	80	56.01%	80	56.09%
38	51.38%	81	47.41%	96	37.46%	125	36.74%	228	45.15%	9	51.43%
56	50.91%	72	47.37%	105	37.20%	124	36.56%	62	43.79%	38	51.38%
62	43.79%	82	43.19%	95	31.85%	116	33.59%	82	43.19%	56	50.91%
60	43.09%	86	42.53%	99	31.57%	117	32.95%	86	42.53%	228	45.15%
61	38.96%	85	39.17%	90	31.42%	118	32.89%	85	39.17%	62	43.79%
11	38.60%	74	38.93%	98	30.33%	129	29.24%	61	38.96%	86	42.53%
16	38.35%	87	36.23%	109	30.05%	127	28.89%	74	38.93%	85	39.16%
45	38.10%	66	34.52%	91	30.01%	130	28.39%	11	38.60%	61	38.96%
13	36.76%	64	33.10%	106	29.41%	146	27.42%	45	38.10%	45	38.10%
4	35.65%	67	30.60%	111	28.25%	148	26.84%	87	36.23%	87	36.23%
63	35.64%	75	29.60%	110	28.06%	147	26.77%	63	35.64%	63	35.64%

*1-63: EEG features; 64-87: EOG features; 88-111: EMG features; 112-148: ECG features; 149-187: Coherence; 188-226: Phase angles; 227-232: Mutual information.

Table 10
Selected features for distinguishing specific pair of sleep stages.

Top fifteen features	Stages	S4-S3	S4-S2	S4-S1	S3-S2	S3-S1	S2-S1	S4-R	S3-R	S2-R	S1-R	S4-W	S3-W	S2-W	S1-W	W-R
	Top.1	80	80	9	80	9	62	9	80	80	80	9	9	80	61	38
	Top.2	9	9	38	9	38	82	38	228	228	62	38	38	9	125	61
	Top.3	38	56	56	56	56	60	56	62	82	85	56	56	56	44	16
	Top.4	56	81	228	228	43	85	228	82	86	125	228	62	81	124	125
	Top.5	81	72	62	62	228	61	62	60	85	124	62	60	72	59	44
	Top.6	72	228	82	82	62	87	82	86	74	64	82	86	85	116	124
	Top.7	228	62	86	60	82	129	60	85	45	95	60	61	61	117	116
	Top.8	86	82	61	86	60	123	86	11	52	129	86	16	74	118	117
	Top.9	85	86	11	85	86	231	61	45	55	110	61	44	16	40	64
	Top.10	61	85	16	61	61	110	11	52	125	187	11	87	44	129	118
	Top.11	74	61	45	13	11	147	16	55	124	147	16	46	87	229	95
	Top.12	11	74	55	87	16	181	45	13	87	181	45	116	46	123	129
	Top.13	13	11	13	129	13	175	13	87	95	175	44	117	116	231	229
	Top.14	87	13	63	175	87	169	4	95	129	169	46	118	117	232	231
	Top.15	63	63	129	169	129	220	63	129	147	220	63	129	129	147	232

*1-63: EEG features (color: yellow); 64-87: EOG features (color: green); 88-111: EMG features ((color: blue)); 112-148: ECG features (color: red); 149-187: Coherence (color: gray); 188-226: Phase angles (color: gray); 227-232: Mutual information (color: gray).

features, median amplitude (P: 95) and spectral entropy (P: 110) behaved well in distinguishing R stage from W and S1.

4. Discussion

In general, a sleep study records several kinds of signals. Taking full advantage of multi-modality signals is advisable to perform a comprehensive sleep assessment. Some studies have reported that features from multi-modality signals are beneficial to the improvement of scoring accuracy[29]. The effectiveness of multi-modality signals' fusion was illustrated in Fig. 3, Table 8 and Table 10. More specifically, EEG signals contain valuable and interpretable information resembling brain activities. The changes of rhythm waves (such as delta, theta, alpha, beta) reflect the alternation of sleep stages leading to the cyclic pattern of sleep [3,7]. As demonstrated in Table 10, EEG features contributed to the discrimination of most stages. The eyes movement, recorded by EOG, may be very frequent in stage W and REM while being rare during NREM stages[16]. Therefore, EOG features have good performance in differentiating NREM stages from stage W and REM. The stage W presents the highest muscular activity in contrast to REM stage which has the lowest EMG activity[16]. As a result, EMG features are good at distinguishing stage W and stage R. Heart rate decreases with less variability in NREM stages, while increases and becomes more unstable during REM sleep[19]. ECG features are useful for differentiating stage R and W from the others, as shown in Table 10. With the deepening of sleep, the frequency of EEG signals attenuates gradually along with rare eye movements, low EMG activity and slow heart rate. There-

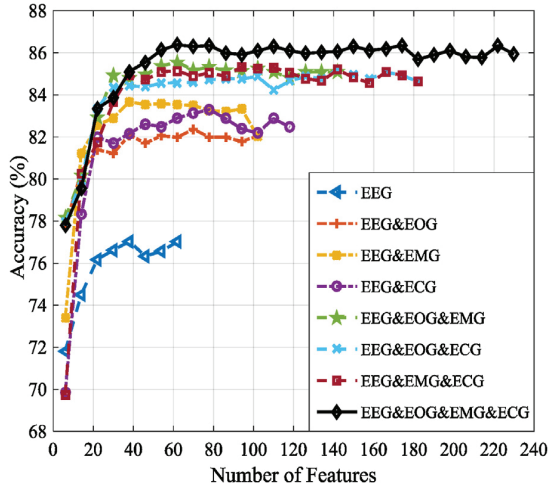


Fig. 3. Mean value of 10-fold cross-validation from different signals' fusions.

fore, four modalities of PSG signals contribute to the discriminant of NREM stage (S1, S2, S3 and S4). Features extracted from the signals of EEG, EOG, EMG and ECG can reveal sleep status from different aspects, which contributes greatly to a higher scoring accuracy in the identification of multi-modality signals.

The present study investigated 232 features (time-based, frequency-based, statistical and nonlinear features) coming from four modalities of PSG signals (EEG, EOG, EMG and ECG) to identify the most discriminative features fusion. The results point to the view that the optimal feature set can be reached by joint fusion of spectral measures (e.g. spectral edge), entropy measures (e.g. approximate entropy, spectral entropy, permutation entropy), fractal measures (e.g. Petrosian fractal dimension) and statistical time-domain features. The good performance of spectral edge has been reported in several studies [23,30,47]. In Fell and his colleagues' study [47], it claimed that delta power and spectral edge were two clinically well established measures used for monitoring sleep cycles. Meanwhile, the time-domain feature named zero-crossings is crucial in automatic sleep scoring [23,47,48]. It has been observed that the zero-crossings, as a rough estimate of average frequency, decreases as sleep goes deeper [48]. Besides, the relative spectral power of theta wave with the frequency band of 4–8 Hz (P: 45) is also important for the automatic sleep scoring since it is a marker of sleep onset [23].

It has been agreed that the classification of S1 is an enormous challenge for virtually every sleep scoring method. From the neurophysiological standpoint, S1 is a transition phase and a mixture of wakefulness and sleep, which is likely to result in the obscurity of neuronal oscillation between S2 and wakefulness (W). Besides, in the REM stage, the cortex generates 40–60 Hz gamma waves, which also occurs in the awake stage [49]. Due to the resemblance of wakefulness and REM, stage S1 is often misclassified as wakefulness or REM by both automatic sleep scoring and human technologist scorers [9], as shown in Table 7. Besides, the unbalanced instances also account for the poor classification accuracy of S1. The duration of each sleep stage varies. As a result, different stages may contain different numbers of epochs. Specifically, as described in Table 1, the number of epochs in S1 and S3 is relatively smaller when compared with other stages. It can be seen that sleep data invariably suffer from the imbalance of instances. As a result, traditional classification models are inclined to classify instances into large groups [50].

5. Conclusion

This study proposed an automatic sleep scoring method, which achieved an encouraging accuracy by employing multiple features from four modalities of polysomnography signals (EEG, EOG, EMG and ECG). In addition, different signals' fusions were investigated and the optimal features were highlighted by searching a large scale of features covering statistical characters, frequency characters, time-frequency characters, fractal characters, entropy characters and nonlinear characters. The present study concluded that the optimal feature set was a joint fusion of multiple characteristics and that the fusion of multi-modality PSG signals contributed to the increasing of classification accuracy. Furthermore, through comparing the performance of different selectors and classifiers, ReliefF and random forest classifier turned out to be reliable candidates for automatic sleep scoring.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 81471742 & 91748105), the Fundamental Research Funds for the Central Universities [DUT16JJ(G)03] in Dalian University of Technology in China, and the scholarships from China Scholarship Council (Nos. 201606060227).

References

[1] S.M. Isa, I. Wasito, A.M. Arymurthy, A. Noviyanto, Kernel dimensionality reduction on sleep stage classification using ECG signal, Int. J. Comput. Sci. Issues 8 (1) (2011) 1178–1181.
[2] S. Khalighi, T. Sousa, G. Pires, U. Nunes, Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels, Expert Syst. Appl. 40 (no. 17) (2013) 7046–7059.
[3] A. Rechtschaffen, A. Kales, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, US National Inst. Heal. Publ., Washington DC, 1968.
[4] A. Karni, D. Tanne, B. Rubenstein, J. Askenasy, D. Sagi, Dependence on REM sleep of overnight improvement of a perceptual skill, Science 265 (5172) (1994) 679–682.
[5] J.E. Kang, et al., Amyloid-β dynamics are regulated by orexin and the sleep-wake cycle, Science 326 (no. 5955) (2009) 1005–1007.
[6] A.V. Nedeltcheva, M.C. Program, C. Disorders, Metabolic effects of sleep disruption, links to obesity and diabetes, Curr. Opin. Endocrinol. Diabetes Obes. 21 (no. 4) (2014) 293–298.
[7] R.B. Berry, et al., Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events, J. Clin. Sleep. Med. 8 (no. 5) (2012) 597–619.
[8] T. Penzel, R. Conradt, Computer based sleep recording and analysis, Sleep. Med. Rev. 4 (2) (2000) 131–148.
[9] A.R. Hassan, M.I.H. Bhuiyan, A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features, J. Neurosci. Methods 271 (2016) 107–118.
[10] A.R. Hassan, M.I.H. Bhuiyan, Automatic sleep scoring using statistical features in the EMD domain and ensemble methods, Biocybern. Biomed. Eng. 36 (1) (2016) 248–255.
[11] B. Şen, M. Peker, A. Çavuşoğlu, F.V. Çelebi, A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms, J. Med. Syst. 38 (3) (2014).
[12] M. Diykh, Y. Li, P. Wen, EEG sleep stages classification based on time domain features and structural graph similarity, IEEE Trans. Neural Syst. Rehabil. Eng. 24 (11) (2016) 1159–1168.
[13] S.F. Liang, C.E. Kuo, Y.H. Hu, Y.H. Pan, Y.H. Wang, Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models, IEEE Trans. Instrum. Meas. 61 (6) (2012) 1649–1657.
[14] S.I. Dimitriadis, C. Salis, D. Linden, A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates, bioRxiv 7 (July4) (2017) 1–41.
[15] J. Shi, X. Liu, Y. Li, Q. Zhang, Y. Li, S. Ying, Multi-channel EEG-based sleep stage classification with joint collaborative representation and multiple kernel learning, J. Neurosci. Methods 254 (2015) 94–101.
[16] E. Estrada, H. Nazeran, J. Barragan, J.R. Burk, E.A. Lucas, K. Behbehani, EOG and EMG: Two important switches in automatic sleep stage classification, Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc. (2006) 2458–2461.
[17] S. Özşen, Classification of sleep stages using class-dependent sequential feature selection and artificial neural network, Neural Comput. Appl. 23 (5) (2013) 1239–1250.

[18] J. Zhang, Y. Wu, J. Bai, F. Chen, Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers, Trans. Inst. Meas. Control. 38 (4) (2016) 435–451.
[19] F. Ebrahimi, S.K. Setarehdan, H. Nazeran, Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs, Biomed. Signal. Process. Control. 18 (2015) 69–79.
[20] I. Hermawan, M.S. Alvissalim, M.I. Tawakal, W. Jatmiko, An integrated sleep stage classification device based on electrocardiograph signal, Adv. Comput. Sci. Inf. Syst. (ICACSIS), 2012 Int. Conf. (2012) 37–41.
[21] B. Yılmaz, et al., Sleep stage and obstructive apneaic epoch classification using single-lead ECG, Biomed. Eng. Online 9 (39) (2010) 39.
[22] K. Šušmáková, A. Krakovská, Discrimination ability of individual measures used in sleep stages classification, Artif. Intell. Med. 44 (3) (2008) 261–277.
[23] A. Krakovská, K. Mezeiová, Automatic sleep scoring: A search for an optimal combination of measures, Artif. Intell. Med. 53 (1) (2011) 25–33.
[24] A.A. Gharbali, S. Najdi, J.M. Fonseca, Investigating the contribution of distance-based features to automatic sleep stage classification, Comput. Biol. Med. 96 (March) (2018) 8–23.
[25] R.S. Vasan, D. Levy, PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals, Heart Fail. (2000) 2118–2121.
[26] M. Giovanni Terzano, et al., Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep, Sleep. Med. 3 (2) (2002) 185.
[27] M.G. Terzano, D. Mancia, M.R. Salati, G. Costani, A. Decembrino, L. Parrino, The cyclic alternating pattern as a physiologic component of normal NREM sleep, Sleep 8 (2) (1985) 137–145.
[28] C. Vidaurre, N. Krämer, B. Blankertz, A. Schlögl, Time domain parameters as a feature for EEG-based brain-computer interfaces, Neural Netw. 22 (no. 9) (2009) 1313–1319.
[29] R. Boostani, F. Karimzadeh, M. Nami, A comparative review on sleep stage classification methods in patients and healthy individuals, Comput. Methods Programs Biomed. 140 (2017) 77–91.
[30] S.A. Imtiaz, E. Rodriguez-Villegas, A low computational cost algorithm for REM sleep detection using single channel EEG, Ann. Biomed. Eng. 42 (11) (2014) 2344–2359.
[31] M.K. Uçar, M.R. Bozkurt, C. Bilgin, K. Polat, Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques, Neural Comput. Appl. (2016) 1–16.
[32] R. Esteller, J. Echauz, T. Tcheng, B. Litt, B. Pless, Line length: an efficient feature for seizure onset detection, Annu. Int. Conf. IEEE Eng. Med. Biol. 2 (January 2015) (2001) 1707–1710.
[33] R. Acharya, U.O. Faust, N. Kannathal, T. Chua, S. Laxminarayan, Non-linear analysis of EEG signals at various sleep stages, Comput. Methods Programs Biomed. 80 (1) (2005) 37–45.
[34] S. Vorobyov, A. Cichocki, Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis, Biol. Cybern. 86 (4) (2002) 293–303.
[35] X.F. Liu, Y. Wang, Fine-grained permutation entropy as a measure of natural complexity for time series, Chin. Phys. B 18 (7) (2009) 2690–2695.
[36] J. Pan, W.J. Tompkins, A Real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. BME-32 (3) (1985) 230–236.
[37] A. Noviyanto, S.M. Isa, I. Wasito, A.M. Arymurthy, Selecting features of single lead ecg signal for automatic sleep stages classification using correlation-based feature subset selection, Int. J. Comput. Sci. Issues 8 (1) (2011) 1178–1181.
[38] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[39] M. Robnik-Siknja, I. Kononeko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23–69.
[40] Y. Lei, Z. He, Y. Zi, A new approach to intelligent fault diagnosis of rotating machinery, Expert Syst. Appl. 35 (4) (2008) 1593–1600.
[41] C.M. Fonseca, P.J. Fleming, Genetic algorithms for multiobjective optimization: formulation, discussion and generalization, Icga 93 (July) (1993) 416–423.
[42] A.W. Whitney, A direct method of nonparametric measurement selection, IEEE Trans. Comput. C-20 (9) (1971) 1100–1103.
[43] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[44] B. Şen, E. Uçar, D. Delen, Predicting and analyzing secondary education placement test scores: A data mining approach, Expert Syst. Appl. 39 (10) (2012) 9468–9476.
[45] S. Isa, M. Fanany, W. Jatmiko, A. Murini, Feature and model selection on automatic sleep apnea detection using ECG, Int. Conf. Comput. Inf. Syst. (February) (2010) 357–362.
[46] C. Cortes, V. Vapnik, Support vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[47] J. Fell, J. Röschke, K. Mann, C. Schäffner, Discrimination of sleep stages: A comparison between spectral and nonlinear EEG measures, ElectroEncephalogr. Clin. Neurophysiol. 98 (no. 5) (1996) 401–410.
[48] F. Machado, Automatic Sleep Staging Based on Classification Methods, 2015.
[49] J. Horne, Why REM sleep? Clues beyond the laboratory in a more challenging world, Biol. Psychol. 92 (2) (2013) 152–168.
[50] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost: A hybrid approach to alleviating class imbalance, IEEE Trans. Syst. 40 (1) (2010) 13.

II

AN AUTOMATIC SLEEP SCORING TOOLBOX: MULTIMODALITY OF POLYSOMNOGRAPHY SIGNALS' PROCESSING

by

Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi & Fengyu Cong 2019

In Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019), pp.301-309.

DOI: 10.5220/0007925503010309

An Automatic Sleep Scoring Toolbox: Multi-modality of Polysomnography Signals' Processing

Rui Yan[1, 2], Fan Li[3], Xiaoyu Wang[2], Tapani Ristaniemi[1] and Fengyu Cong[1, 2]
[1]*Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland*
[2]*School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China*
[3]*School of Information Science and Engineering, Dalian Polytechnic University, 116034, Dalian, China*

Keywords: Polysomnography, Multi-modality Analysis, MATLAB Toolbox, Automatic Sleep Scoring.

Abstract: Sleep scoring is a fundamental but time-consuming process in any sleep laboratory. To speed up the process of sleep scoring without compromising accuracy, this paper develops an automatic sleep scoring toolbox with the capability of multi-signal processing. It allows the user to choose signal types and the number of target classes. Then, an automatic process containing signal pre-processing, feature extraction, classifier training (or prediction) and result correction will be performed. Finally, the application interface displays predicted sleep structure, related sleep parameters and the sleep quality index for reference. To improve the identification accuracy of minority stages, a layer-wise classification strategy is proposed according to the signal characteristics of sleep stages. The context of the current stage is taken into consideration in the correction phase by employing a Hidden Markov Model to study the transition rules of sleep stages in the training dataset. These transition rules will be used for logic classification results. The performance of proposed toolbox has been tested on 100 subjects with an average accuracy of 85.76%. The proposed automatic scoring toolbox would alleviate the burden of the physicians, speed up sleep scoring, and expedite sleep research.

1 INTRODUCTION

Sleep covers almost one-third of human lifespan. Adequate and high-quality sleep is vital to our physical and mental well-being (Pagel and Pandi-Perumal, 2014). However, and likely because of our ephemeral lifestyle in modern society, sleep disorder complaints increase dramatically among people. Assessing sleep behaviour and analysing the sleep structure, therefore, become more and more crucial. Up to now, the conventional visual scoring method is still the main method in most clinical and sleep research labs worldwide.

Visual scoring, mainly based on the rules of Rechtschaffen & Kales (R&K) (Rechtschaffen and Kales 1968) and the recently updated American Academy of Sleep Medicine rules (AASM) (Berry et al., 2012), requires at least one registered sleep technologist (RST) who has sufficient expertise and experience in sleep scoring. Generally, the annotation of 8-h recording requires approximately 2-4 hours (Hassan and Bhuiyan, 2016a), which is rather time-consuming. Besides, visual scoring in some degree is subjective, as the inter-scorer reliability among trained technologists is less than 90% (Danker-Hopfe et al., 2009). In contrast, automatic sleep scoring has demonstrated advantages of cost-effective and preferable scoring performance.

Electroencephalogram (EEG) signals are mainly used in automatic sleep scoring since they contain valuable and interpretable information resembling brain activities (Boostani et al., 2017). According to the morphological characteristics of EEG signals, sleep EEG waves are mainly composed by α wave, β wave, θ wave and δ wave, K complex, sleep spindles and saw-tooth (Niedermeyer and da Silva, 2005). These rhythm waves form the foundation of sleep scoring. Some studies (Hassan et al., 2015; Hassan and Bhuiyan, 2016b) tried to extract statistical and spectral features from these rhythm waves to perform an automatic sleep scoring. Cross frequency coupling estimated between rhythm waves also showed high classification accuracy (Dimitriadis et al., 2018). Instead of traditional linear features, multiscale entropy and autoregressive models for single-channel EEG were employed in Liang et.al' s study, obtaining a good scoring performance (Liang et al., 2012).

Sleep is a complex process involving multiple

Yan, R., Li, F., Wang, X., Ristaniemi, T. and Cong, F.
An Automatic Sleep Scoring Toolbox: Multi-modality of Polysomnography Signals' Processing.
DOI: 10.5220/0007925503010309
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 301-309
ISBN: 978-989-758-378-0

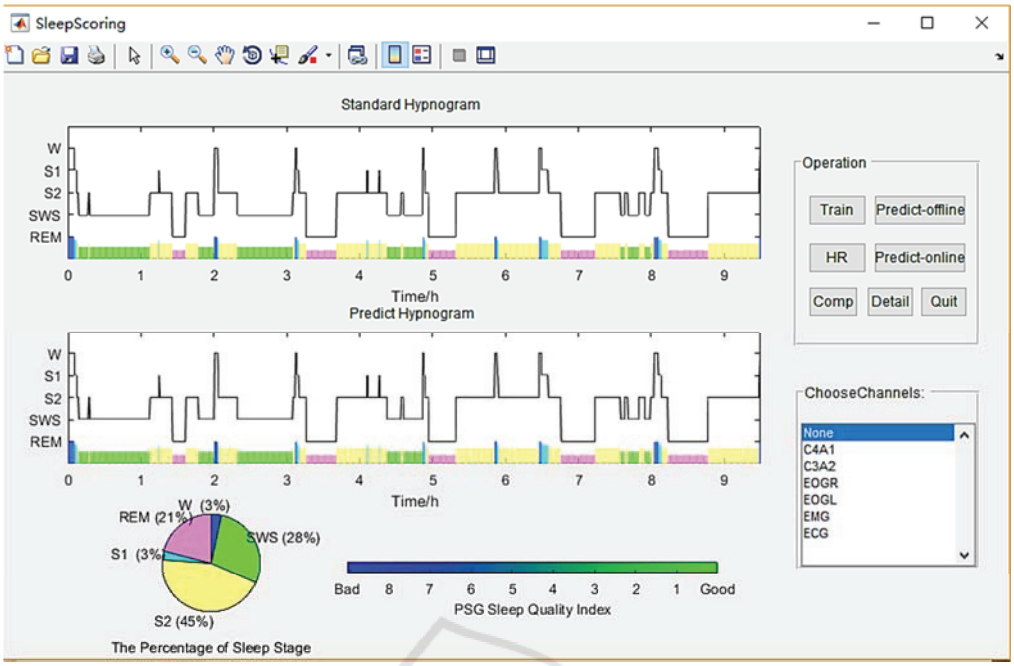


Figure 1: The interface of sleep scoring toolbox.

organs. Signals recorded from different physical areas change with the sleep cycle. The multi-modality signals' contribution to sleep scoring has been explored in several studies (Gharbali et al., 2018; Yan et al., 2019; Šušmáková and Krakovská, 2008). Özşen concluded that as the sleep deepen, the frequency of EEG signals attenuated gradually, along with rare eye movements, low electromyography (EMG) activity and slow heart rate (Özşen, 2013). Ebrahimi and his colleagues found that under the control of parasympathetic nervous system and sympathetic nervous system, cardiovascular and respiratory behaviours fluctuated with the alternation of sleep stage (Ebrahimi et al., 2015). It has demonstrated that features from multi-modality signals were beneficial to the improvement of scoring accuracy (Boostani et al., 2017).

Although there are many studies on automatic sleep scoring, the available software and toolbox is limited. Given that, this study aims to develop an automatic sleep scoring toolbox with the capability of multi-signal processing, see Figure 1. The main contributions of this work are presented as following:

a) An automatic sleep scoring toolbox is proposed which supports multiple sleep signals and two data formats.
b) An interactive interface is provided which allows the user to select the number of target classes, change signal types and visualize various analysis results.
c) A layer-wise classification strategy is proposed which can significantly improve the classification accuracy of minority stages without compromising the accuracy of other classes.
d) A correction procedure is proposed to make classification results logical.

The article is organized as follows: Section 2 explains the details of experimental data and methodology of this study. Section 3 demonstrates the performance of proposed toolbox. Section 4 provides discussions of results and limitations of this study. Finally, section 5 gives conclusions of this paper.

2 MATERIALS AND METHODS

2.1 System Overview

The proposed toolbox consists of a training module, an offline prediction module, an online prediction module and several parameter panels, as shown in Figure 1. Their functions are briefly described in the following lines. The specific model structure will be introduced in detail in section 2.5.

Training Module: The objective of the training module is to train a classifier based on the user's selection. The user can choose signal types and the number of target stages as required. The software automatically performs signal pre-processing, feature

extraction and classifier training. The output of this module is a trained model which can be used to predict sleep structures.

Prediction Module: The aim of this module is to predict sleep structure based on the predefined model or user-specified model. The module automatically checks if the user has trained a model, and allows the user to determine if the predefined model is needed. Once the model selected, the module automatically processes the test data based on model parameters. Finally, the application interface displays the predicted sleep structure, related sleep parameters and a sleep quality index as a reference to sleep quality. If a hypnogram (e.g., labels scored by RST) is available for the test data, the interface would display both the hypnogram and predicted labels together, and highlight the disagreement by pressing the button named "Comp".

Online Prediction Module: The module is similar to the offline prediction process except for the real-time updating results. The module can be connected to a sleep monitoring device in order to realize the real-time analysis of sleep signals and to visualize sleep structures. The updated sleep signal will be saved as a TXT file in storage.

2.2 Description of Experiment Data

The sleep data for this investigation was provided by the Sleep Heart Health Study (SHHS) database. We used only the first round (SHHS-1) due to its wide age range. The recordings employed in this study were selected by considering a Respiratory Disturbance Index 3 Percent (RDI3P) < 5 to have near-normal characteristics. Moreover, subjects did not use beta-blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease, or history of stroke. Given that, a total number of 100 subjects were selected with the total duration of 816 hours and 43 minutes. The age of subjects ranged from 40 to 54 years, with a mean value of 47 years and a standard deviation of 4.3 years. Each record was scored by the experienced research assistant or sleep technologist according to the R&K rules. The sleep recordings were segmented into 30-second per epoch and labelled as wakefulness (W), non-rapid eye movement stage (NREM, containing S1, S2, S3 and S4) and rapid eye movement stage (REM). The deepest NREM stage, namely S3 and S4, were collectively referred to as "slow wave sleep" (SWS), based on a prevalence of low-frequency oscillations (Berry et al., 2012). A detailed description of SHHS was given in the study (Quan et al., 1997) .

2.3 Pre-processing

For the predefined model and the following experiments, four modalities of polysomnography (PSG) signals were considered: EEG channels (C4-A1 and C3-A2, following the 10-20 international electrode placement system), two electrooculography (EOG) channels (named: ROC, LOC), one submental electromyography (EMG) channel and one electrocardiography (ECG) channel. All the aforementioned signals were fully included within the evaluation process without discarding any recorded segments, thereby to have a near-clinical situation.

In order to remove noise and artefacts, a notch filter, a high-pass filter with a cut-off frequency of 0.3Hz and a low-pass filter with a cut-off frequency of 30Hz were applied to the signals of EEG, EOG and ECG. In terms of EMG, a notch filter, a high-pass filter with a cut-off frequency of 10Hz and a low-pass filter with a cut-off frequency of 75Hz were performed. The whole night recordings were smoothed by its mean value ±5×standard deviation to remove the outliers. In order to eliminate individual differences, the sleep signals were normalized to [-100, 100]. Afterwards, all the signals were divided into 30-second epochs, each epoch corresponding to a single sleep stage.

2.4 Feature Extraction

The features, employed in this study, involves a variety of traditional and modern characteristics serving as distinctive markers for various psycho-physiological states. They are summarized in Table 1. Some of the parameters are introduced in the following, and the others can be found in Yan et al.'s research (Yan et al. 2019).

2.4.1 Time Domain Parameters

Some statistical parameters, such as minimum value, maximum value, standard deviation, arithmetic mean, variance, skewness, kurtosis and median are derived from signal segments. These statistical parameters are good indicators of the amplitude and distribution of time series (Şen et al. 2014). Percentile analysis is known as the most effective time domain measures for EEG signals (Boostani et al. 2017). Hjorth parameters (i.e., activity, mobility and complexity) represent the signal power, the mean frequency and frequency changes (Vidaurre et al. 2009).

Table 1: Parameter list.

Type	Feature Name
Statistical measures	Minimum Value (MinV), Maximum Value (MaxV), Arithmetic Mean(AM), Median(M), Standard Deviation (SD), Variance(V), Skewness(S), Kurtosis(K), The 5th Percentile (Pre5), The 25th Percentile (Pre25), The 75th Percentile (Pre75), The 95th Percentile (Pre95), Hjorth Parameters（HA, HM, HC）, Zero-Crossing(ZC)
Spectral measures	Power Spectral Density(PSD), Mean Value of PSD (mPSD), Median Value of PSD (mdPSD), Power Ratio(PR), Absolute and Relative Spectral Power (APSD, RPSD), Brain Rate (BR), Spectral Centroid (Sc), Spectral Width (Sw), Spectral Asymmetry (Sa), Spectral Flatness (Sk), Spectrum Flatness (Sf), Spectral Slope (Ss), Spectral Decrease (Sd), Edge_D, Spectral Edge Frequency at 90% and 50%
Nonlinear measures	Mean teager energy (MTE), Mean Energy (E), Mean curve length (CL), SecD, The 4th Power,
Fractal measures	Petrosian fractal dimension (PFD)
Entropy measures	Spectral Entropy(SpE)
Mutual measures	Coherence

2.4.2 Spectral Features

The calculation of spectral measures is based on Fourier transform using hamming window in the time domain. The following spectral measures are considered.

Power spectral density is calculated based on the following formula. Meanwhile, its mean value and median value are also considered.

$$PSD = F(\omega) \times F^*(\omega)/N \quad (1)$$

where ω is the frequency, * representing the complex conjugate, and N is the length of time series.

Spectral edge is defined as the frequencies corresponding to 90% and 50% of the total spectral power (Imtiaz and Rodriguez-Villegas 2014). The difference between the two frequencies (edge_D) is also considered.

$$\sum_{f=f_{min}}^{edge} P(f) = p \sum_{f=f_{min}}^{30Hz} P(f) \quad (2)$$

where p is equal to 0.9 or 0.5, f_{min} is 0.3Hz in terms of EEG, EOG and ECG, and 10Hz in EMG.

Absolute and relative spectral power are obtained from seven frequency bands of EEG, namely, 0.3-4Hz (delta), 2-3.9Hz (K complex), 2-6Hz (saw-tooth), 4-8Hz (theta), 8-12Hz (alpha), 14-30Hz (beta), and 12-16Hz (spindle). Absolute spectral power is spectral power within the specific frequency bands. The relative value is defined as the ratio of the absolute value to the total spectral power. The total spectral powers of EEG, EOG and ECG signals are computed within the range of 0.3-30Hz, and 10-30Hz for EMG signals.

Power ratios are computed based on absolute spectral powers in aforementioned frequency bands. The following power ratios are computed: delta/theta, delta/alpha, delta/beta, theta/alpha, theta/beta, alpha/beta, alpha/(theta + delta), delta/(theta + alpha) and theta/(beta + delta).

Brain rate estimates the EEG mean frequency weighted over the brain spectrum distribution (Pop-Jordanova and Pop-Jordanov 2005).

$$BR = \sum_i^M f_i \times P_i \,/\, \sum_i^M P_i \quad (3)$$

where M is the number of frequency bins, i the sub-band, P_i the power of the spectral distribution corresponding to frequency band i, and f_i is the frequency at bin i.

Spectral centroid is defined as the frequency-weighted sum of the magnitude spectrum of the signal normalized by its unweighted sum, indicating the

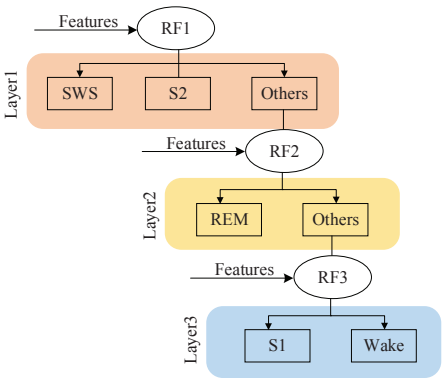


Figure 2: Layer-wise classifier.

location of the spectrum centre (Hassan et al. 2015). Spectral width is the wavelength interval over which the magnitude of all spectral components is equal to or greater than a specified fraction of the magnitude of the component having the maximum value. Spectral asymmetry represents the asymmetry in the distribution of the spectrum of eigenvalues of an operator. Spectral flatness, measured in decibels, provides a way to quantify how noise-like a sound is (Dubnov 2004). Spectrum flatness defines the planeness properties from an audio signal's spectrum, which shows how the power spectrum of a signal deviates from a frequency of a flat shape (Lazaro et al. 2017). Spectral slope is a measure of the slope of the spectral shape (Hassan et al. 2015). The steepness of the decrease of the spectral envelope of the signal with respect to its frequency is defined as spectral decrease (Hassan et al. 2015). The detailed definition of these parameters can be found in Chen et al.'s study (Chen et al. 2018).

2.5 Classification

It is well-known that the distribution of epochs among sleep stages are highly imbalanced. Unfortunately, the traditional classifier is kind of sensitive to the distribution of data sets. When instances of one class in the training set vastly outnumber the instances of other classes, the classifier inclines to classify instances as belonging to the majority class and ends up creating suboptimal classification models in the process (Hassan and Bhuiyan 2016c). After studying the characteristics of sleep stages, we find that the REM, S1 and wakefulness present a certain similarity leading to misclassification. For example, the level of brain activity and eye movements increase in REM stage which is similar to the waking period. In addition, S1 is a transition phase of wakefulness and sleep, along with the ambiguous neuronal oscillation, that makes the detection of S1 is the most problematic of the sleep stages. For S2 and SWS, with the deepening of sleep, the activity levels of various organs decrease to some extent.

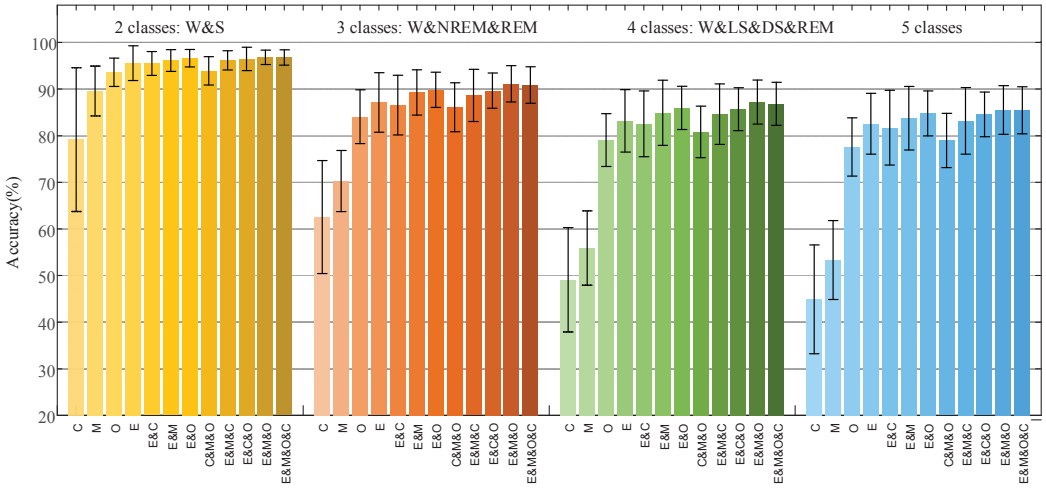
Based on these characteristics of the sleep stages, we develop a layer-wise classification strategy (See Figure 2) which is used in this toolbox to train and predict sleep structures. The strategy uses three random forest classifiers. The first layer is a multi-class classifier dividing the sleep sequences into SWS, S2, and others. The second layer is a two-class classifier, which aims to distinguish the REM stage according to its lowest EMG activity and obvious eye movements. The third layer discriminates the characteristics of S1 and awake stage. Experiments have confirmed that the structure can significantly improve the recognition accuracy of the minority sleep stages, such as S1, without significantly reducing the classification accuracy of other classes.

2.6 Result Correction

Studies have found that sleep transition is not a random process. However, the traditional classifier can only give its decision according to the information of the current stage, but can't remember the context. Therefore, a correction process is applied to classification results. Firstly, the Hidden Markov Model is used to learn the transition rules among sleep stages in the training data. Then, the correction rule can be derived according to these transition rules and some natural characters of sleep. These rules refer to the epochs prior to and posterior to the current epoch. The development of correction rules is inspired by the studies of Liang et al. (Liang et al. 2012) and Li et al. (Li et al. 2018). More specifically, the stage sequences, like $[S_{i-1}, S_i, S_{i+1}]$, are smoothed by the rules proposed in the study (Liang et al. 2012) to correct some sudden changes in predicted results. For the stage sequences that do not meet the aforementioned smooth rules, the transition rules derived from Hidden Markov Model will be used to analyse the rationality of the stage transitions.

Table 2: Sleep parameters and its definition.

Sleep Parameters	Definition
Time in bed	From light off to getting up
Sleep period time	From sleep onset to sleep end, in minutes
Sleep efficiency	Total sleep time / Bed time
Sleep onset latency	From recording start to sleep onset, in minutes
REM latency	From sleep onset to the occurrence of the first REM period, in minutes
Stage shifts/h	Number of sleep stage shifts after sleep onset per hour
Waking times	Number of awakenings after sleep onset per hour
Waking time	Wakefulness after sleep onset, percentage of sleep period time
Number of REM	Number of REM periods
Stage time	Specific stage time after sleep onset, in minutes
Stage percentage	Specific stage time in percentage of sleep period time



*C: single-modality ECG; M: single-modality EMG; O: single-modality EOG; E: single-modality EEG; &: combination signals; 2 classes: wakefulness and sleep (W&S); 3 classes: wakefulness, non-rapid eye movement sleep and rapid eye movement sleep (W&NREM&REM); 4 classes: wakefulness, light sleep (containing S1 and S2), deep sleep (SWS) and rapid eye movement sleep (W&LS&DS &REM); 5 classes: W, S1, S2, SWS and REM

Figure 3: The classification accuracy for different signal fusions and target class.

Table 3: Selected features for distinguishing specific pair of sleep stages.

Sleep stages		SWS-S2	SWS-S1	SWS-R	SWS-W	S2-S1	S2-R	S2-W	S1-R	S1-W	R-W
Top 15 features	Top1	C.Per25	C.Per25	C.Per25	M.ZC	C.PFD	O.ZC	C.ZC	O.ZC	E.SPE	O.ZC
	Top2	M.ZC	C.Per75	C.Per75	C.ZC	E.PSD	O.Sw	M.ZC	O.Per75	C.PSD	O.Per75
	Top3	C.Per75	M.ZC	C.Per95	M.PFD	C. mPSD	E.PR	E.SPE	E.K	O.Ss	O.Per25
	Top4	C.ZC	C.Per95	M.Per25	C.Per25	E.PR	E.PFD	C.PFD	O.Per25	O.Sf	O.PFD
	Top5	C.Per5	M.Sf	M.ZC	C.PFD	C.HM	O.edge90	O.Sf	O.PFD	E.MTE	M.ZC
	Top6	C.Per95	E.PSD	C.PSD	E.PSD	M.PFD	C. mPSD	M.PFD	O.HC	O.MTE	C.ZC
	Top7	C.K	C.Per5	E.Per95	C.Per75	O.Sd	C.K	O.Ss	O.HM	C.ZC	E.Per5
	Top8	O.K	M.Per25	C.Per5	M.Sf	M.Sf	O.Sd	E.PFD	O.SPE	C.Ss	O.Ss
	Top9	E.Per5	O.PFD	O.Power4	O.edge.D	E.PFD	O.PFD	C.HM	O.K	O.MaxV	O.RPSD
	Top10	M.Per25	E.RPSD	M.Per75	O.Sk	O.BR	E.S	C.PSD	O.K	E.D	O.CL
	Top11	M.K	C.ZC	E.PR	E.PR	O.PFD	O.BR	E.edge90	O.CL	C.Sf	C.PSD
	Top12	O.Per75	C.HM	C.K	E. mdPSD	C.ZC	C. mdPSD	E.PR	O.RPSD	O.RPSD	E.CL
	Top13	E.Per25	M.HM	C.HA	M.HM	O.edge90	O.HC	C.R R	O.SecD	E.MaxV	O.SecD
	Top14	M.HM	E.RPSD	O.Sf	E.PR	C.mPSD	O.S	C.mPSD	O.MaxV	M.ZC	O.Sf
	Top15	E.K	O.edge.D	C.PFD	E.PR	O.Sw	E.PR	M.HM	O.Sc	C.PFD	E.Sc

EEG features (colour: yellow); EOG features (colour: green); EMG features (colour: red); ECG features (colour: blue).

2.7 PSG Sleep Quality Index

Usually, sleep quality is evaluated by the standardized questionnaire, such as the Pittsburgh Sleep Quality Index (PSQI), the Berlin Questionnaire, and so on. These self-report questionnaires are subjective, and can be easily exaggerated or minimized by the person completing them. Furthermore, some items of questionnaires are challenging to self-evaluation. For example, the PSQI needs to evaluate the time it takes to fall asleep and the actual sleep time per night. Some papers claimed that the correspondence between the objective measurement and a person's subjective assessment of the sleep quality is surprisingly small, if existent (Sohn et al., 2012). In order to overcome the uncertainty of subjective assessment, the toolbox proposed a sleep quality index. The algorithm will calculate various sleep parameters (summarized in Table 2) according to the predicted sleep stages. Based on these sleep parameters, PSG sleep quality index is statistically calculated and displayed in a bar in the lower-right corner of the interface. Detailed sleep parameters can be obtained by pressing the button "Detail".

3 PERFORMANCE ASSESSMENT

3.1 Influence of Signal Types

In order to explore the relationship between signal types and classification accuracy, we performed a greedy search for several signal fusions referring to different target classes. The result was shown in Figure 3 where the column denoted the mean accuracy of 10-fold cross-validation and the bars represented the standard deviation. Four categories were considered, highlighted in different colours in Figure 3. For each category, twelve signal fusions were listed along X-axis where signals' names were abbreviated to its middle letter.

Figure 3 depicted the uncertainty or variation of classification accuracy under each condition. Collectively, with the enrichment of signal types, the mean value of accuracy increased, and the uncertainty decreased to some extent. More specifically, Figure 3 indicated that the required signal types varied with the number of target classes. If sleep recordings were classified into two classes, namely wakefulness (W) and sleep (S), all considered signal fusions gave satisfactory results. With the increasing number of target classes, the number of required signals increased accordingly.

From the perspective of signal types, the signal fusions containing EEG signals showed better identification accuracy, indicating a crucial role of EEG signals in sleep scoring. Furthermore, the discriminative information provided by ECG and EMG channels was inferior to that from EEG and EOG signals.

3.2 Feature Evaluation

To further elucidate the contributions of features and signals, the important features, measured by their contribution to distinguishing each pair of sleep stages, were derived from random forest classifier. The top 15 features were shown in Table 3, where features sorted in descending order of discriminative capability. As can be seen from Table 3, the features from EEG contributed to the recognition of most stages. Meanwhile, ECG features demonstrated its contribution to the discrimination of SWS from the others. For EOG signal, its features were good at distinguishing REM stage and wakefulness. In terms of feature types, the top 15 features indicated that the optimal feature subset was a fusion of statistical measures (e.g. Percentiles, Hjorth parameters, Zero-Crossing), spectral measures (e.g. spectral edge, power spectral density), entropy measures (e.g. spectral entropy), fractal measures (e.g. Petrosian fractal dimension) and nonlinear measures (e.g. mean curve length, the 4th Power).

4 DISCUSSION

PSG, the golden standard for measuring sleep qualitatively, is a traditional technology which is time-consuming and has barely changed over the years. Burgeoning public interest in sleep quality improves a strong impetus for a robust, easily implemented and rapid sleep scoring system. Limited toolbox or software is available for automatic sleep scoring, although there are many theoretical researches in this field. In previous studies, some portable devices were developed based on ECG and respiration. For example, Hermawan et al. (Hermawan et al., 2012) developed a real-time sleep stage classification device which classified sleep recordings into 2 stages (wakefulness and sleep) with an average precision of 0.941. Recently, some deep learning-based scoring tools sprouted out, such as SLEEPNET (Biswal et al., 2017) and SeqSleepNet (Phan et al., 2019). The classification accuracy of these deep learning-based tools was about 0.85 with the support of tremendous training data and highly configured computer (like GPU or server).

Compared with previous studies, the proposed toolbox provides comparable precision and greater freedom. The toolbox, based on MATLAB, allows users to select the available signal types and the number of target classes according to their condition and need. Meanwhile, it supports two popular data formats (MAT file and EDF file) that make data transfer easy. This offline prediction module is helpful for researchers, especially the newcomers in this field, to accelerate their understanding of sleep structures. It can also be used in clinic to speed up the annotation of PSG records, thus alleviating the burden of the physicians. The online prediction module provides the potential to control sleep tasks automatically by combining the toolbox with sleep experiments.

Even though our results are encouraging, our model still has several limitations. One of them is that the performance of proposed toolbox is affected by the data property. As our model learns from training data, it might not perform well when the trained model is applied to the data with different properties. For example, a scoring model trained by healthy subjects may not perform well for the analysis of patients' sleep structure. To achieve better results in

that condition, the model might have to be re-trained or fine-tuned.

5 CONCLUSIONS

This paper proposed an automatic sleep scoring toolbox that supported four types of sleep signals and two data formats. The toolbox provided an interface for user-friendly operation. Sleep recordings could be automatically analysed to reveal multiple sleep parameters and sleep quality index. A layer-wise classification strategy was proposed to improve the classification accuracy of minority stages. In addition, a Hidden Markov Model was used to make classification results logic. Compared with manual scoring, the proposed automatic scoring toolbox is cost-effective, which would alleviate the burden of the physicians, speed up sleep scoring and expedite sleep research.

ACKNOWLEDGEMENTS

The authors would like to thank the SHHS for providing the polysomnographic data. This work was supported by the scholarships from China Scholarship Council (Nos. 201606060227).

REFERENCES

Berry, R.B., Budhiraja, R., Gottlieb, D.J., Gozal, D., Iber, C., Kapur, V.K., Marcus, C.L., Mehra, R., Parthasarathy, S., Quan, S.F. and Redline, S., 2012. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine*, *8*(05), pp.597-619.

Biswal, S., Kulas, J., Sun, H., Goparaju, B., Westover, M.B., Bianchi, M.T. and Sun, J., 2017. SLEEPNET: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262*.

Boostani, R., Karimzadeh, F. and Nami, M., 2017. A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine*, *140*, pp.77-91.

Chen, T., Huang, H., Pan, J. and Li, Y., 2018, May. An EEG-based brain-computer interface for automatic sleep stage classification. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)* (pp. 1988-1991). IEEE.

Danker-hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S. and Saletu, B., 2009. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of sleep research*, *18*(1), pp.74-84.

Dimitriadis, S.I., Salis, C. and Linden, D., 2018. A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates. *Clinical Neurophysiology*, *129*(4), pp.815-828.

Dubnov, S., 2004. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, *11*(8), pp.698-701.

Ebrahimi, F., Setarehdan, S.K. and Nazeran, H., 2015. Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs. *Biomedical Signal Processing and Control*, *18*, pp.69-79.

Gharbali, A.A., Najdi, S. and Fonseca, J.M., 2018. Investigating the contribution of distance-based features to automatic sleep stage classification. *Computers in biology and medicine*, *96*, pp.8-23.

Hassan, A.R., Bashar, S.K. and Bhuiyan, M.I.H., 2015, August. On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2238-2243). IEEE.

Hassan, A.R. and Bhuiyan, M.I.H., 2016. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *Journal of neuroscience methods*, *271*, pp.107-118.

Hassan, A.R. and Bhuiyan, M.I.H., 2016. Automatic sleep scoring using statistical features in the EMD domain and ensemble methods. *Biocybernetics and Biomedical Engineering*, *36*(1), pp.248-255.

Hassan, A.R. and Bhuiyan, M.I.H., 2016. Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. *Biomedical Signal Processing and Control*, *24*, pp.1-10.

Hermawan, I., Alvissalim, M.S., Tawakal, M.I. and Jatmiko, W., 2012, December. An integrated sleep stage classification device based on electrocardiograph signal. In *2012 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 37-41). IEEE.

Imtiaz, S.A. and Rodriguez-Villegas, E., 2014. A low computational cost algorithm for REM sleep detection using single channel EEG. *Annals of biomedical engineering*, *42*(11), pp.2344-2359.

Lazaro, A., Sarno, R., Andre, R.J. and Mahardika, M.N., 2017, October. Music tempo classification using audio spectrum centroid, audio spectrum flatness, and audio spectrum spread based on MPEG-7 audio features. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*(pp. 41-46). IEEE.

Li, X., Cui, L., Tao, S., Chen, J., Zhang, X. and Zhang, G.Q., 2017. Hyclasss: A hybrid classifier for automatic

sleep stage scoring. *IEEE journal of biomedical and health informatics*, *22*(2), pp.375-385.

Liang, S.F., Kuo, C.E., Hu, Y.H. and Cheng, Y.S., 2012. A rule-based automatic sleep staging method. *Journal of neuroscience methods*, *205*(1), pp.169-176.

Liang, S.F., Kuo, C.E., Hu, Y.H., Pan, Y.H. and Wang, Y.H., 2012. Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. *IEEE Transactions on Instrumentation and Measurement*, *61*(6), pp.1649-1657.

Niedermeyer, E. and da Silva, F.L. eds., 2005. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

Özşen, S., 2013. Classification of sleep stages using class-dependent sequential feature selection and artificial neural network. *Neural Computing and Applications*, *23*(5), pp.1239-1250.

Pagel, J.F. and Pandi-Perumal, S.R. eds., 2014. *Primary Care Sleep Medicine: A Practical Guide*. Springer.

Phan, H., Andreotti, F., Cooray, N., Chén, O.Y. and De Vos, M., 2019. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Pop-Jordanova, N. and Pop-Jordanov, J., 2005. Spectrum-weighted EEG frequency ("brain-rate") as a quantitative indicator of mental arousal. *Prilozi*, *26*(2), pp.35-42.

Quan, S.F., Howard, B.V., Iber, C., Kiley, J.P., Nieto, F.J., O'Connor, G.T., Rapoport, D.M., Redline, S., Robbins, J., Samet, J.M. and Wahl, P.W., 1997. The sleep heart health study: design, rationale, and methods. *Sleep*, *20*(12), pp.1077-1085.

Rechtschaffen, A. & Kales, A., 1968. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Washington DC: US National Institute of Health Publication*.

Şen, B., Peker, M., Çavuşoğlu, A. and Çelebi, F.V., 2014. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. *Journal of medical systems*, *38*(3), p.18.

Sohn, S.I., Kim, D.H., Lee, M.Y. and Cho, Y.W., 2012. The reliability and validity of the Korean version of the Pittsburgh Sleep Quality Index. *Sleep and Breathing*, *16*(3), pp.803-812.

Šušmáková, K. and Krakovská, A., 2008. Discrimination ability of individual measures used in sleep stages classification. *Artificial intelligence in medicine*, *44*(3), pp.261-277.

Vidaurre, C., Krämer, N., Blankertz, B. and Schlögl, A., 2009. Time domain parameters as a feature for EEG-based brain–computer interfaces. *Neural Networks*, *22*(9), pp.1313-1319.

Yan, R., Zhang, C., Spruyt, K., Wei, L., Wang, Z., Tian, L., Li, X., Ristaniemi, T., Zhang, J. and Cong, F., 2019. Multi-modality of polysomnography signals' fusion for automatic sleep scoring. *Biomedical Signal Processing and Control*, *49*, pp.14-23.

III

AUTOMATIC SLEEP SCORING TOOLBOX AND ITS APPLICATION IN SLEEP APNEA

by

Rui Yan, Fan Li, Xiaoyu Wang, Tapani Ristaniemi & Fengyu Cong 2020

Obaidat M. (eds) E-Business and Telecommunications. ICETE 2019.

Communications in Computer and Information Science, vol 1247, pp.256-275. Springer, Cham.

DOI: 10.1007/978-3-030-52686-3_11

Automatic Sleep Scoring Toolbox and Its Application in Sleep Apnea

Rui Yan[1, 2], Fan Li[1], Xiaoyu Wang[1], Tapani Ristaniemi[2], and Fengyu Cong[1, 2]

[1]School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China
[2]Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland
ruiyanmodel@foxmail.com

Abstract. Sleep scoring is a fundamental but time-consuming process in any sleep laboratory. Automatic sleep scoring is crucial and urgent to help address the increasing unmet needs for sleep research. Therefore, this paper aims to develop an automatic sleep scoring toolbox with the capability of multi-signal processing. The toolbox allows the user to choose signal types and the number of target classes. In addition, a user-friendly interface is provided to display sleep structures and related sleep parameters. The proposed approach employs several automatic processes including signal preprocessing, feature extraction and classification in order to save labor costs without compromising accuracy. For the phase of feature extraction, a huge number of features are considered including statistical characters, frequency characters, time-frequency characters, fractal characters, entropy characters and nonlinear characters. Their contribution to distinguishing between different sleep stages are compared in this article. The classifier we used for sleep stages discrimination is the random forest algorithm. The performance of the proposed approach is tested on the patients with sleep apnea by assessing accuracy, sensitivity and precision. The model achieves an accuracy of 82% to 86% for patients with varying degrees of sleep-disordered breathing, which indicates that sleep-disordered breathing does not significantly affect the performance of the proposed model. The proposed automatic scoring toolbox would alleviate the burden of the physicians, speed up sleep scoring, and expedite sleep research.

Keywords: Polysomnography, Multi-modality analysis, MATLAB toolbox, Automatic sleep scoring, Sleep-disordered breathing.

1 Introduction

Sleep covers almost one-third of human lifespan. Adequate and high-quality sleep is vital to our physical and mental well-being[1]. However, and likely because of our ephemeral lifestyle in modern society, sleep disorder complaints increase dramatically among people. One of the most common sleep disorders seen in sleep medicine centers is sleep apnea, which is characterized by repetitive cessations of the respiratory flow during sleep[2]. These nocturnal respiratory disturbances can induce sleep-EEG

changes, which promotes sleep fragmentation and an increase of arousals[3]. Studies have found that the distortion of sleep signals aggravates the burden of sleep scoring[4], especially reducing inter-scorer agreement. Clinically, the gold standard of diagnosis of sleep apnea is overnight polysomnography (PSG) (sleep study), which is carried out in a specialized hospital-based sleep laboratory. Polysomnography records simultaneously tens of sleep signals including electroencephalograms (EEG), electromyograms (EMG), electrooculogram (EOG), electrocardiogram (ECG), pulse oximetry, airflow, respiratory effort etc. Generally, according to the rules of Rechtschaffen & Kales (R&K)[5] and the most recently updated American Academy of Sleep Medicine rules (AASM)[6], these PSG recordings are scored manually by at least one registered sleep technologist (RST) to obtain the sequence of sleep stages and related sleep parameters.

The R&K rules divide sleep into five distinct stages: non-rapid eye movement (NREM) stages 1, 2, 3 and 4 and rapid eye movement stage (stage R), while the most recently developed standard AASM merges stages 3 and 4 into N3 due to their prevalent low-frequency oscillations. Generally, sleep experts divide sleep recordings into 30-second intervals called one epoch and classify each epoch based on its amplitude and frequency characteristics, as described in Table 1. The process of assigning a sleep stage to every epoch of polysomnographic recordings is called sleep scoring. Sleep scoring is a very important step in sleep research and clinical interpretation of polysomnography. However, the process of sleep scoring is labor-intensive, as studies have revealed that the annotation of 8-h recording requires approximately 2-4 hours[7]. Following the development of computerized methods, interests have been initiated to score automatically polysomnographic recordings, allowing the expert to avoid spending too much time on this time-consuming work. Nevertheless, it is a challenging problem because of the nonstationary nature of EEG signals and the complexity of sleep phenomena.

Numerous attempts have been made to automate sleep scoring[8, 9]. These methods were usually composed of two main components: feature generation or extraction and classification. For the phase of feature extraction, various signal processing techniques were explored such as wavelet transform[10, 11], empirical mode decomposition[12, 13], Hilbert-Huang transform[14, 15], Fourier transform[16] and short-time Fourier transform (STFT)[9, 17]. Then, diverse features, such as statistic features, frequency features and nonlinear features, were extracted from the transformed or decomposed signals of EEG, EOG and/or EMG [18, 19]. For classification, support vector machine (SVM)[20], random forest[19], K-nearest neighbor classifier[21], Naive Bayes[22], artificial neural network[23] etc. have been employed in literatures. In these studies, the agreement between automatic methods and human experts ranged from 80% to 90%[15, 23].

Besides traditional approaches, Liang and his colleagues[24] integrated multiscale entropy and autoregressive models for single-channel EEG achieving good performance. Another recent attempt proposed by Dimitriadis et.al[22] calculated the cross-frequency coupling of predefined frequency pairs, which demonstrated the effectiveness of phase-to-amplitude coupling and amplitude-to-amplitude coupling for the automatic classification of sleep stages. In addition, some studies based on waveform

detection have emerged, the most famous of which was deep learning methods[25]. CNNs were especially promising because they can learn complex patterns and 'look' at the data in a similar way as a 'real brain' although working with raw data required a huge amount of training data and computational resources[26].

In previous studies, ECG signals, as a substitute for standard techniques for determining sleep stages, were mainly used in portable sleep monitoring systems[27, 28]. Some studies proposed that sleep scoring with ECG is less complex but equally accurate when compared to PSG analysis[8]. Krakovská and Mezeiová [29] found the ECG feature named zero-crossing rate performed well in automatic sleep scoring, but was still inferior to those from the signals of EEG, EOG and EMG. Yan et.al[19] revealed that ECG features were useful in the recognition of stage R and W from other stages. Moreover, evidence from sleep physiology also suggested that the autonomous nervous system was regulated in totally different ways during wakefulness, slow-wave sleep and REM sleep[30]. Heart rate (HR) and arterial blood pressure decreased during non-REM sleep, while increased significantly during REM sleep, due to the modulations in sympathetic and parasympathetic activity [31]. Analysis of heart-rate variations was beneficial for us to track the transition from wakefulness to sleep. Therefore, the authors believed that ECG had promising prospects in automatic sleep scoring.

In our previous study[32], we proposed an automatic sleep scoring toolbox with the capability of multi-signal processing. Following automatic signal processing and analysis, the proposed toolbox achieved an average accuracy of 85.76% for 100 healthy subjects. Moreover, a user-friendly interface was provided to display sleep structures and related sleep parameters. Given the complexity of clinical conditions, the present article aims to extend the toolbox to patients with sleep apnea. More specifically, this work extends our preliminary work [32] in four aspects. First, the proposed toolbox covers multiple target domains so users can choose the number of target categories according to their needs. Second, more non-statistical features are considered to effectively track the variance of signals. Third, our previous study[32] performed feature evaluations based on a single experiment. The random-forest classifier accepts the number of features selected at random for each decision split, and therefore, it is necessary to perform 100 repeated experiments to obtain reliable results. Fourth, we investigate the influence of sleep-disordered breathing on classification. Healthy subjects and patients are diverging in their sleep structures, and therefore, it is important to examine whether these dissimilarities give rise to any difference in classification performance. To evaluate the performance of the proposed toolbox, various parameters such as accuracy, sensitivity and precision, will be considered in the following article.

The article is organized as follows: Section 2 gives an overview of the proposed toolbox. Experimental data and corresponding methodology are described in Section 3. Section 4 demonstrates the performance of the proposed toolbox and the model transferability between patients with different degrees of sleep apnea. Section 5 provides discussions of the results and limitations of this study. Finally, conclusions are given in Section 6.

Table 1. The characteristics of adult sleep records during each sleep stage[6]

Sleep stage	Description
Stage W	a) The EEG shows mixed beta and alpha activities as the eyes open and close, and predominantly alpha activity (8–13 Hz) when the eyes remain closed. b) The EOG channels show eye blinking, reading eye movement or rapid eye movement. c) Submental EMG is relatively high tone.
Stage N1	a) The EEG pattern is characterized by low amplitude, predominantly 4–7 Hz activity (theta wave). Vertex sharp waves with duration <0.5 seconds may occur in stage N1. b) The EOG often shows slow eye movements in stage N1. c) The EMG amplitude is variable but often lower than in stage W.
Stage N2	a) The EEG is characterized by predominant theta activity and occasional quick bursts of faster activity. Sleep spindles (commonly 12–14 Hz) and K complexes may appear here. b) The EOG usually shows no eye movement activity. c) The chin EMG shows variable amplitude, but it is usually lower than in stage W, and maybe as low as in stage R sleep.
Stage N3	a) The EEG activity is marked by high-amplitude slow waves. b) Eye movements are not typically seen during stage N3 sleep. c) The chin EMG is of variable amplitude, often lower than in stage N2 sleep and sometimes as low as in stage R sleep.
Stage R	a) It is characterized by relatively low-amplitude, mixed-frequency EEG activity without K complexes or sleep spindles. Sawtooth waves (2-6Hz, sharply contoured triangular) are strongly supportive of the presence of stage R sleep. b) Rapid eye movements are characteristic of stage R sleep. c) The chin EMG usually shows the lowest level of the entire recording.

2 System Overview

The toolbox proposed in this article aims to accelerate the process of sleep scoring, alleviating the burden of physicians. With this consideration in mind, we design a user-friendly operation interface, as shown in Fig.1, in order to facilitate operations and visualize analysis results. The interface consists of a training module, an offline prediction module, an online prediction module, several parameter panels and visualization panels. The following lines briefly describe the operation of the toolbox and the functions of the main modules. To run the proposed toolbox, a MATLAB environment was required which can be found at this link https://www.mathworks.com/.

If you have enough training data and want to train your model, the training module can help you. It should be mentioned that the user needs to select the signal type (multiple choice), the target number of stages (single selection) and the power frequency noise (a number). The software then automatically performs signal pre-processing, feature extraction and classifier training. The trained model would be stored in a folder named "Current Results" in the current working directory.

If you want to evaluate the sleep structure for new sleep recordings, the prediction module can help you whether or not you have a trained model. This module automatically checks if the user has a trained model (searches for a folder named “Current

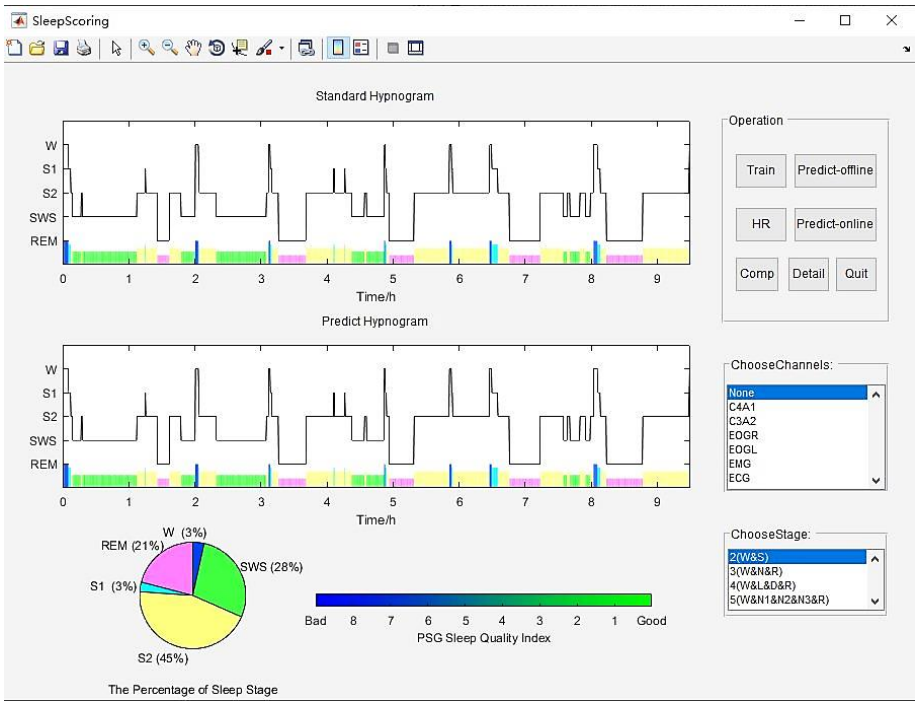


Fig. 1. The interface of sleep scoring toolbox

Results"). If no model is found in the folder, it would allow the user to specify a folder or load a predefined model. Once the model is determined, the module will automatically process the test data based on model parameters. Finally, the application interface displays the predicted sleep structure, related sleep parameters and the estimated sleep quality index. If a hypnogram (e.g., labels scored by RST) is available for the test data, the interface would display both hypnogram and predicted labels together. The disagreement between hypnogram and predicted labels would be highlighted by pressing a button named "Comp". Clicking the "Detail" button would pop up a new panel to display related sleep parameters. A button named "HR" is used to analyze the variance of heartbeat during sleep.

The online prediction module is similar to the offline prediction process except for the real-time updating results. This module can be connected to a sleep monitoring device to perform real-time analysis of sleep signals and real-time visualization sleep structures. The real-time update sleep signal will be saved in the memory as a TXT file.

3 Materials and Methods

3.1 Experiment Data

The all-night sleep recordings are provided by the Sleep Heart Health Study (SHHS) database, of which only the first round (SHHS-1) is selected in this study. The Sleep Heart Health Study (SHHS) is a multi-center cohort study held during 1995-1998 to investigate whether sleep-disordered breathing is associated with a higher risk of various cardiovascular diseases. In that study, 6441 individuals aged 40 years or older were recruited to undergo an overnight PSG and physical examination. Full details of SHHS study designs can be found in Quan et al.'s study[33].

Subjects employed in the present study did not use beta-blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease, or history of stroke. Patients with or without sleep-disordered breathing were randomly selected from the SHHS1 dataset. Table 2 summarized the characteristics of the employed subjects.

Each record was scored by an experienced research assistant or sleep technologist according to the R&K rules. Sleep recordings were segmented into 30-second per epoch and labelled as wakefulness (W), non-rapid eye movement stage (NREM, containing N1, N2, N3 and N4) and rapid eye movement stage (R). According to the recently updated AASM standards, NREM stages 3 and 4 were merged into N3 in the present article.

Table 2. Subject characteristics

Category	Number	Age	AHI range	AHI	Epoch
Normal breathing	100	46.86 (4.22)	AHI<5	2.11 (1.46)	97,514
Mild sleep apnea	30	48.50 (4.37)	5≤AHI<15	8.95 (2.62)	28,995
Moderate sleep apnea	30	48.53 (3.57)	15≤AHI<30	22.08 (4.83)	28,953
Severe sleep apnea	30	48.07 (4.86)	AHI≥30	41.38 (10.34)	28,713

3.2 Signal Pre-processing

For the predefined model and the following experiments, four modalities of polysomnography (PSG) signals were analyzed: two EEG channels (C4-A1 and C3-A2, following the 10-20 international electrode placement system), two electrooculography (EOG) channels (ROC and LOC), one submental electromyography (EMG) channel and one electrocardiography (ECG) channel. All of the aforementioned signals were fully included in the evaluation process without discarding any recorded segments, thereby to have a near-clinical situation.

In order to remove noise and artefacts, a notch filter, a high-pass filter with a cut-off frequency of 0.3Hz and a low-pass filter with a cut-off frequency of 30Hz were applied to the signals of EEG, EOG and ECG. In terms of EMG, a notch filter, a high-pass filter with a cut-off frequency of 10Hz and a low-pass filter with a cut-off frequency of 75Hz were performed. If the sampling frequency is less than twice the cut-

off frequency, the corresponding filtering process would be skipped. Overnight night recordings were smoothed by their mean value ±5×standard deviation to remove outliers. In order to eliminate individual differences, sleep signals were normalized to [-100, 100]. All signals were divided into 30-second epochs, each epoch corresponding to a single sleep stage. Signal interruption due to electrode fall-off or amplifier failure can be identified by judging the standard deviation of epochs. If $\mathrm{std}(x_{epoch}) < 0.01$, the epoch was judged as an interruption and was eliminated.

3.3 Feature Extraction

This study considers a variety of conventional and neoteric characteristics serving as distinctive markers for various psycho-physiological states. They are summarized in Table 3. Some of the parameters are introduced in the following, and the others can be found in our previous studies[19, 32].

In order to reduce the influence of extreme values, percentiles, instead of the maximum and minimum values, are used to measure the signals' amplitude. Besides, variance, skewness, kurtosis etc. are calculated which have proven to be good indicators of the amplitude distribution[11].

The calculation of all spectral measures is based on the short-time Fourier transform (STFT). The signal epoch is separated by a 3-second hamming moving window with 2-second overlap. Each partition is zero-padded to 5 times the length of sampling points, then a Fourier transform is calculated. The final spectral density is achieved by averaging spectral densities of the twenty-nine partitions.

Given the important role of rhythm waves and sleep events, the relative spectral power and spectral entropy are computed in the corresponding sub-frequency bands. Table 4 lists the EEG rhythm waves and sleep events together with their frequency ranges. Power ratios are computed based on the relative spectral power of rhythm waves and sleep events. The following power ratios are computed: delta/theta, delta/alpha, delta/beta, theta/alpha, theta/beta, alpha/beta, alpha/(theta + delta), delta/(theta + alpha) and theta/(beta + delta).

Nonlinear features are good methods to quantify the complexity of time series. Some nonlinear measures are described as follows.

The second difference of raw signal is defined as,

$$\mathrm{D} = 1/(N-2)\sum_{n=1}^{N-2}|x(n+2)-x(n)| \tag{1}$$

where N is the length of time series $x(n)$.

The fourth power of raw data,

$$\mathrm{P} = log_{10}\sum_{n=1}^{N} x(n)^4 \tag{2}$$

Entropy is a concept addressing irregularity or predictability. The greater entropy is often associated with more randomness and less system order. Recently, several different estimators of entropy have been introduced to quantify the oscillation of time series. In this paper, spectral entropy, Shannon entropy, Tsallis entropy ($\mathrm{q} = 2$) and

Renyi entropy ($\alpha = 2$) [34] are considered. Their corresponding definitions are listed below.

Spectral entropy,

$$SpE = \frac{1}{\ln(N)} \sum_{f=f_{min}}^{30Hz} P(f) * ln\big(P(f)\big) \quad (3)$$

where N is the length of the time series, $P(f)$ indicating the power spectral power, and $f_{\min}$ is set as 0.3Hz in terms of the signals of EEG, EOG and ECG, and 10Hz of EMG.

Table 3. Parameter list

Type	Feature Name
Statistical measures	The 5th, 25th, 75th, 95th percentile (Pre5, Pre25, Pre75, Pre95), Variance(V), Root mean square(Rms), Skewness(S), Kurtosis(K), Hjorth parameters(HA, HM, HC), Zero-Crossing(ZC)
Spectral measures	Power spectral density(PSD), Mean value of PSD (mPSD), Median value of PSD (mdPSD), Power ratio(PR), Relative spectral power(P), Mean of instantaneous frequency(Ifq), Mean frequency(mF), Median frequency(mdF), Brain rate (BR), Spectral centroid (Sc), Spectral width (Sw), Spectral asymmetry (Sa), Spectral flatness (Sk), Spectrum flatness (Sf), Spectral slope (Ss), Spectral decrease (Sd), Spectral edge frequency at 90% and 50%(edge90, edge50), Difference of spectral edges(Edge_D)
Nonlinear measures	Mean Teager energy (MTE), Mean energy (E), Mean curve length (CL), the second differences(SecD), The 4th Power(P4)
Fractal measures	Petrosian fractal dimension (PFD)
Entropy measures	Spectral entropy(spE), Shannon entropy(dnE), Renyi entropy (ryE), Tsallis entropy(tsE), Log energy entropy(lgE)
Mutual measures	Coherence(Ch), Coefficient (Cf)

Table 4. Rhythm waves and events of sleep EEGs

Name	Frequency band
Slow wave activity	0.5–2.0 Hz
Delta	0–3.99 Hz
Theta	4–7.99 Hz
Alpha	8–13 Hz
Beta	13–30 Hz
Sawtooth waves	2-6Hz
Sleep spindles	12–14 Hz
K complex	0.5–1.5 Hz

Shannon entropy,

$$\mathrm{ShnE} = -\sum_{n=1}^{N} p_i ln(p_i) \tag{4}$$

where $p_i = n_i/N$ is the histogram distribution of the time series x, N the length of the signal x, and n_i is the number of samples within the i^{th}bin.

Tsallis entropy,

$$\mathrm{TslE} = \frac{1}{1-q}\left(1 - \sum_{n=1}^{N} p_n^q\right) \tag{5}$$

where q is entropic index.

Renyi entropy[15],

$$\mathrm{RyiE} = \frac{1}{1-\alpha} log_2\left(\sum_{n=1}^{N} p_n^{\alpha}\right) \tag{6}$$

where α is the Renyi's entropy order.

3.4 Feature normalization

After all of the features extracted, post-processing is performed, which helps reduce the influence of signal noise. The outlier of each feature is defined as the element that is more than three standard deviations from the mean. All outliers would be replaced by their nearest non-outlier value.

In order to balance numerical ranges and reduce the impact of variability between-subjects or within-subjects, each feature vector of individuals is separately normalized to [0, 1] by the following formula,

$$\bar{p}_{i,j} = \left[p_{i,j} - min(p_j)\right] / \left[max(p_j) - min(p_j)\right] \tag{7}$$

where p_j denotes an independent feature vector from an individual, and $p_{i,j}$ is an element in the j^{th} feature vector.

3.5 Classification

In order to capture the characters of sleep stages and to predict new instances, the present article employs a random forest (RF) classifier because it is relatively parameter-free, robust to outliers, fast to train and resistant to overfitting[35]. This study designs a classifier with 200 trees, where each tree is created using a randomly selected set of m features ($m = \sqrt{M}$, where M is the total number of features.). These features will be used to design decision nodes of each tree. Those decision nodes divide test samples into specific categories, and the final classification results are achieved by the most votes in the forest. Moreover, the RF classifier is able to store the out-of-bag information for each node, which provides an estimation of the importance of each feature. The classifier is trained using 10-fold cross-validation. The test set is completely independent, so the test set consists of data that the model has never seen before.

3.6 Result Correction

Studies have found that sleep transitions are not a random process. However, traditional classifiers can only give their decisions based on information of the current stage, but can't remember the context. The objective of this phase is to utilize the strengths of the hidden Markov model to complement the weaknesses of conventional classifiers. Firstly, hidden Markov model (HMM) is trained using the classification results of validation dataset. The HMM parameters, namely transition matrix, emission matrix and initial matrix, are described as follows. Concerning the initial matrix, all states have equal probabilities to be the initial state.

The transition matrix $\mathrm{A} = \{a_{ij}\}$ stores the probability of stage s_j following stage s_i. a_{ij} can be calculated by the following formula,

$$a_{ij} = \frac{\sum s_i \to s_j}{\sum_{t=1}^{N} s_t = S_i} \tag{8}$$

where s_t denotes sleep stage of each epoch, N the number of epochs, and $S_i, S_j \in \{W, R, N1, N2, N3\}$.

The emission matrix[36] is calculated from the confusion matrix $\mathrm{C} = \{c_{ij}\}$ where c_{ij} represents the probability of classifying stage S_i as stage S_j.

$$a_{ij} = \frac{N_{s_i \to s_j}}{N_{s_i}} \tag{9}$$

where N_{s_i} denotes the total number of stage S_i, $N_{s_i \to s_j}$the number of stage S_i classified to stage S_j, and $S_i, S_j \in \{W, R, N1, N2, N3\}$.

After determining the HMM model, the Viterbi algorithm[37] is used to find the most likely sequence of states through the trellis. The correction algorithm takes a sequence of observations (classifier output) as input and returns a sequence of states as output (correct stages). The correction process refer to the observations prior to and posterior to the current epoch, which makes the corrected stages avoid unreasonable stages transitions. More specifically, the stage transition $[W, R, N2]$ would be corrected to $[W, N1, N2]$. The study by Liang et al.[38] illustrated some unreasonable transitions. Experiments have proved that these unreasonable transitions can be corrected by the proposed method.

3.7 PSG Sleep Quality Index

In order to provide an objective estimation of sleep quality, the toolbox proposes a sleep quality index which is built according to items of the Pittsburgh Sleep Quality Index (PSQI). PSQI is a standard self-report questionnaire for assessing sleep quality during the previous month[39]. The self-report questionnaire is relatively subjective and can be easily exaggerated or minimized by the person completing it. Therefore, the toolbox calculates various sleep parameters (Table 2 in reference [32]) from the predicted sleep structure. Afterwards, statistical scoring is performed based on these sleep parameters. The score is then normalized into a range of [0, 9] that is PSG sleep

quality index. The PSG index is displayed in the bar in the lower-right corner of the interface. Detailed sleep parameters can be obtained by pressing the button “Detail”.

4 Performance Assessment

To demonstrate the generalizability of the proposed toolbox, the proposed model was tested on subjects with normal, mild, moderate and severe sleep-disordered breathing. Model performance was evaluated by accuracy, sensitivity and precision. Accuracy indicated the fraction of correct detections. Sensitivity represented the proportion of positive epochs that are correctly identified. Precision is the proportion of epochs predicted as belonging to the positive class to epochs that actually belong to the positive class. Their definitions[11] were listed below.

$$Acc = \frac{TP+TN}{TP+FN+FP+TN}(\%) \tag{10}$$

$$Sen = \frac{TP}{TP+FN}(\%) \tag{11}$$

$$Pre = \frac{TP}{TP+FP}(\%) \tag{12}$$

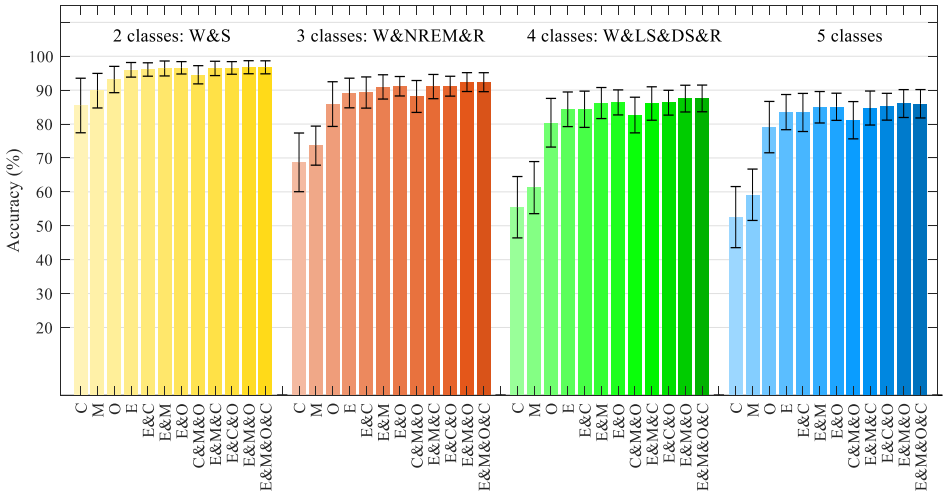
where TP, TN, FP and FN respectively denote true positives, true negatives, false positives and false negatives.

4.1 Influence of Signal Types

In order to explore the relationship between signal types and classification accuracy, we performed a greedy search on signal fusions for two to six sleep stages. The results were shown in Fig.2 where the column denoted the mean accuracy of 100 normal sleep-breathing subjects and the bars represented the standard deviation. Four target domains were highlighted in different colors in Fig.2. For each target domain, twelve signal fusions were considered and displayed along X-axis where signals’ name was abbreviated with its middle letter.

As can be seen from Fig.2, with the enrichment of signal types, the mean value of accuracy increased, and the uncertainty decreased to some extent. More specifically, Fig.2 indicated that the required signal types increased with the fineness of sleep scoring. If sleep recordings were classified into two stages, namely wakefulness (W) and sleep (S), all considered signal fusions gave satisfactory results. As the number of sleep stages increased, the number of required signals increased accordingly if a satisfactory accuracy was desired.

From the perspective of signal types, the signal fusions containing EEG signals showed better classification accuracy, indicating that EEG signals played a vital role in sleep scoring. The discriminative information provided by ECG and EMG channels was inferior to that from EEG and EOG signals. Nevertheless, the richness of signal modalities contributed to the increasing accuracy of sleep stages classification, but up to a certain point.



*C: single-modality ECG; M: single-modality EMG; O: single-modality EOG; E: single-modality EEG; &: combination signals; 2 classes: wakefulness and sleep (W&S); 3 classes: wakefulness, non-rapid eye movement sleep and rapid eye movement sleep (W&NREM&REM); 4 classes: wakefulness, light sleep (containing N1 and N2), deep sleep (N3) and rapid eye movement sleep (W&LS&DS &REM); 5 classes: W, N1, N2, N3 and R

Fig. 2. The classification accuracy for different signal fusions and target class[32].

4.2 Feature Evaluation

To further elucidate the features' contribution to automatic sleep scoring, the importance of features, measured by their contribution to distinguishing each pair of sleep stages, were derived from the random forest classifier. Given random-forest classifier randomly selected a subset of features for each decision split, and therefore, we performed 100 repeated tests for each experiment. The final importance of each feature was the average of 100 tests. The top 15features was highlighted in Fig.3, where features were sorted in descending order of discriminative ability.

Fig.3 indicated that the optimal feature subset was a fusion of statistical measures (e.g. Percentiles, Hjorth parameters, Zero-Crossing), spectral measures (e.g. spectral edge, power spectral density), entropy measures (e.g. spectral entropy), fractal measures (e.g. Petrosian fractal dimension) and nonlinear measures (e.g. mean curve length, The 4th Power). Varied features captured signals from multiple aspects, thereby improving accuracy.

Fig.4 listed the distribution of feature importance in distinguishing each pair of sleep stages. As can be seen from the Fig.4, the features from EEG contributed to the recognition of most stages. EMG features were good at distinguishing stage R and wakefulness, partly because of the apparent change of EMG amplitude in stage R and stage W. ECG features were inferior to those from the signals of EEG, EOG and EMG, which is consistent with the previous studies[29]. Nevertheless, the ECG features showed their contribution in distinguishing stage N3, R and W.

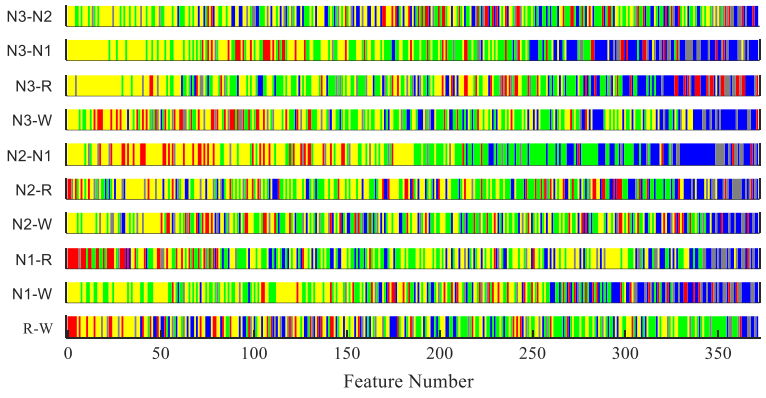
Sleep stages	N3-N2	N3-N1	N3-R	N3-W	N2-N1	N2-R	N2-W	N1-R	N1-W	R-W
Top1	E.K	E.lgE	E.lgE	E.ZC	E.Sf	M.lgE	E.P-St	M.lgE	E.P-St	M.lgE
Top2	E.lgE	E.Sf	E.per75	E.PFD	E.PFD	M.per25	E.spE-St	M.per25	E.spE-St	M.K
Top3	E.per75	E.per75	E.K	E.ifq	E.P-K	Cf	E.PFD	M.per75	E. θ/β	M.per25
Top4	O.ifq	E.K	Cf	O.P4	E.ryE	E.spE-sp	E.ZC	M.per5	E. θ (α+β)	M.per75
Top5	O.mdF	E.ZC	E.PFD	E.Sf	E.spE-K	M.per5	E.ifq	M.SecD	E.ifq	M.per5
Top6	E.Sf	E.spE-K	E.Sf	O.ryE	E.P4	O.spE-α	O.P4	M.K	O.P4	E.P-St
Top7	O.K	E.ifq	E.per25	E.θ/β	O.PFD	E.P-sp	E.θ/(α+β)	Ch	E. θ/α	M.CL
Top8	E.PFD	E.P-K	E.spE-K	E.spE-β	Ch	M.per75	O.ryE	M.CL	O.spE-δ	E.spE-St
Top9	E.Sw	E.spE-β	E.Rms	M.per25	E.δ/θ	O.P-α	E.P-δ	M.mdPSD	E.P-δ	E. θ/(α+β)
Top10	Cf	E.PFD	E.per95	E.Sw	E.Rms	M.K	E. δ/α	M.per95	O.SecD	M.SecD
Top11	O.ZC	E.per25	E.P-β	M.P4	Cf	E.spE-K	E. δ/(α+β)	O.spE-α	O.ryE	E. θ/α
Top12	E.per25	E.HM	E.spE-β	M.per5	M.P4	E.ryE	Ch	M.ryE	E.ZC	M.per95
Top13	O.edge50	E.per95	E.spE-St	M.per95	E.per95	Cf	E.spE-δ	M.Sw	E. δ/(α+β)	E.PFD
Top14	E.SecD	E.per5	E.ryE	E.K	E.lgE	Ch	O.S	O.P-α	E.P-K	E. θ/β
Top15	E.spE-θ	E.Sw	E.P-δ	E.HM	E.spE-θ	E.PFD	Ch	M.Sk	O.S	Ch

Top 15 features

*EEG features (colour: yellow); EOG features (colour: green); EMG features ((colour: red)); ECG features (colour: blue); Mutual information (color: gray).

*-K: K-complex; -sp: sleep spindle; -St: Sawthooth;

Fig. 3. Selected features for distinguishing specific pair of sleep stages



*EEG features (colour: yellow); EOG features (colour: green); EMG features ((colour: red)); ECG features (colour: blue); Mutual information (color: gray). Decreasing importance from left to right.

Fig. 4. Feature distributions for distinguishing each pair of sleep stages

4.3 Performance on transfer learning

Table 5 presented the confusion matrix of classification results. Except for stage N1, all other stages were correctly classified with a precision of 80% or more. Stage N1 was considered as a transient state between wakefulness and “real” sleep, so it usually

contained information of two or three sleep stages. As a result, the scoring of N was quite obscure, even for sleep scoring experts[29].

The model's generalizability was assessed by patients with mild, moderate, and severe sleep-disordered breathing. It was worth noting that the model was trained on 100 subjects with normal sleep-breathing. Therefore, all patients with sleep-disordered breathing consisted of information that the model had never seen before. Table 6-8 displayed the confusion matrix of test subjects. As can be seen from tables, the model performed slightly better in normal and mild apnea cases than in moderate and severe cases. Nevertheless, even for patients with severe sleep-disordered breathing, the proposed model could provide per-epoch sensitivity and precision comparable to state of the art, indicating that the proposed model had good transferability between healthy populations and patients.

Table 5. Confusion matrix for AHI<5 subjects

		Technologists' score stage					**Pre.**
	Stage	**W**	**R**	**N1**	**N2**	**N3**	
Proposed	**W**	18370	134	716	504	23	93.0%
	R	589	15817	504	1727	0	84.9%
	N1	274	98	750	382	0	49.9%
	N2	1177	906	1044	38035	2726	86.7%
	N3	11	2	0	2202	11523	83.9%
Sen.		90.0%	93.3%	24.9%	88.8%	80.7%	
Acc.							86.65%

Table 6. Confusion matrix for 5<AHI<15 subjects

		Technologists' score stage					**Pre.**
	Stage	**W**	**R**	**N1**	**N2**	**N3**	
Proposed	**W**	5849	70	328	155	13	91.2%
	R	119	4261	163	315	0	87.7%
	N1	122	32	303	161	0	49.0%
	N2	325	459	333	11110	930	84.4%
	N3	3	0	0	501	3443	87.2%
Sen.		91.1%	88.4%	26.9%	90.8%	78.5%	
Acc.							86.1%

Table 7. Confusion matrix for 15<AHI<30 subjects

		Technologists' score stage					**Pre.**
	Stage	**W**	**R**	**N1**	**N2**	**N3**	
Proposed	**W**	5290	169	411	469	13	83.3%
	R	133	4089	135	286	0	88.1%
	N1	87	175	273	295	0	32.9%
	N2	498	460	367	11268	1000	82.9%
	N3	23	2	0	654	2856	80.8%
Sen.		87.7%	83.5%	23.0%	86.9%	73.8%	
Acc.							82.1%

Table 8. Confusion matrix for AHI>30 subjects

		Technologists' score stage					Pre.
	Stage	**W**	**R**	**N1**	**N2**	**N3**	
Proposed	**W**	5735	177	352	440	3	85.5%
	R	266	3490	129	226	0	84.9%
	N1	122	172	247	165	0	35.0%
	N2	735	582	406	11439	678	82.7%
	N3	30	0	1	727	2591	77.4%
Sen.		83.2%	79.0%	21.8%	88.0%	79.2%	
Acc.							81.9%

5 Discussion

Burgeoning public interest in sleep health had provided a strong impetus for the study of automatic sleep scoring. To the best of our knowledge, limited toolbox or software was available for automatic sleep scoring, although there were many theoretical studies in this field. In previous studies, some portable devices were developed based on ECG and respiration. For example, Hermawan et al. [28] developed a real-time monitoring device that classified sleep into 2 stages (wakefulness and sleep) with an average precision of 0.941. Recently, some deep-learning-based scoring tools sprouted out, such as SLEEPNET [40] and SeqSleepNet [41]. The classification accuracy of these deep-learning-based tools was not less than 0.85 with the support of tremendous training data and highly configured computer (like GPU or server). Compared with previous studies, the proposed toolbox provided comparable or better precision and greater freedom. Even for the subjects with severe sleep-disordered breathing, the model performance met or exceeded the accepted benchmark of $Acc = 80\%$ between trained human scorers[42]. In addition, the proposed toolbox evaluated all epochs, including severely contaminated or distorted epochs, thereby to have a near-clinical condition.

The MATLAB-based toolbox allowed users to select the available signal types and the number of target classes. Based on our aforementioned analysis, it was worth adding multi-modality sensors, if a satisfactory accuracy was desired. This conclusion was consistent with previous research that claimed features from multi-modality signals were beneficial to the improvement of scoring accuracy[9], but up to a certain point. Compared to our previous results[32], the importance of features from 100 repeated tests was slightly different from the results from a single experiment. One possible explanation was that the random forest randomly selected features for decision splits. Besides that, it can also be attributed to changes of features. In the present article, we refined the frequency bands of EEG signals and calculated more non-statistical features for EEG rhythm waves and sleep events in the present paper, thereby making the contribution of EEG signals even more prominent. Unfortunately, after 100 repeated tests, we did not find an ECG feature in the first 15 features, although some ECG features performed well in several tests. That may be attributed to the characteristics of the classifier. The random forest classifier tended to select fea-

tures containing many distinct values over those containing few distinct values[43]. Compared to changes of EEG and EMG signals during sleep, changes of heart rate seems to be slow and gradual. Nevertheless, we found that the ECG features performed relatively well in distinguishing stage N3, R and W, as shown in Fig.4.

From the perspective of applications, the proposed toolbox would be helpful for researchers, especially the newcomers in this field, to accelerate their understanding of sleep structures. It can also be used in clinic to expedite the annotation of PSG records, thereby alleviating the burden of the physicians. The online prediction module provides the potential to automatically control sleep tasks by combining the toolbox with sleep experiments.

Even though our results are encouraging, our model still has several limitations. One of them is that the performance of proposed toolbox is affected by the data property. Since our model learns from the training data, it might not perform well if the test data and training data have different properties. For example, a scoring model trained by healthy subjects may not perform well on patients. In that case, the model might need to be re-trained or fine-tuned if a good precision is desired.

Another complaint about the proposed algorithm might be worse N1 precision. A possible cause is the obscure character of stage N1. From the neurophysiological standpoint, N1 is a transition phase between wakefulness and sleep, which contains information of two or three sleep stages. The recognition of stage N1 is also an enormous challenge for other published models, even for sleep scoring experts[7]. Another reason affecting the precision of stage N1 might be unbalanced instances. It is caused by the natural asymmetric distribution of sleep stages. In this article, we did not consider any sample-balance strategy, such as resampling, since according to our study, the improved precision of stage N1 comes at the expense of other stages. We believe that preserving the natural sleep structure best represents how the model would perform in a production setting.

The proposed model performs slightly better in normal and mild apnea cases than moderate and severe cases. This can trace back to different pathological manifestations of patients with normal breathing and patients with severe apnea. Studies have shown that patients with severe sleep apnea suffer from sleep fragmentation and an increase of arousals, which pose a challenge to sleep scoring. Other physiological changes in sleep apnea can’t be excluded. Some studies found significant changes in EEG spectral power as disease progression [44]. Other studies have shown that patients with severe obstructive apnea had higher activity level in the sympathetic nervous system[45], which modulate the activity of brain and cardiopulmonary system. In addition to pathological factors, another explanation for the model's poor performance in severe apnea cases is model-mismatch. The author believes that the model accuracy can be further improved by training with sufficient data on severe apnea cases.

6 Conclusions

This study proposed an automatic sleep scoring toolbox with the capability of multi-signal processing. It allowed users to choose the signal type and target domain ac-

cording to their needs. Sleep recordings can be automatically analyzed in batches or individually. In addition, different signals' fusions and a huge number of characteristics were investigated in present studies to highlight their contribution to automatic sleep scoring. The performance of the proposed model has been tested on patients with varying degrees of sleep apnea. The patient-specific accuracy ranged from 82% to 86%, indicating good generalizability of the proposed model. Compared with manual scoring, the proposed automatic scoring toolbox was cost-effective, which would alleviate the burden of the physicians, speed up sleep scoring and expedite sleep research.

Acknowledgements

The authors would like to thank the SHHS for providing the polysomnographic data. This work was supported by the scholarships from China Scholarship Council (Nos. 201606060227).

References

1. Strine, T.W., Chapman, D.P.: Associations of frequent sleep insufficiency with health-related quality of life and health behaviors. Sleep Med. 6, 23–27 (2005). https://doi.org/10.1016/j.sleep.2004.06.003
2. Pagel, J.F., Pandi-Perumal, S.R.: Primary care sleep medicine: A practical guide. Springer (2014)
3. Nano, M.-M., Long, X., Werth, J., Aarts, R.M., Heusdens, R.: Sleep apnea detection using time-delayed heart rate variability. 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC). IEEE. 7679–7682 (2015). https://doi.org/10.1109/EMBC.2015.7320171
4. Schluter, T., Conrad, S.: An approach for automatic sleep stage scoring and apnea-hypopnea detection. In: 2010 IEEE International Conference on Data Mining. pp. 230–241 (2010)
5. Rechtschaffen, A., Kales, A.: A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Washingt. DC US Natl. Inst. Heal. Publ. (1968)
6. Berry, R.B., Budhiraja, R., Gottlieb, D.J., Gozal, D., Iber, C., Kapur, V.K., Marcus, C.L., Mehra, R., Parthasarathy, S., Quan, S.F., Redline, S., Strohl, K.P., Ward, S.L.D., Tangredi, M.M.: Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events. J. Clin. Sleep Med. 8, 597–619 (2012). https://doi.org/10.5664/jcsm.2172
7. Hassan, A.R., Bhuiyan, M.I.H.: A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. J. Neurosci. Methods. 271, 107–118 (2016). https://doi.org/10.1016/j.jneumeth.2016.07.012
8. Faust, O., Razaghi, H., Barika, R., Ciaccio, E.J., Acharya, U.R.: A review of automated sleep stage scoring based on physiological signals for the new millennia. Comput. Methods Programs Biomed. 176, 81–91 (2019). https://doi.org/10.1016/j.cmpb.2019.04.032

9. Boostani, R., Karimzadeh, F., Nami, M.: A comparative review on sleep stage classification methods in patients and healthy individuals. Comput. Methods Programs Biomed. 140, 77–91 (2017). https://doi.org/10.1016/j.cmpb.2016.12.004
10. Khalighi, S., Sousa, T., Pires, G., Nunes, U.: Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. Expert Syst. Appl. 40, 7046–7059 (2013). https://doi.org/10.1016/j.eswa.2013.06.023
11. Şen, B., Peker, M., Çavuşoğlu, A., Çelebi, F. V.: A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J. Med. Syst. 38, 18 (2014). https://doi.org/10.1007/s10916-014-0018-0
12. Hassan, A.R., Bhuiyan, M.I.H.: Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. Biomed. Signal Process. Control. 24, 1–10 (2016). https://doi.org/10.1016/j.bspc.2015.09.002
13. Hassan, A.R., Hassan Bhuiyan, M.I.: Automatic sleep scoring using statistical features in the EMD domain and ensemble methods. Biocybern. Biomed. Eng. 36, 248–255 (2016). https://doi.org/10.1016/j.bbe.2015.11.001
14. Li, Y., Yingle, F., Gu, L., Qinye, T.: Sleep stage classification based on EEG hilbert-huang transform. 2009 4th IEEE Conf. Ind. Electron. Appl. ICIEA 2009. 3676–3681 (2009). https://doi.org/10.1109/ICIEA.2009.5138842
15. Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., Dickhaus, H.: Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. Comput. Methods Programs Biomed. 108, 10–19 (2012). https://doi.org/10.1016/j.cmpb.2011.11.005
16. Otzenberger, H., Simon, C., Gronfier, C., Brandenberger, G.: Temporal relationship between dynamic heart rate variability and electroencephalographic activity during sleep in man. Neurosci. Lett. 229, 173–176 (1997). https://doi.org/10.1016/S0304-3940(97)00448-5
17. Álvarez-Estévez, D., Fernández-Pastoriza, J.M., Hernández-Pereira, E., Moret-Bonillo, V.: A method for the automatic analysis of the sleep macrostructure in continuum. Expert Syst. Appl. 40, 1796–1803 (2013). https://doi.org/10.1016/j.eswa.2012.09.022
18. Šušmáková, K., Krakovská, A.: Discrimination ability of individual measures used in sleep stages classification. Artif. Intell. Med. 44, 261–277 (2008). https://doi.org/10.1016/j.artmed.2008.07.005
19. Yan, R., Wei, L., Zhang, C., Li, X., Ristaniemi, T., Spruyt, K., Wang, Z., Tian, L., Cong, F., Zhang, J.: Multi-modality of polysomnography signals' fusion for automatic sleep scoring. Biomed. Signal Process. Control. 49, 14–23 (2019). https://doi.org/10.1016/j.bspc.2018.10.001
20. Seifpour, S., Niknazar, H., Mikaeili, M., Nasrabadi, A.M.: A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal. Expert Syst. Appl. 104, 277–293 (2018). https://doi.org/10.1016/j.eswa.2018.03.020
21. Gharbali, A.A., Najdi, S., Fonseca, J.M.: Investigating the contribution of distance-based features to automatic sleep stage classification. Comput. Biol. Med. 96, 8–23 (2018). https://doi.org/10.1016/j.compbiomed.2018.03.001
22. Dimitriadis, S.I., Salis, C., Linden, D.: A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates. Clin. Neurophysiol. 129, 815–828 (2018). https://doi.org/10.1101/160655

23. Özşen, S.: Classification of sleep stages using class-dependent sequential feature selection and artificial neural network. Neural Comput. Appl. 23, 1239–1250 (2013). https://doi.org/10.1007/s00521-012-1065-4
24. Liang, S.F., Kuo, C.E., Hu, Y.H., Pan, Y.H., Wang, Y.H.: Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. IEEE Trans. Instrum. Meas. 61, 1649–1657 (2012). https://doi.org/10.1109/TIM.2012.2187242
25. Chambon, S., Thorey, V., Arnal, P.J., Mignot, E., Gramfort, A.: DOSED: a deep learning approach to detect multiple sleep micro-events in EEG signal. J. Neurosci. Methods. 321, 64–78 (2019)
26. Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., Achermann, P.: Automatic Human Sleep Stage Scoring Using Deep Neural Networks. Front. Neurosci. 12, 781 (2018). https://doi.org/10.3389/fnins.2018.00781
27. Ebrahimi, F., Setarehdan, S.K., Nazeran, H.: Automatic sleep staging by simultaneous analysis of ECG and respiratory signals in long epochs. Biomed. Signal Process. Control. 18, 69–79 (2015). https://doi.org/10.1016/j.bspc.2014.12.003
28. Hermawan, I., Alvissalim, M.S., Tawakal, M.I., Jatmiko, W.: An integrated sleep stage classification device based on electrocardiograph signal. 2012 Int. Conf. Adv. Comput. Sci. Inf. Syst. 37–41 (2012)
29. Krakovská, A., Mezeiová, K.: Automatic sleep scoring: A search for an optimal combination of measures. Artif. Intell. Med. 53, 25–33 (2011). https://doi.org/10.1016/j.artmed.2011.06.004
30. Roebuck, A., Monasterio, V., Gederi, E., Osipov, M., Behar, J., Malhotra, A., Penzel, T., Clifford, G.D.: A review of signals used in sleep analysis. Physiol. Meas. 35, R1 (2013). https://doi.org/10.1088/0967-3334/35/1/R1
31. Trinder, J., Kleiman, J., Carrington, M., Smith, S., Breen, S., Tan, N., Kim, Y.: Autonomic activity during human sleep as a function of time and sleep stage. J. Sleep Res. 10, 253–264 (2001). https://doi.org/10.1046/j.1365-2869.2001.00263.x
32. Yan, R., Li, F., Wang, X., Ristaniemi, T., Cong, F.: An automatic sleep scoring toolbox: Multi-modality of Polysomnography Signals' Processing. ICETE 2019 - Proc. 16th Int. Jt. Conf. E-bus. Telecommun. 1, 307–315 (2019). https://doi.org/10.5220/0007925503010309
33. Quan, S.F., Howard, V., Iber, C., Kiley, J.P., Nieto, J., Connor, G.T.O., Rapoport, D.M., Redline, S., Samet, I.M., Wahl, P.W.: The sleep heart health study: design, rationale, and methods. Sleep. 20, 1077–1085 (1997). https://doi.org/10.1093/sleep/20.12.1077
34. Khalighi, S., Sousa, T., Santos, J.M., Nunes, U.: ISRUC-Sleep: A comprehensive public dataset for sleep researchers. Comput. Methods Programs Biomed. 124, 180–192 (2016). https://doi.org/10.1016/j.cmpb.2015.10.013
35. Cooray, N., Andreotti, F., Lo, C., Symmonds, M., Hu, M.T.M., De Vos, M.: Automating the detection of REM sleep behaviour disorder. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS. 2018–July, 1460–1463 (2018). https://doi.org/10.1109/EMBC.2018.8512539
36. Esmael, B., Arnaout, A., Fruhwirth, R.K., Thonhauser, G.: Improving time series classification using Hidden Markov Models. Proc. 2012 12th Int. Conf. Hybrid Intell. Syst. HIS 2012. 502–507 (2012). https://doi.org/10.1109/HIS.2012.6421385

37. Hagenauer, J., Germany, W.: A viterbi algorithm with soft-decision outputs and its applications. 1989 IEEE Glob. Telecommun. Conf. Exhib. Technol. 1990s Beyond. IEEE. 1680–1686 (1989)
38. Liang, S.F., Kuo, C.E., Hu, Y.H., Cheng, Y.S.: A rule-based automatic sleep staging method. J. Neurosci. Methods. 205, 169–176 (2012). https://doi.org/10.1016/j.jneumeth.2011.12.022
39. Sohn, S. Il, Kim, D.H., Lee, M.Y., Cho, Y.W.: The reliability and validity of the Korean version of the Pittsburgh Sleep Quality Index. Sleep Breath. 16, 803–812 (2012). https://doi.org/10.1007/s11325-011-0579-9
40. Biswal, S., Kulas, J., Sun, H., Goparaju, B., Westover, M.B., Bianchi, M.T., Sun, J.: SLEEPNET: Automated sleep staging system via deep learning. arXiv Prepr. arXiv1803.01710. (2017)
41. Phan, H., Andreotti, F., Cooray, N., Chén, O.Y., De Vos, M.: SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans. Neural Syst. Rehabil. Eng. 27, 400–410 (2019). https://doi.org/10.1109/TNSRE.2019.2896659
42. Zhang, L., Fabbri, D., Upender, R., Kent, D.: Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. Sleep. 42, zsz159 (2019). https://doi.org/10.1093/sleep/zsz159
43. Loh, W.Y., Shih, Y.S.: Split selection methods for classification trees. Stat. Sin. 7, 815–840 (1997)
44. Rahman, M.J., Mahajan, R., Morshed, B.I.: Exacerbation in obstructive sleep apnea: Early detection and monitoring using a single channel EEG with quadratic discriminant analysis. 2019 9th Int. IEEE/EMBS Conf. Neural Eng. (NER). IEEE. 85–88 (2019). https://doi.org/10.1109/NER.2019.8717054
45. Van Steenkiste, T., Groenendaal, W., Deschrijver, D., Dhaene, T.: Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. IEEE J. Biomed. Heal. Informatics. 23, 2354–2364 (2018). https://doi.org/10.1109/JBHI.2018.2886064

IV

A DEEP LEARNING MODEL FOR AUTOMATIC SLEEP SCORING USING MULTIMODALITY TIME SERIES

by

Rui Yan, Fan Li, Dongdong Zhou, Tapani Ristaniemi & Fengyu Cong 2020

28th European Signal Processing Conference (EUSIPCO 2020), 5 pages, Amsterdam, Netherlands.

A Deep Learning Model for Automatic Sleep Scoring using Multimodality Time Series

Rui Yan
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
ruiyanmodel@foxmail.com

Fan Li
School of Biomedical Engineering
Dalian University of Technology
Dalian, China
lifandlpu@foxmail.com

DongDong Zhou
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
dongdongzhou1017@foxmail.com

Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
tapani.e.ristaniemi@jyu.fi

Fengyu Cong
School of Biomedical Engineering
Dalian University of Technology
Dalian, China
fengyu.cong@foxmail.com

***Abstract*—Sleep scoring is a fundamental but time-consuming process in any sleep laboratory. Automatic sleep scoring is crucial and urgent to help address the increasing unmet need for sleep research. Therefore, this paper aims to develop an end-to-end deep learning architecture using raw polysomnographic recordings to automate sleep scoring. The proposed model adopts two-dimensional convolutional neural networks (2D-CNN) to automatically learn features from multi-modality signals, together with a "squeeze and excitation" block for recalibrating channel-wise feature responses. The learnt representations are finally fed to a softmax classifier to generate predictions for each sleep stage. The model performance is evaluated on two public sleep datasets (SHHS and Sleep-EDF) with different available channels. The results have shown that our model achieves an overall accuracy of 85.2% on the SHHS dataset and an accuracy of 85% on the Sleep-EDF dataset. We have also demonstrated that the proposed architecture not only is able to handle various numbers of input channels and several signal modalities from different datasets but also exhibits short runtimes and low computational cost.**

Keywords—polysomnography, automatic sleep scoring, multi-modality analysis, deep learning, transfer learning

I. INTRODUCTION

Adequate and high-quality sleep is vital to our physical and mental well-being. Nowadays, and likely because of our ephemeral lifestyle in modern society, complaints about sleep disorders increase dramatically among people. An effective way to diagnose sleep disorders and monitor sleep quality is overnight polysomnography (PSG), which is carried out in a specialized hospital-based sleep laboratory. A PSG test simultaneously records dozens of sleep signals including electroencephalograms (EEG), electrooculogram (EOG), electromyograms (EMG), electrocardiogram (ECG), pulse oximetry, airflow, respiratory effort etc. The standard method of analyzing PSG recordings is based on the criteria proposed by Rechtschaffen and Kales (R&K)[1] and the recently updated American Academy of Sleep Medicine (AASM) standards[2].

Based on the amplitude and frequency characteristics of sleep signals, the R&K rules divide sleep into five distinct stages: non-rapid eye movement (NREM) stages 1, 2, 3 and 4 and rapid eye movement stage (stage R), but the recently updated AASM standard merges stages 3 and 4 into N3 due to their prevalent low-frequency oscillations. The process of assigning a sleep stage to each sleep segment is called sleep scoring, which is a fundamental step in sleep research. However, the process of sleep scoring is labor-intensive, as studies have revealed that the annotation of an 8-h recording requires approximately 2-4 hours[3]. With the development of computerized methods, there is a growing interest in automatic scoring of PSG recordings.

Numerous attempts so far have been made to automate sleep scoring[4]. Conventional machine-learning methods mainly consist of two main components: feature extraction and classification. For the step of feature extraction, diverse features, such as statistic features, frequency features and nonlinear features, are extracted from the transformed or decomposed signals of EEG, EOG and/or EMG[5]. For classification, support vector machine, random forest, K-nearest neighbor classifier, Naive Bayes, artificial neural network etc. have been employed in the existing literature[6], [7]. In these studies, the agreement between automatic methods and human experts ranged from 0.8 to 0.9 and that value highly relied on the validity of employed features.

Most recently, in the field of automatic sleep scoring, there have sprung up many algorithms that adopted deep learning networks since it did not require explicit feature extraction and was especially suitable for big data approach[8]. Convolutional neural network (CNN) had been used on raw EEG signals to extract features automatically[9], which offered competitive scoring performance on a large multi-center sleep dataset. PSG signals were also transformed into time-frequency images using short-time Fourier transform[10] or wavelet transformation[11], given the superiority of CNNs in image processing. Moreover, deep learning algorithms had introduced some novel classification schemes to mimic the way sleep experts performed manual sleep scoring, such as one-input-multi-output scheme[12] and sequence-to-sequence model[13]. The novel classification schemes were impossible for conventional machine learning paradigms. Attempts on deep learning had yielded exciting results, although training models from scratch required a huge amount of training data and computational resources[14].

In practice, however, some sleep studies may only focus on a small cohort, in which case the network's performance would decline significantly. Besides, different in monitor devices and specific experimental motivations cause channel mismatch, limiting model application across tasks[15]. To solve the above problems, this work proposes a deep learning

This work is supported by the scholarship from China Scholarship Council (Nos. 201606060227).

approach that consists of a very low number of layers, thus resulting in low computation cost compared to other deep learning approaches. Moreover, the proposed approach constructs an end-to-end structure without computing spectrograms or hand-crafted features. One of the most key contributions of this study is that the proposed model can handle various numbers of input channels and several signal modalities from different datasets without changing the model structure and hyperparameters to accommodate channel mismatch.

The article is organized as follows: Section 2 details the experimental data and the proposed deep learning architecture. Section 3 demonstrates the performance of the proposed model. Section 4 discusses the results and limitations of this study. Finally, section 5 gives conclusions.

II. Materials and Methods

A. Sleep Datasets

This study employed two common datasets to evaluate the proposed deep-learning architecture. The first one was from the Sleep Heart Health Study (SHHS)[16], in which only the first round (SHHS-1) was selected in this study. The SHHS dataset was a multi-center cohort study to investigate whether sleep-disordered breathing was associated with a higher risk of various cardiovascular diseases. Subjects employed in the present study were selected by restricting the Respiratory Disturbance Index 3 Percent (RDI3P) < 5 to have near-normal characteristics. In addition, the selected subjects did not use beta-blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease, or history of stroke.

The second one was the Sleep-EDF dataset[17], [18], of which the Sleep Cassette (SC) subset was adopted. It consisted of 20 subjects aged 25-34 years. Each subject had two PSG recordings from two consecutive day-night periods, except for one subject (subject 13) who had only the first-night data. PSG recordings from the second night were employed in this study, and thus 19 recordings were included. TABLE I summarized the characteristics of employed recordings.

Each recording was scored by an experienced research assistant or sleep technologist according to R&K rules. Sleep recordings were segmented into 30-second per epoch and labelled as wakefulness (W), non-rapid eye movement stage (NREM, containing N1, N2, N3 and N4) and rapid eye movement stage (R). According to the recently updated AASM standard, NREM stages 3 and 4 were merged into N3 in the present article. As signals sampled at different rates, we up-sampled those with sampling rates lower than 125 Hz to accommodate data from different datasets. In order to remove noise and artefacts, a simple filtering process, including notch filters, low-pass filters and high-pass filters, was performed. The filtered frequency bands of each signal were summarized in TABLE II. In addition, the long awake period before and after sleep was trimmed so that the number of awake epochs was not dominant in sleep cycles. Except that, the whole sleep recording was fully included in the analysis without discarding any recorded segments, thereby to have a near-clinical situation. To eliminate individual differences, sleep signals were normalized by mapping its mean to 0 and its deviation to 1. Then, signals were divided into 30-second per epoch, and each epoch corresponds to one sleep stage.

TABLE I. THE DESCRIPTION OF SUBJECTS FROM TWO DATASETS.

Parameters		SHHS	Sleep-EDF
Subjects		100	19
Age		46.86 (4.22)	28.74(2.99)
Power Frequency		60Hz	50Hz
Employed Channels		C3, C4, EOGR, EOGL, EMG, ECG	Fpz-Cz, Pz-Oz, EOG (horizontal)
Amplitude	***EEG***	[-26.0, 20.7]	[-208.2, 204.5]
	EOG	[-17.3, 17.7]	[-478.5, 449.6]
	EMG	[-22.3, 22.0]	--
	ECG	[-39.0, 44.2]	--
Sampling Freq.	***EEG***	125Hz	100Hz
	EOG	50Hz	100Hz
	EMG	125Hz	1Hz
	ECG	125Hz	--

B. Model Architecture

The proposed method expands raw EEG signals to multi-modality PSG signals consisting of EEG, EOG, EMG and ECG. The idea is to mimic the way sleep experts perform manual sleep scoring. When sleep experts label a 30-second PSG epoch, they visually inspect amplitude and frequency characteristics of EEG signals and sleep-related events such as spindles and K-complexes [1]. They also check eye movements and muscle activity levels as a reference for labelling some stages, such as stage R[1], [2]. Recent studies have revealed that analysis of heart-rate variations enables us to track the transition from wakefulness to sleep[19]. Similarly, the proposed model jointly processes multi-modality signals, thereby providing a comprehensive analysis.

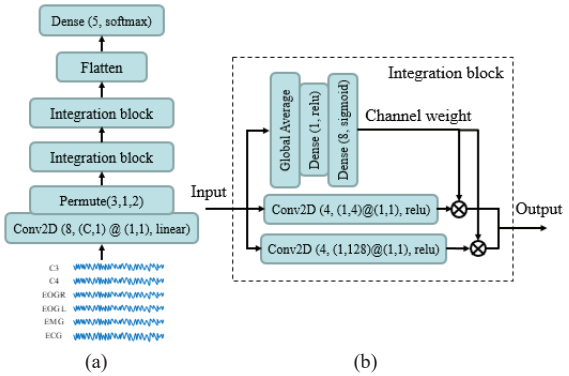
The proposed model, shown in Fig. 1, utilizes CNNs to extract features from raw PSG signals. The size of input data is $N \times 3750 \times C \times 1$ where N is the number of samples and C is the number of input channels. The first convolutional layer filters the input data using 8 kernels of size $C \times 1$ with a stride size of 1 point. The activation function of the first layer is a time-independent linear operation that projects diverse inputs into an optimal virtual space by adjusting weights and biases during model training. The dimension of virtual space is determined by the kernels of the first convolutional layer. To reduce the change of model parameters in transfer learning, we fix the dimension of virtual space. A permutation layer[20] is followed to hold channel information of virtual space and to transfer subsequent operations to the time domain.

The third layer is an integration block with three key components: a "squeeze and excitation" block to estimate channel weight[21], a convolutional layer with a smaller size to capture local features and a convolutional layer with a larger size for capturing the big context. Given the local receptive field of convolution operations, a global average pooling is used to squeeze global information which is then excited to generate channel-wise statistics[21]. We employ two CNNs

TABLE II. FILTERED FREQUENCY BANDS FOR EACH SIGNAL.

Signals	Frequency Band
EEG	0.5Hz-30Hz
EOG	0.5Hz-10Hz
EMG	10Hz-fs/2[a]
ECG	0.5Hz- 30Hz

[a] fs denotes sampling frequency.



(a) (b)

Fig. 1. Overview of the proposed architecture.

with small and large filter sizes to extract nonlinear features from the input data. Previous research[22] found that smaller filters are better to capture local contexts (e.g., when certain EEG patterns appear), while larger filters are better to capture big contexts. The outputs of these two CNNs are weighted and then concatenated into the final output of integration block. Two integration modules are adopted, each followed by a max-pooling layer with a size of (1,4), a dropout layer with a drop rate of 0.15 and a batch-normalization layer.

There is a dropout layer with a drop rate of 0.5 before the decision layer. The decision layer is a fully-connected layer with 5 units, which is activated by the softmax function. The output of the proposed architecture is a probability matrix with size $N \times 5$, where N is the number of samples and 5 is the number of sleep stages. The stage prediction for each sample corresponds to the stage with the maximum probability in the probability matrix.

C. Model training

Only PSG recordings from the SHHS dataset were used to determine the structure and hyperparameters of the proposed model. The whole dataset was split into train, validation and test sets. We used PSG recordings from 20 subjects as the final test set, and recordings from 80 subjects for training and validation. It should be noted that only the data from the training set and validation set was used in the process of parameters selection and model training, and thereby the test set was completely independent. In order to find the best hyperparameters for the proposed architecture, we performed a random search using a Python package named hyperopt[23]. For each set of hyperparameters, we used 5-fold cross-validation to train and evaluate the classifier (64 subjects for training and 16 subjects for evaluation). TABLE III summarized the distribution of each hyperparameter value. The parameter set leading to the highest accuracy and the least variability was adopted. Finally, the optimal model was achieved by using Adam optimizer with a learning rate of 0.002 and a batch size of 256. The network was trained by minimizing categorical cross-entropy. The code was written in Keras[24] with a Tensorflow backend[25].

TABLE III. DISTRIBUTION OF HYPERPARAMETERS.

Hyperparameter		Distribution
First CNN	***Filters***	[4, 6, 8, 16, 32]
	Strides	[1, 2, 3, 5, 7]
Integration Block	***Filters***	[4, 8, 16, 32, 64, 128]
	Smaller Kernel size	[2, 4, 8, 16, 32]
	Bigger Kernel size	[32, 64, 128, 256, 512]
	Strides	[1, 2, 3, 5, 7]
	Activation	{'relu',' tanh'}
Pooling Size		[2, 3, 4, 5]
Dropout Rate		[0.05, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5]
Learning Rate		[0.001, 0.002, 0.003, 0.004, 0.005, 0.01]
Optimizer		{'Adam', 'SGD'}
Batch Size		[64, 128, 256, 512]

D. Transfer learning on small datasets

Usually, training of deep learning networks required large amounts of data, which was expensive and difficult for many sleep studies. Thus, model transferability became crucial because it made deep learning research on a small cohort a reality. To demonstrate the transferability of the proposed model, we evaluated model performance with the SHHS dataset as the source domain and the Sleep-EDF dataset as the target domain. The proposed model was firstly trained using data from six channels of 80 subjects from the SHHS dataset and then transferred the model to three channels of data from the Sleep-EDF dataset. Note that we adopted not only different signal modalities but also different numbers of channels for the source and target domains on purpose because we wanted to enforce more channel mismatches.

To evaluate the efficiency of transfer learning in the target domain, a leave-one-out cross-validation was conducted. It means that for each iteration, there are 18 recordings for fine-tuning the entire pre-trained network and one independent recording for testing. It's worth mentioning that fine-tuning does not change the model structure and hyperparameters, but only adjusts weights and biases of notes. That iteration repeats 19 times. The aggregated performance of 19 recordings will be reported in the next section.

III. PERFORMANCE ASSESSMENT

Model performance was evaluated by accuracy, sensitivity, precision, Cohen's kappa and F1 score. The detailed definition can be found in previous studies[7], [26], [27].

A. Performance on the SHHS dataset

To illustrate model performance, TABLE IV showed the confusion matrix obtained by test subjects from the SHHS dataset, where we can verify the distribution of samples that were correctly or incorrectly classified. As can be seen from Table IV, the overall classification accuracy was 85.2%, which exceeded the accepted benchmark $Acc = 80\%$ among trained human scorers[28]. The most correctly classified stage was wakefulness with a precision of 92.6%. It was followed by N3 (88.0%), R (86.2%) and N2 (85.1%). Stage N1 was the hardest to classify with 35.2% of samples correctly assigned. 34% of samples were misclassified as N2, 19% as R and 12% as W. That result was consistent with previous results. Stage N1 was considered as a transition state between wakefulness and "real" sleep, thereby including information from two or three sleep stages. As a result, the scoring of N1 was quite obscure, even for sleep scoring experts[29]. Closer inspection of TABLE IV showed that most misclassifications occurred in contiguous stages in the sleep cycle. For example, N3 was most often misclassified as N2, and rarely as N1. This error was mainly due to similar electrophysiological characteristics between adjacent stages, rather than defects of model design.

TABLE IV. CONFUSION MATRIX OF TEST RECORDINGS FROM THE SHHS DATASET.

		Technologists' score stage					Pre.
	Stage	***W***	***N1***	***N2***	***N3***	***R***	
Proposed	***W***	3369	68	119	3	80	92.6%
	N1	125	187	92	0	128	35.2%
	N2	139	184	7986	804	275	85.1%
	N3	0	0	325	2392	0	88.0%
	R	34	106	467	2	3018	83.2%
Sen.		91.9%	34.3%	88.8%	74.7%	86.2%	
Acc.							85.2%

B. Transfer learning

Model generalizability was tested on the Sleep-EDF dataset. As shown in TABLE I, the available channel, amplitude distribution and acquisition environment were significantly different between the two datasets, which may limit the use of models proposed by some studies. TABLE V displayed the performance of the transfer learning scenario (indicated by Sleep-EDF[b]) compared to the model trained from scratch using only the Sleep-EDF data (shown by Sleep-EDF[a]). The results in TABLE V showed that the proposed model outperformed all other state-of-the-art results on the Sleep-EDF dataset, whether training from scratch or using transfer learning. Moreover, despite the serious channel mismatch, transferring the knowledge of the source domain (the SHHS dataset) to the target domain (the Sleep-EDF dataset) brought up the accuracy compared to a network trained from scratch using only data from the target domain.

TABLE V also listed the model architecture, its approach, input channels, input types, the number of subjects and other parameters for comparison. As can be seen from TABLE V, our method achieved a comparable or better performance compared to state-of-the-art methods that used the same dataset but more complex model structure. For example, the deep-learning architecture proposed in Sors et al.'s study [30] had up to 10^6 parameters, while the proposed architecture did not exhibit more than 10^4 parameters. Note that this was at least two orders of magnitude lower than the architecture proposed by Sors et al. Moreover, to the best of our knowledge, the proposed model was the first one that can handle different numbers of input channels without changing model structure and hyperparameters. The compact and versatile structure was conducive to clinical applications.

IV. DISCUSSION

In this work, we presented a convolutional network to automatically classify sleep stages which would help alleviate the burden of practitioners. Most automatic methods reported so far were based on hand-crafted features or designed for certain datasets. Thus, it was hard to find a method that generalized correctly to other datasets, especially when the channel did not match. To solve this problem, we proposed a compact end-to-end recognition structure that can handle various numbers of input channels and several signal modalities without changing any layer or hyperparameter values. Experimental results have demonstrated that the proposed model exhibited strong classification performance and low computational cost on both datasets compared with state-of-the-art results. More importantly, the proposed model showed potential transferability on data with different channels. As shown in TABLE V, transfer learning brought a slight accuracy improvement compared to a network trained from scratch using only data from the target domain, and the authors believed that using a huge and high-quality source dataset contributes to performance improvement of transfer learning.

Few studies tested their models on recordings collected from different record environments and hardware platforms. Zhang et al. [28] did so, demonstrating generalizability by testing model on two novel datasets without using transfer learning. In this article, it was impossible to directly test model performance on the Sleep-EDF dataset due to different channel numbers and signal modalities. Phan et al.[15] evaluated different fine-tuning strategies of transfer learning using the SeqSleepNet+ model and the DeepSleepNet+ architecture. However, they only used two signal modalities (EEG and EOG) and the same number of input channels. The proposed model can simultaneously handle four commonly used signal modalities without limiting the number of input channels. In addition, we adopted a "squeeze and excitation" block[21] to adaptively recalibrate channel-wise feature responses, thereby making full use of channel information.

TABLE V. PERFORMANCE COMPARISON.

Ref.	**Dataset**	**Subject**	**Input Channel**	**Input Type**	**Deep Learning Architecture**		**Approach**	**Result**			
					Structure	***Layer***		***Accuracy***	***K***	***Macro F1***	***Micro F1***
Ref[31]	SHHS	1000	EEG: C3, C4 EMG 2 EOGs	Time series	1DCNN	37 CNN	One-to-one	0.78	0.83	0.76	--
Ref [30]	SHHS-1	5728	C4-A1	Time series	1DCNN	12 CNN	Many-to-one	0.87	0.81	0.78	0.87
Ref [32]	Sleep-EDF	20	Fpz-Cz	Time series	CNN+LSTM	--	Many-to-one	0.84	0.78	0.78	--
Ref [12]	Sleep-EDF	20	EEG EOG	Spectrogram	2DCNN	2 CNN	One-to-many	0.82	0.75	0.75	--
Proposed	SHHS	100	EEG: C3, C4 2 EOGs EMG ECG	Time series	2DCNN	5 CNN	One-to-one	0.85	0.79	0.76	0.85
	Sleep-EDF [a]	19	EEG: FpzCz, PzOz 1 EOG					0.84	0.77	0.78	0.84
	Sleep-EDF [b]	19	EEG: FpzCz, PzOz 1 EOG					0.85	0.79	0.80	0.85

LSTM: Long short-term memory unit which is an artificial recurrent neural network (RNN) architecture.

Sleep-EDF [a]: The model was trained from scratch.

Sleep-EDF [b]: The model was fine-tuned for the Sleep-EDF dataset.

Even though our results are encouraging, the proposed model is still subject to several limitations. Firstly, our model requires to be trained with a sufficient amount of sleep data to improve generalization. Secondly, it is worth to explore the addition of different deep learning modules, such as a combination of CNN and LSTM, since studies have found that the use of temporal context can significantly improve model performance. Thirdly, more fine-tuning strategies will be explored in our future work. For example, training only the first and decision layers can help speed up the use of transfer learning in the target domain.

V. CONCLUSION

This paper proposed an automatic sleep scoring model based on raw PSG recordings. The deep-learning network was composed of two parallel convolution layers with different filter sizes for capturing both fine and coarse temporal features, and a "squeeze and excitations" block to recalibrate channel-wise feature responses. Experiments on two common sleep datasets showed that the model achieved comparable performance and low computational cost compared to state-of-the-art methods. In addition, our results proved that the proposed model was able to handle various numbers of input channels and several signal modalities from different datasets without changing model architecture and hyperparameters. The versatile model can be integrated with diverse sleep monitoring devices, thereby facilitating sleep research in clinical or routine care.

ACKNOWLEDGEMENT

The authors would like to thank the Sleep Heart Health Study(SHHS)[16] and Physionet[17], [18] for providing the polysomnographic data.

REFERENCES

[1] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," Washingt. DC US Natl. Inst. Heal. Publ., 1968.

[2] R. B. Berry et al., "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events," J. Clin. Sleep Med., vol. 8, no. 5, pp. 597–619, 2012.

[3] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," J. Neurosci. Methods, vol. 271, pp. 107–118, 2016.

[4] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," Comput. Methods Programs Biomed., vol. 176, pp. 81–91, 2019.

[5] K. Šušmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," Artif. Intell. Med., vol. 44, no. 3, pp. 261–277, 2008.

[6] S. Özşen, "Classification of sleep stages using class-dependent sequential feature selection and artificial neural network," Neural Comput. Appl., vol. 23, no. 5, pp. 1239–1250, 2013.

[7] R. Yan et al., "Multi-modality of polysomnography signals' fusion for automatic sleep scoring," Biomed. Signal Process. Control, vol. 49, pp. 14–23, 2019.

[8] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "DOSED: a deep learning approach to detect multiple sleep micro-events in EEG signal," J. Neurosci. Methods, vol. 321, pp. 64–78, 2019.

[9] Z. Mousavi, T. Yousefi Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi, "Deep convolutional neural network for classification of sleep stages from single-channel EEG signals," J. Neurosci. Methods, vol. 324, pp. 108312, 2019.

[10] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," Sleep, vol. 41, no. 5, pp. zsy041, 2018.

[11] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," arXiv Prepr. arXiv1610.01683, 2016.

[12] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 1285–1296, 2019.

[13] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, 2019.

[14] A. Malafeev et al., "Automatic human sleep stage scoring using deep neural networks," Front. Neurosci., vol. 12, pp. 781, 2018.

[15] H. Phan et al., "Towards more accurate automatic sleep staging via deep transfer learning," arXiv Prepr. arXiv1907.13177, 2019.

[16] D. A. Dean et al., "Scaling up scientific discovery in sleep medicine: The National Sleep Research Resource," Sleep, vol. 39, no. 5, pp. 1151–1164, 2016.

[17] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[18] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," IEEE Trans. Biomed. Eng., vol. 47, no. 9, pp. 1185–1194, 2000.

[19] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," Physiol. Meas., vol. 36, no. 10, pp. 2027–2040, 2015.

[20] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 4, pp. 758–769, 2018.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. IEEE Conf. Comput. Vis. pattern Recognit., pp. 7132–7141, 2018.

[22] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, 2017.

[23] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," 30th Int. Conf. Mach. Learn. ICML 2013, no. PART 1, pp. 115–123, 2013.

[24] F. Chollet, "Keras: Deep learning library for theano and tensorflow," URL: https//keras. io/k, vol. 7, no. 8, pp. T1, 2015.

[25] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, pp. 265–283, 2016.

[26] E. Fernandez-Blanco, D. Rivero, and A. Pazos, "Convolutional neural networks for sleep stage scoring on a two-channel EEG signal," Soft Comput., vol. 24, no. 6, pp. 4067–4079, 2020.

[27] R. Yan, F. Li, X. Wang, T. Ristaniemi, and F. Cong, "An automatic sleep scoring toolbox: multi-modality of polysomnography signals' processing," ICETE 2019 - Proc. 16th Int. Jt. Conf. E-bus. Telecommun., vol. 1, pp. 301–309, 2019.

[28] L. Zhang, D. Fabbri, R. Upender, and D. Kent, "Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks," Sleep, vol. 42, no. 11, pp. 1–10, 2019.

[29] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," Artif. Intell. Med., vol. 53, no. 1, pp. 25–33, 2011.

[30] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," Biomed. Signal Process. Control, vol. 42, pp. 107–114, 2018.

[31] I. Fernández-Varela, E. Hernández-Pereira, and V. Moret-Bonillo, "A convolutional network for the classification of sleep stages," Multidiscip. Digit. Publ. Inst. Proc., vol. 2, no. 18, pp. 1174, 2018.

[32] S. Back, S. Lee, H. Seo, D. Park, T. Kim, and K. Lee, "Intra- and Inter-epoch Temporal Context Network (IITNet) for automatic sleep stage scoring," arXiv Prepr. arXiv1902.06562, 2019.

V

AUTOMATIC SLEEP SCORING: A DEEP LEARNING ARCHITECTURE FOR MULTI-MODALITY TIME SERIES

by

Rui Yan, Fan Li, Dongdong Zhou, Tapani Ristaniemi & Fengyu Cong 2020

Manuscript submitted to Journal of Neuroscience Methods, Under Review.

Automatic Sleep Scoring: A Deep Learning Architecture for Multi-modality Time Series

Rui Yan[a,d], Fan Li [a], DongDong Zhou[a,d], Tapani Ristaniemi[d], Fengyu Cong[a,b,c,d]

[a] *School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China*

[b] *School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China*

[c] *Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province. Dalian University of Technology, 116024, Dalian, China*

[d] *Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland*

Abstract

Background: Sleep scoring is an essential but time-consuming process, and therefore automatic sleep scoring is crucial and urgent to help address the growing unmet needs for sleep research. This paper aims to develop a versatile deep-learning architecture to automate sleep scoring using raw polysomnography recordings.
Method: The model adopts a linear function to address different numbers of inputs, thereby extending model applications. Two-dimensional convolution neural networks are used to learn features from multi-modality polysomnographic signals, a "squeeze and excitation" block to recalibrate channel-wise features, together with a long short-term memory module to exploit long-range contextual relation. The learnt features are finally fed to the decision layer to generate predictions for sleep stages.
Result: Model performance is evaluated on three public datasets. For all tasks with different available channels, our model achieves outstanding performance not only on healthy subjects but even on patients with sleep disorders (SHHS: Acc-0.87, K-0.81; ISRUC: Acc-0.86, K-0.82; Sleep-EDF: Acc-0.86, K-0.81). The highest classification accuracy is achieved by a fusion of multiple polysomnographic signals.
Comparison: Compared to state-of-the-art methods that use the same dataset, the proposed model achieves a comparable or better performance, and exhibits low computational cost.
Conclusions: The model demonstrates its transferability among different datasets, without changing model architecture or hyper-parameters across tasks. Good model transferability promotes the application of transfer learning on small group studies with mismatched channels. Due to demonstrated availability and versatility, the proposed method can be integrated with diverse polysomnography systems, thereby facilitating sleep monitoring in clinical or routine care.
Keywords: polysomnography; automatic sleep scoring; multi-modality analysis; deep learning

1. Introduction

Sleep is a vital physiological process as it covers approximately one-third of the human lifespan. Adequate and high-quality sleep is essential for physical restoration[1], memory processing[2] and metabolism[3]. Nowadays, probably due to our hectic lifestyle in modern society, complaints about sleep problems increase dramatically among people. An effective way to monitor sleep quality and diagnose sleep problems is overnight polysomnographic (PSG) test. The PSG test simultaneously records dozens of sleep signals, including electroencephalograms (EEG), electrooculogram (EOG), electromyograms (EMG), electrocardiogram (ECG), airflow and respiratory effort. These recorded signals are generally analyzed by sleep experts based on the R&K rules[4] and recently updated American Academy of Sleep Medicine (AASM) standard [5].

Based on the amplitude and frequency characteristics of PSG signals, the R&K rules divide sleep into five distinct stages: non-rapid eye movement (NREM) stages 1, 2, 3 and 4 and rapid eye movement stage (stage R). The most recent AASM standard merges NREM stages 3 and 4 into N3 due to their prevalent low-frequency oscillations in EEG signals. Assigning a sleep stage to each sleep segment, called sleep scoring, is a very important step in any sleep research. However, the manual sleep scoring is labor-intensive and subjective. Previous studies have reported that the annotation of an 8-h recording requires approximately 2-4 hours[6], and the inter-scorer reliability of sleep scorings is about 0.8[7]. Therefore, automatic scoring is deemed as a promising approach due to its cost efficiency and high precision.

Numerous attempts[8] so far have been made in the field of automatic sleep scoring. Scoring methods based on conventional machine-learning methods were prevalent, which usually included two main components: feature extraction and classification. There was a wide variety of techniques for feature extraction, including but not limited to statistic methods[9], Fourier transforms[10], wavelet analysis[11] and Hilbert transform[12]. These techniques were responsible for describing sleep signals from multiple aspects. In order to obtain an evaluation of sleep stages, these extracted features were then fed to a classifier[13], such as support vector machine[14], random forest[15], K-nearest neighbor classifier [16], Naive Bayes[10], artificial neural network[17]. These studies' accuracy ranged from 0.8 to 0.9 and highly depended on the validity of employed features.

Recently, approaches based on deep learning have sprung up since it avoided explicit feature extractions commonly seen in conventional machine-learning methods, and was especially suitable for big data approach[18]. Mousavi et al. [19] proposed a convolutional neural network (CNN) to automate sleep scoring using EEG time series, which achieved competitive performance in the classification of 2 to 6 classes of sleep stages. Instead of raw signal inputs, time-frequency images, generated by short-time Fourier transform [20] or

wavelet transformation[21], were also explored in several studies. Zhang et al. [22] even compared these two different input representations and concluded that the network performance using the spectrogram as inputs was superior to that using time series as inputs, which was attributed to the compact information and less artifact in the spectrogram. Although CNN gave the most convincing performance in some fields, for example, computer vision and image recognition, it still suffered from some problems, such as tricky hyper-parameters, feature redundancy, and vanishing gradients[23], which challenged the construction of deep convolutional networks.

Recurrent neural networks (RNN) was also important in deep learning because of their good performance in capturing temporal correlations of inputs[24]. One of the most popular was the long short-term memory network (LSTM) that solved the problem of vanishing gradients and long-term dependence in traditional RNN. The LSTM module had made great progress in the application of natural language processing[25]. In the field of automatic sleep scoring, some studies had revealed that the application of LSTM module helped to capture the inter-segments temporal context[26]. However, the LSTM module required to calculate a lot of parameters and was prone to overfitting. In practical applications, the LSTM module was usually combined with CNN modules[27] or conventional techniques of feature extraction[28] to compress the input, thereby saving computational cost.

Moreover, studies based on deep learning had introduced some novel classification schemes to mimic the way sleep experts performed in manual sleep scoring, such as one-input to multi-output schemes[29] and sequence-to-sequence models[30]. These novel schemes explicitly utilized the dependence of consecutive segments, which were impossible for conventional machine learning paradigms. According to their experiment results, the long-term dependence between segments led to significant performance improvement. In short, attempts on deep learning had yielded exciting results, although training models from scratch required a huge amount of training data and computational resources[31].

However, in terms of conventional machine-learning methods in automatic sleep scoring, their classification performances highly rely on extracted features. The elaborate features may underperform in other datasets, thus limiting model generalizability. For automatic sleep scoring methods based on deep learning, most models are designed for specific datasets and certain input signals, which requires task-specific modification when their models are used in different tasks. Moreover, that modification is difficult and even inefficient, especially for sleep studies focused on a small group because of insufficient training data. In practical applications, differences in monitor device and experimental design induce channel mismatch, which challenges the application of transfer learning[32]. To tackle the above problems, this work proposes a simple but versatile deep learning architecture that does not require task-specific modifications to model architecture or hyper-parameter. The proposed architecture employs very few numbers of layers, thus resulting in low computation cost compared to other deep learning approaches. The main contributions of this work are presented as follows.

a) A deep learning architecture is proposed to automate sleep scoring using multi-modality PSG signals.
b) A linear activation function is adopted in the first CNN layer to accommodate different numbers of input channels, which helps to address channel mismatches.
c) One LSTM module and two CNN modules with different kernels sizes are employed to capture information across temporal and spatial scales.
d) The proposed model achieves good performance on three disparate datasets with different subject attributions, thereby demonstrating model generalizability on different disease populations.
e) Model transferability is demonstrated across three datasets with different input channels and signal modalities.

The article is organized as follows: Section 2 presents details of experimental data and the proposed deep learning architecture. Section 3 demonstrates the performance of the proposed model. Section 4 discusses the results and limitations of this study. Finally, section 5 gives conclusions.

2. Methodology

2.1 Data description

This study adopted three public datasets to evaluate model performance. The first one was from the Sleep Heart Health Study (SHHS)[33], in which only the first round (SHHS-1) was selected in this study. The SHHS dataset recruited thousands of participants from nine existing epidemiological studies to investigate the relationship between sleep-disordered breathing and various cardiovascular diseases. A total of 100 subjects were selected out by restricting the respiratory disturbance index (RDI3P) < 5 to have near-normal characteristics. Besides, the selected subjects did not use beta-

Table 1 . Subject characteristics.

Para.		SHHS	ISRUC	Sleep-EDF
Subjects		100	99	19
Attribute		Near-health	Sleep disturbance	Health
Age		46.86 ±4.22	51±16	28.74±2.99
Criterion		R&K	AASM	R&K
Power Frequency		60Hz	50Hz	50Hz
Employed Channels		C3, C4, EOGR, EOGL, EMG, ECG	F3, C3, O1, F4, C4, O2, ROC, LOC, EMG, ECG	Fpz-Cz, Pz-Oz, EOG (horizontal)
Amplitude	EEG	[-26.0, 20.7]	[-149.4, 151.8]	[-208.2, 204.5]
	EOG	[-17.3, 17.7]	[-138.0, 146.1]	[-478.5, 449.6]
	EMG	[-22.3, 22.0]	[-524.7, 518.8]	--
	ECG	[-39.0, 44.2]	[-145.3, 121.0]	--
Sampling	EEG	125Hz	200Hz	100Hz
	EOG	50Hz	200Hz	100Hz
	EMG	125Hz	200Hz	1Hz
	ECG	125Hz	200Hz	--

Note: Unless specifically indicated, the above EEG channels were referred to the left or the right mastoids (M1 or M2) according to the 10–20 international electrode placement system.

blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease and stroke.

The second one was ISRUC-Sleep dataset[34], of which subgroup 1 was chosen in the present article. This subgroup included 100 PSG recordings from healthy subjects, patients with sleep disorders and patients under the effect of sleep medication. Subject 8 was excluded due to the lack of required channels, and therefore only 99 subjects were analyzed in the following experiments. Each recording was visually labelled by two sleep experts according to the AASM standard[5]. To improve signal quality, dataset providers had filtered all signals by a 50Hz notch filter. In addition, the signals of EEG and EOG were filtered between 0.3Hz and 35Hz, and EMG signals were filtered between 10Hz and 70Hz.

The third dataset was the Sleep-EDF dataset[35], [36], in which the sleep cassette (SC) subset was adopted. It consisted of 20 healthy subjects whose age ranged from 25 years old to 34 years old. Each subject had 2 PSG recordings about 20 hours each, except for subject 13 who had only the first-night recording. The recorded two PSG recordings for each subject were from two consecutive day-night periods at subjects' home. To avoid "the first night effect", PSG recordings from the second night were employed in the present study, and thus a total of 19 recordings were analyzed. Table 1 summarized the characteristics of employed recordings, where the age was shown as mean age ± standard deviation.

To accommodate data from different datasets, all signals were sampled or resampled to 125Hz. In order to remove noise and artefacts, all signals were filtered by a notch filter, a high-pass filter and a low-pass filter. The effective frequency band of EEG and ECG signals was limited to 0.5Hz-30Hz, 0.5Hz-10Hz for EOG signals, and only information above 10Hz was retained for EMG signals. In addition, for recordings in the Sleep-EDF dataset, the long awake period before and after sleep was trimmed to restrict our analysis to nocturnal sleep. In order to minimize the variability between recordings, each signal was normalized by mapping its mean to 0 and its deviation to 1. Table 1 displayed the amplitude range of signals after preprocessing. Afterwards, all the signals were divided into 30-second segments, each segment corresponding to a single sleep stage. For PSG recordings scored using R&K rules, NREM stages 3 and 4 were merged into N3 in the present article according to the recently updated AASM standard.

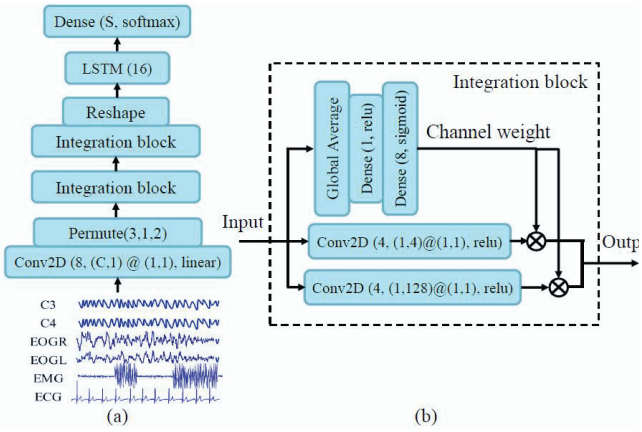


Figure 1. Overview of the proposed architecture.

2.2 Model Architecture

The proposed deep-learning architecture extends the input from EEG signals to a fusion of multiple PSG signals. The idea imitates the way sleep experts perform manual sleep scoring. Besides the characters of EEG signals, sleep experts also check eye movements and muscle activities as a reference when they label a 30-second PSG segment[4], [5]. For example, Stage R is characterized by low-amplitude and mixed-frequency EEG activities, rapid eye movements and the lowest EMG activity level; Stage N3 is marked by high-amplitude slow waves and rare eye movement. Recent studies have revealed that analysis of cardiac electrophysiological activity enables us to track the transition from wakefulness to sleep[37]. Hence, there is every reason to believe that the joint processing of multiple PSG signals contributes to comprehensive sleep analysis.

Figure 1 shows a schematic diagram of the proposed model to provide an intuitive manner to capture the model structure. Detailed model parameters and layer outputs are provided in Table 2. As can be seen from Figure 1, the proposed architecture comprises several CNN modules and one LSTM module to extract spatial and temporal features from raw PSG signals. The size of the input data is $T \times C \times 1$ where T is the number of time points and C is the number of input channels. Since the sampling rate of employed signals is set to 125Hz and each sample lasts 30 seconds, hence, $T = 3750$ in the present article. The proposed architecture does not restrict the number of input channels C which may be diverse in different datasets.

The first convolution layer filters the input data using 8 kernels of size $C \times 1$ with the stride size of 1 point. The activation function of the first layer is a time-independent linear operation. Similar linear functions have been used as spatial filters in the study of Chambon et al. [38]. Here, we use linear functions to accommodate mismatched input channels,

Table 2. Architecture detail

Layer	Type	Units	Size	Stride	Activation	Output size
Input						(3750, C, 1)
1	Conv2D	8	(1, C)	(1, 1)	linear	(3750, 1, 8)
2	Permute					(8, 3750, 1)
3	Integration block					(8, 3750, 8)
4	Max-pooling		(1, 16)			(8, 234, 8)
5	Dropout (0.15)					(8, 234, 8)
6	Batch normalization					(8, 234, 8)
7	Integration block					(8, 234, 8)
8	Max-pooling		(1, 16)			(8, 14, 8)
9	Dropout (0.15)					(8, 14, 8)
10	Batch normalization					(8, 14, 8)
11	Permute					(14, 8, 8)
12	Reshape					(14, 64)
13	LSTM	16			tanh	16
14	Dense	5			softmax	5

and thus subsequent model parameters can be free from the influence of varying numbers of input channels. The output of the first layer is a set of linear combinations of input signals. The optimal combinations can be achieved by adjusting weights and biases of kernels during model training. This operation can be considered as a projection that maps diverse inputs into the optimal virtual space, thereby compensating channel mismatch. In addition, in order to prevent the model from overfitting[39], we apply a L2 weight regularization with a value of 0.01 in the first convolution layer. A permutation layer[38] is followed to hold channel information of the virtual space and to transfer subsequent operations to the time domain.

The third layer is an integration block with three key components: a "squeeze and excitation" block to estimate channel weights, a convolution layer with a smaller kernel size to capture local features and a convolution layer with a larger kernel size for capturing the big context. In view of the local receptive field of convolution operations[40], global information is required to evaluate channel weights, which is achieved by a global average pooling. Two fully-connected layers followed the global average pooling is to excite the nonlinearity of among weights[40]. We employ two CNNs with small and large filter sizes to extract nonlinear features from its input. The previous study[39] has found that smaller kernels are better to capture local contexts (i.e., when certain of EEG patterns appear), while larger kernels are conducive to capturing big contexts. The outputs of the two CNN modules are weighted by channel-wise statistics and then concatenated into the final output of the integration block. Two integration modules are adopted, each followed by a max-pooling layer with a size of (1, 16), a dropout layer with a drop rate of 0.15 and a batch-normalization layer. Here, the large pooling size is to compress temporal information, thereby reducing model parameters and memory requirements. The layers of dropout and batch-normalization help to control overfitting.

The long short-term memory (LSTM) module is arranged before the decision layer to dig up long-range contextual information. A typical LSTM unit has a memory cell and three gates, namely an input gate, an output gate and a forget gate, to regulate the retention or discard of information flow. The unique mechanism allows LSTM units to selectively remember the previous information, thereby facilitating the current decision. Previous studies[39] have revealed that the context information helps to capture the transition rules among sleep stages. The conclusion is consistent with manual scoring rules[4], [5]. The transition rules allow sleep experts to predict possible sleep stages for the current segment based on a sequence of PSG segments. These transition rules are especially helpful for decision-making when signal characters of the current segment are ambiguous.

The final layer is the decision layer, which is a fully-connected layer activated by the softmax function. The number of units is equal to the number of classes. In the present article, we split sleep segments into five sleep stages, namely W, N1, N2, N3 and R, and therefore $S = 5$, The output of the decision layer is a probability matrix with size $N \times S$, where N is the number of samples (or sleep segments) and S is the number of sleep stages. The stage prediction for each sample corresponds to the stage with the maximum probability.

2.3 Hyper-parameter optimization

The selection of hyper-parameters was carried out on only the SHHS dataset via 5-fold cross-validation. The whole dataset was split into five subsets, each with 20 subjects. For a given hyper-parameter set, the proposed model was trained on data from 4 subsets and tested on data from the remaining subset. In addition, we used 20% of training data for model validation. This process was repeated 5 times, with each subset being used as test data once. The final performance on this hyper-parameter set was determined by the aggregated test performance across all five folds. It should be noted that once the optimal hyper-parameter set determined, it would be used in all experiments. Therefore, there was no task-specific modification to model structure and hyper-parameters, except for the kernel size of the first convolution layer that was determined by the number of input channels.

In order to find the best hyper-parameters for the proposed architecture, we performed a random search using a Python package named hyperopt[41]. The number of iteration was set to 50. The search space of hyper-parameters was summarized in Table 3. The parameter set leading to the highest accuracy and the lowest variability was adopted as the optimal parameters. If two sets of parameters gave a similar performance, the one with lower computational costs would be selected. Finally, the optimal model was achieved by using Adam optimizer with a learning rate of 0.002 and a batch size of 256. The network was trained by minimizing categorical cross-entropy. The code was written using the Keras package[42] with the Tensorflow backend[43].

3. Performance assessment

Model performance is evaluated by accuracy, precision, recall, F1 score and Cohen's kappa.

Accuracy (Acc.) measures the proportion of samples that the model correctly predicted.

Precision (P) is the fraction between true positives and the predicted positives.

Table 3. Search set for hyper-parameters

Hyper-parameter		Distribution
First CNN	Filters	[4, 6, 8, 16, 32]
	Strides	[1, 2, 3, 5, 7]
Integration Block	Filters	[4, 8, 16, 32, 64, 128]
	Smaller Kernel size	[2, 4, 8, 16, 32]
	Bigger Kernel size	[32, 64, 128, 256, 512]
	Strides	[1, 2, 3, 5, 7]
LSTM	Unit	[6, 8, 16, 32, 64, 128]
	Activation	{'relu',' tanh'}
Pooling Size		[2, 3, 4,...,15, 16]
Dropout Rate		[0.05, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5]
Learning Rate		[0.001, 0.002, 0.003, 0.004, 0.005, 0.01]
Optimizer		{'Adam', 'SGD'}
Batch Size		[64, 128, 256, 512]

Recall (R), also named sensitivity, calculates the percentage of actual positives that the model correctly identified.

F1 score (F1) represents the harmonic mean between precision and sensitivity.

Kappa (K) is an agreement measure between the proposed model and a human expert, which takes into account the chances of random agreement. A Large value indicates a high agreement between two classification results, and the perfect agreement gets a value of 1.

3.1 Classification performance

In order to illustrate model performance, Table 4 showed the aggregated confusion matrix from 5-fold cross-validation on the SHHS dataset. The confusion matrix clarified the distribution of samples that were correctly or incorrectly classified. From Table 4, we can see that the total classification accuracy was 0.87, which exceeded the accepted benchmark $Acc = 80\%$ among trained human scorers[7]. The best classification was wakefulness with the precision of 0.93. It followed by N2 (0.87), N3 (0.87) and R (0.83). Stage N1 was the hardest class to classify, with 31% of samples correctly assigned. There were 25% of N1 samples misclassified as R, 25% as N2 and 19% as W. The low precision of N1 stage was common in studies. Stage N1 was considered a transition state between wakefulness and "real" sleep, thereby including information from two or three sleep stages. As a result, the scoring of N1 was quite obscure, even for sleep scoring experts[44]. Closer inspection of Table 4 showed that most misclassifications occurred in contiguous stages in the sleep cycle. For example, N3 was often misclassified as N2, and rarely misclassified as N1. These misclassifications were mainly due to similar or mixed electrophysiological characteristics between adjacent stages, rather than the defect of model design.

To test the generalization capabilities, the proposed model was further evaluated on two independent datasets, the Sleep-EDF and the ISRUC dataset, in which subjects in the ISRUC study suffered from diverse sleep disorders. As shown in Table 1, the available channels, amplitude distributions and acquisition environment were significantly different among these three datasets. Besides, the model architecture and hyper-parameters were determined by recordings from the SHHS dataset, and they would remain unchanged in the classification of sleep segments from the other two datasets. In terms of the Sleep-EDF dataset, signals from three available channels (FpzCz, PzOz, EOG) were employed as model inputs, and a leave-one-out cross-validation was performed to evaluate model performance. For the ISRUC dataset, 10 available channels were adopted including six EEG, two EOG channels, one EMG channel and one ECG channel. Model performance was evaluated using 5-fold cross-validation to provide a generalized model evaluation. Table 5 and Table 6 presented the confusion matrix obtained on test recordings from the datasets of Sleep-EDF and ISRUC, respectively.

Comparing Table 4, Table 5 and Table 6, we can get that the proposed model gave outstanding performance on three disparate datasets, no matter healthy subjects from the Sleep-EDF dataset or patients with complex sleep disturbances from the ISRUC dataset. For recordings from the ISRUC dataset or Sleep-EDF dataset, the N1 stage got acceptable precision despite its small sample size, which further demonstrated that our method could tackle the problem of unbalanced classes.

Furthermore, we displayed the learning curve of the proposed model on three datasets. Figure 2 showed the changes of accuracy versus the number of iterations for

Table 4. Confusion matrix for test recordings from the SHHS dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	18925	575	456	9	281	0.93	0.93	0.93
	N1	330	937	493	0	224	0.47	0.31	0.37
	N2	712	740	38442	3599	882	0.87	0.90	0.88
	N3	21	0	1512	10662	1	0.87	0.75	0.81
	R	433	762	1947	2	15569	0.83	0.92	0.87
Accuracy									0.87
Kappa									0.81

Table 5. Confusion matrix for test recordings from the Sleep-EDF dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	2347	81	22	3	8	0.95	0.84	0.89
	N1	232	901	333	6	107	0.57	0.58	0.57
	N2	34	227	7430	234	95	0.93	0.86	0.89
	N3	5	8	392	2476	1	0.86	0.91	0.88
	R	166	347	422	3	3750	0.80	0.95	0.87
Accuracy									0.86
Kappa									0.81

Table 6. Confusion matrix for test recordings from the ISRUC dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	22804	905	174	19	127	0.95	0.94	0.94
	N1	1091	7021	1708	13	631	0.67	0.68	0.67
	N2	166	1703	24175	2196	410	0.84	0.88	0.86
	N3	18	19	865	12637	7	0.93	0.84	0.89
	R	246	749	612	132	7723	0.82	0.87	0.84
Accuracy									0.86
Kappa									0.82

training data and validation data. As it is seen, the network accuracy improved with increasing numbers of iteration (indicating by "Epoch Number" in Figure 2). Since model hyper-parameters were selected based on the data from the SHHS dataset, the convergence speed of the network was the fastest on the SHHS dataset, followed by the ISRUC dataset, and that on the Sleep-EDF dataset was the slowest. Nevertheless, using early stopping with a patience of 10 epochs to monitor the validation loss, the model training could complete within 100 iterations. Given the limited quantity of the Sleep-EDF dataset, the final accuracy was inferior to the SHHS dataset and the ISRUC dataset. In order to improve classification accuracy of the Sleep-EDF dataset, we applied a fine-tuning strategy, which would be introduced in detailed in section 3.4.

3.2 Model performance under different signals' fusions

In order to explore the effect of signal type on classification performance, we investigated different fusions of PSG signals. The analyzed signals included EEG, EOG, EMG and ECG. This experiment was performed on the ISRUC dataset due to its abundant channels. To provide an unbiased estimate of the model performance, we conducted a subject independent 5-fold cross-validation on 99 recordings. The results were shown in Figure 3, where the column represented the average accuracy of 5-fold cross-validation and the bar denoted the standard deviation. For the fusions of more than two signals, the signal name was abbreviated to its middle letter, such as C&E denoted the fusion of ECG signals and EEG signals, and M&O&E meant the fusion of signals of EMG, EOG and EEG.

As can be seen from Figure 3, abundant signals were conducive to improving accuracy and reducing uncertainty.

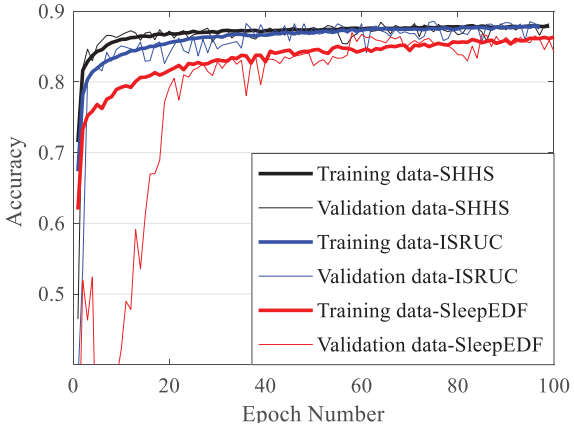


Figure 2. Train curve of three datasets

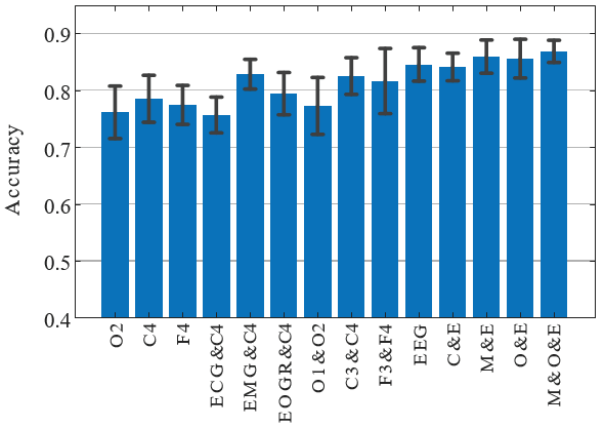


Figure 3. The classification accuracy for different signal fusions

Table 7. Performance comparison

Ref.	Dataset	Subjects	Input Channel	Input Type	Architecture		Approach	Result	
					Structure	*Layers*		*Acc.*	*Kappa*
Ref[48]	SHHS	1000	EEG: C3, C4 EOG: ROC, LOC EMG	Time series	1DCNN	37 CNN	One-to-one	0.78	0.83
Ref [47]	SHHS-1	5728	C4-A1	Time series	1DCNN	12 CNN	Many-to-one	0.87	0.81
Ref [45]	Sleep-EDF	20	Fpz-Cz	Time series	CNN+LSTM	--	Many-to-one	0.84	0.78
Ref [29]	Sleep-EDF	20	EEG EOG	Spectrogram	2DCNN	2 CNN	One-to-many	0.82	0.75
Ref[46]	ISRUC	40	EEG: F3, C3, O1, F4, C4, O2 EOG: ROC, LOC EMG	Features	Random forest	--	--	0.82	--
Proposed	SHHS	100	EEG: C3, C4 EOG: ROC, LOC EMG ECG	Time series	2DCNN+LSTM	5 CNN	One-to-one	0.87	0.81
	ISRUC	99	EEG: F3, C3, O1, F4, C4, O2 EOG: ROC, LOC EMG ECG					0.86	0.82
	Sleep-EDF	19	EEG: FpzCz, PzOz EOG					0.86	0.81

Note: Unless specifically indicated, the above EEG channels were referred to the left or the right mastoids (M1 or M2) according to the 10–20 international electrode placement system.

Specifically, in the perspective of single-channel EEG inputs, time series from the C4 channel achieved the best performance with the mean accuracy of 0.78 and the standard deviation of 0.004. Time series from the O2 channel performed the worst, which may be attributed to the poor signal quality caused by uncomfortable electrodes location. Adding EEG channels or other PSG modalities enhanced model performance, but up to a certain extent. The fusion of EEG, EOG and EMG signals produced the best performance in this experiment, with the average accuracy of 0.87 and the standard deviation of 0.002. In terms of signal types, the performance of EMG signals and EOG signals was superior to ECG signals, likely due to the morphological difference of ECG signals.

3.3 Performance comparison

The performance of the proposed model was compared with recent studies that used the same datasets. Table 7 showed model performance, together with model architectures, their approaches, input channels, input types, subject numbers and other parameters for comparison. What stood out in Table 7 was that our method achieved a comparable or better performance compared to the state-of-the-art methods that used the same dataset but more complex model structure.

More specifically, for studies on the Sleep-EDF dataset, our model achieved an accuracy of 0.86 and a kappa value of 0.81, which exceeded 2% on accuracy and 3% on kappa value compared to the "many to one" classification scheme proposed by Back et al.[45]. For studies on the ISRUC dataset, there was a significant improvement (+4% on accuracy) between the proposed model and Khalighi et al.'s methods[46]. For studies on the SHHS dataset, our model obtained comparable performance with Sors et al.'s study. However, the deep-learning architecture proposed in Sors et al.'s study[47] employed 12 convolution layers and two fully-connected layers with about 10^6 parameters, while the proposed model exhibited about 10^4 parameters. Note that this was at least two order of magnitude lower than the model proposed by Sors and his colleagues. The compact structure helped saving training time and computational cost, thus facilitating clinical practice.

Moreover, few studies [29], [45]–[48] had tested their model on diverse datasets with different sample attributes, input channels and disease populations. The proposed model shows stable performance on three datasets with completely different attributes, indicating good model generalization in different datasets and sample populations.

3.4 Evaluation of model transferability

In order to test classification performance of the trained model against data that the model had never seen before, we tested model transferability among three datasets. In terms of channel-matched cases, six matched channels were extracted from the datasets of SHHS and ISRUC in this experiment. After the model was trained on one dataset, the trained model was directly used to predict sleep stages for recordings from the other dataset. It was worth noting that the trained model did not suffer any modification for test data. Table 8 showed the classification results. As can be seen from Table 8, the direct prediction achieved moderate classification accuracy,

Table 8. Model generalizability

Model		Direct predict		Fine tuning with 20 subjects	
		Acc.	K	Acc.	K
Training	SHHS	0.73	0.64	0.84	0.79
Testing	ISRUC				
Training	ISRUC	0.66	0.55	0.84	0.77
Testing	SHHS				

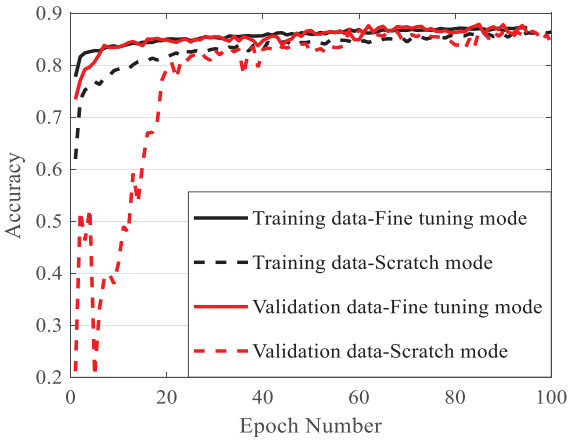


Figure 4. Train curve comparison between fine-tuning mode and scratch mode on the Sleep-EDF dataset

which may be attributed to the lack of huge training dataset. Nevertheless, fine-tuning the trained model with a small amount of test data, the accuracy can be significantly improved. In addition, the SHHS dataset contained near-healthy participants, while the ISRUC dataset involved patients with complex sleep disturbance. The results indicated good model transferability between different disease populations.

In the case with channel mismatch, the direct prediction was impossible. Here, we tried two classification strategies on the Sleep-EDF dataset: fine-tuning a trained model or training a new model from scratch. For a fair comparison, we used leave-one-out cross-validation and the same set of model parameters for these two classification strategies. The adopted model parameters were the same as those described in Section 2 and those used in previous experiments. The model for fine-tuning was trained on the SHHS dataset. Figure 4 displayed the learning curve of these two classification strategies. As can be seen from Figure 4, the fine-tuning strategy resulted in a faster and smoother convergence curve compared to that of the model trained from scratch. Classification performance improved by 1.6% on accuracy and 2.7% on kappa using the fine-tuning strategy. Table 5 showed the detailed confusion matrix under the fine-tuning strategy.

3.5 Model visualization

In order to illustrate how well each layer distinguishes sleep stages, we visualized layer outputs using t-Distributed Stochastic Neighbor Embedding (t-SNE)[49]. The t-SNE can

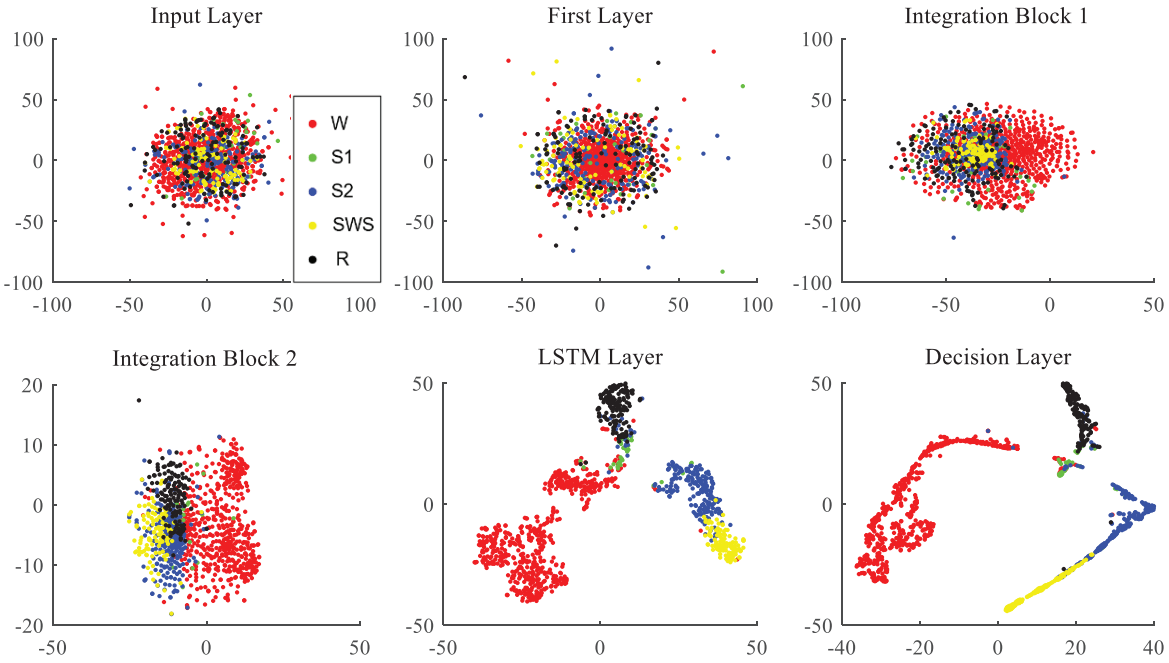


Figure 5. Model visualization using t-SNE method.

transform high-dimensional data into two-dimensional data to facilitate data visualization. Figure 5 displayed compressed layer outputs when the trained model predicted subject "shhs1-204846" from the SHHS dataset.

It can be seen from the first map in Figure 5 that the distribution of input data was random. The first layer with linear activation functions regularized the inputs. As moving forward from the integration block 1 to the decision layer, five sleep stages were more clearly separated. In particular, the LSMT layer led to a significant increase in separability, and the decision layer resulted in further clear separation.

4. Discussion

In this work, we developed a deep learning network for the automatic classification of sleep stages. Most of the automatic methods reported so far are based on human-engineered features or designed for a specific dataset. Thus, these models are hard to generalize correctly and easily to other datasets, especially when the channels do not match. To address these problems, we propose a compact and versatile end-to-end architecture to automate sleep scoring. We think two characteristics propel our model better than state-of-the-art methods. The first is good generalization and transferability of the proposed model. The above experiments have demonstrated that our model achieves strong classification performance on three disparate datasets, no matter whether it is from the healthy or the patients with sleep disturbance. This indicates good model transferability and generalization among different datasets and disease populations. The characteristic avoids cumbersome task-specific adjustments to model architecture and hyper-parameters, thereby facilitating clinical applications. Moreover, the proposed structure is conducive to the fine-tuning strategy, especially in the cases with limited training data and mismatched channels, which can significantly improve classification accuracy. Secondly, the proposed model exhibits a relatively low number of parameters, which drastically reduce training time, thereby saving computational resources.

The proposed architecture takes raw PSG signals as input without any human-engineered features, thereby preserving the coherence among multi-modality signals. There is no elaborate processing on raw PSG signals, except for a simple filtering process to improve the signal-to-noise ratio. Besides, the whole PSG recording is fully included in the analysis without discarding any recorded segments, even severely contaminated segments. The crude pre-processing enhances model robustness, and therefore the proposed model is more easily adaptable to noisy clinical applications. We have noticed that some studies claimed the network using raw PSG signals as inputs showed inferior performance[22] and was more prone to overfitting [32], compared with that using spectrograms as inputs. Therefore, we adopt several strategies to control overfitting, such as the L2-regularization in the first CNN layer, dropout layer and batch-normalization. To improve model performance, CNN modules with different kernel sizes and LSTM modules are employed to capture information across temporal and spatial scales. Experimental results prove the feasibility of these strategies.

The proposed model is capable of coping with multiple PSG signals. Experiments have demonstrated that the input of multi-modality signals is conducive to the improvement of model performance. This conclusion is consistent with the findings of our previous research[15] and the manual scoring standards[4], [5]. Sleep experts inspect multiple PSG channels, including EEG (records of brain activity), EOG (records of eye movement) and EMG (records of muscle activity). The additional EOG and EMG channels usually provide important information to distinguish sleep stages, especially when EEG

activity is ambiguous, such as wakefulness and REM stages. The results in Figure 3 show that the addition of EMG and EOG produces a better and more stable model performance.

Although ECG is not recommended for manual sleep scoring in the scoring standards of AASM or R&K, it is undeniable that ECG is one of the most commonly used tools in clinical to monitor vital signs. In sleep scoring, the application of ECG channels facilitates distinguish of signal artifacts. Besides, according to our previous research[15], ECG signals perform well in distinguishing sleep and wakefulness. Given that, we train the model to recognize ECG signals so that it can contribute to the discrimination of sleep stages in different classification problems, for example, binary classification of sleep segments. In addition, by changing the number of units in the decision layer, the proposed model can be easily applied to different classification problems of sleep stages, such as distinguishing sleep state and awake state, the recognition of light sleep and deep sleep.

Few studies tested their model on PSG recordings collected from a variety of recording environments and hardware platforms. Zhang et al. [22] did so, where they trained a model on 461 recordings from the SOF dataset and then tested the trained model on the SHHS dataset, achieving a kappa value of 0.53. In the present article, direct testing of a trained model on different datasets yielded moderate accuracy, which is less satisfactory. A possible reason is the lack of sufficient training data since we cannot train the model on a huge dataset due to limited computation resource. In addition, model performance on independent datasets depends on the similarity between the training set and the test set, while the employed three datasets have disparate attributes, as shown in Table 1. Nevertheless, our model is promising and worthwhile training it on a huge and high-quality dataset in our future research, which helps to improve model generalization. Moreover, it would be interesting to explore model performance on large populations with diverse sleep problems, given the complex and diverse clinical symptoms of suspected patients.

5. Conclusion

The present paper proposed a deep learning model for automatic sleep scoring, which took raw PSG signals as input without any human-engineered features. The model employed two parallel convolution layers with different filter sizes and one LSTM layer to exploit information across temporal and spatial scales, thereby enhancing model performance. Moreover, the unique structure allowed the model to cope with various input channels and several signal modalities from different datasets without task-specific modifications to model architecture and hyper-parameters. Model generalization and model transferability were tested on participants with distinct characters, even subjects with complex sleep disturbances. Results evaluated on three public datasets showed that the model achieved a comparable or better performance compared to the state-of-the-art methods, and the highest classification accuracy was achieved by the fusion of multiple PSG signals. Future work will require huge and high-quality datasets to improve the robustness and generalization of the proposed model.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 81471742&91748105), the Fundamental Research Funds for the Central Universities [DUT2019] in Dalian University of Technology in China, and the scholarship from China Scholarship Council (Nos. 201606060227).

Declaration of interest

None.

References

[1] M. Dattilo *et al.*, "Sleep and muscle recovery: endocrinological and molecular basis for a new and promising hypothesis," *Med. Hypotheses*, vol. 77, no. 2, pp. 220–222, 2011.

[2] R. Stickgold and M. P. Walker, "Sleep-dependent memory consolidation and reconsolidation," *Sleep Med.*, vol. 8, no. 4, pp. 331–343, 2007.

[3] L. Xie *et al.*, "Sleep drives metabolite clearance from the adult brain," *Science (80-.).*, vol. 342, no. 6156, pp. 373–377, 2013.

[4] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Washingt. DC US Natl. Inst. Heal. Publ.*, 1968.

[5] R. B. Berry *et al.*, "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events," *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 597–619, 2012.

[6] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, 2016.

[7] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009.

[8] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Comput. Methods Programs Biomed.*, vol. 176, pp. 81–91, 2019.

[9] K. Šušmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," *Artif. Intell. Med.*, vol. 44, no. 3, pp. 261–277, 2008.

[10] A. Procházka, J. Kuchyňka, O. Vyšata, P. Cejnar, M. Vališ, and V. Mařík, "Multi-Class Sleep Stage Analysis and Adaptive
 Pattern Recognition," *Appl. Sci.*, vol. 8, no. 5, p. 697, 2018.

[11] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, "Sleep stage classification using single-channel EOG," *Comput. Biol. Med.*, vol. 102, no. August, pp. 211–220, 2018.

[12] S. I. Dimitriadis, C. Salis, and D. Linden, "A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates," *Clin. Neurophysiol.*, vol. 129, no. 4, pp. 815–828, 2018.

[13] S. Sheykhivand, T. Y. Rezaii, A. Farzamnia, and M. Vazifehkhahi, "Sleep Stage Scoring of Single-Channel EEG Signal based on RUSBoost Classifier," in *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2018, pp. 1–6.

[14] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, 2018.

[15] R. Yan *et al.*, "Multi-modality of polysomnography signals' fusion for automatic sleep scoring," *Biomed. Signal Process. Control*, vol. 49, pp. 14–23, 2019.

[16] S. Güneş, K. Polat, and Ş. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7922–7928, 2010.
[17] S. Özşen, "Classification of sleep stages using class-dependent sequential feature selection and artificial neural network," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1239–1250, 2013.
[18] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.
[19] Z. Mousavi, T. Yousefi Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi, "Deep convolutional neural network for classification of sleep stages from single-channel EEG signals," *J. Neurosci. Methods*, vol. 324, p. 108312, 2019.
[20] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, p. zsy041, 2018.
[21] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks," *arXiv Prepr. arXiv1610.01683*, 2016.
[22] L. Zhang, D. Fabbri, R. Upender, and D. Kent, "Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks," *Sleep*, vol. 42, no. 11, p. zsz159, 2019.
[23] J. Zhang, R. Yao, W. Ge, and J. Gao, "Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 183, p. 105089, 2020.
[24] X. Zhang, W. Kou, E. I.-C. Chang, H. Gao, Y. Fan, and Y. Xu, "Sleep Stage Classification Based on Multi-level Feature Learning and Recurrent Neural Networks via Wearable Device," *Comput. Biol. Med.*, vol. 103, pp. 71–81, 2018.
[25] S. Wang, J. Cao, and P. S. Yu, "Deep Learning for Spatio-Temporal Data Mining : A Survey," pp. 1–21.
[26] Y. Liu, R. Fan, and Y. Liu, "Deep Identity Confusion for Automatic Sleep Staging Based on Single-Channel EEG," *Proc. - 14th Int. Conf. Mob. Ad-Hoc Sens. Networks, MSN 2018*, pp. 134–139, 2018.
[27] X. Chen, J. He, X. Wu, W. Yan, and W. Wei, "Sleep staging by bidirectional long short-term memory convolution neural network," *Futur. Gener. Comput. Syst.*, vol. 109, pp. 188–196, 2020.
[28] P. Fonseca *et al.*, "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population," *Sleep*, no. April, pp. 1–10, 2020.
[29] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2019.
[30] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, 2019.
[31] A. Malafeev *et al.*, "Automatic Human Sleep Stage Scoring Using Deep Neural Networks," *Front. Neurosci.*, vol. 12, p. 781, 2018.
[32] H. Phan *et al.*, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," *arXiv Prepr. arXiv1907.13177*, 2019.
[33] D. A. Dean *et al.*, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016.
[34] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016.
[35] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[36] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000.
[37] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiol. Meas.*, vol. 36, no. 10, pp. 2027–2040, 2015.
[38] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018.
[39] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, 2017.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 7132–7141, 2018.
[41] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," *30th Int. Conf. Mach. Learn. ICML 2013*, no. PART 1, pp. 115–123, 2013.
[42] F. Chollet, "Keras: Deep learning library for theano and tensorflow," *URL https//keras. io/k*, vol. 7, no. 8, p. T1, 2015.
[43] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*, pp. 265–283, 2016.
[44] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," *Artif. Intell. Med.*, vol. 53, no. 1, pp. 25–33, 2011.
[45] S. Back, S. Lee, H. Seo, D. Park, T. Kim, and K. Lee, "Intra- and Inter-epoch Temporal Context Network (IITNet) for Automatic Sleep Stage Scoring," *arXiv Prepr. arXiv1902.06562*, 2019.
[46] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 7046–7059, 2013.
[47] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, 2018.
[48] I. Fernández-Varela, E. Hernández-Pereira, and V. Moret-Bonillo, "A Convolutional Network for the Classification of Sleep Stages," *Multidiscip. Digit. Publ. Inst. Proc.*, vol. 2, no. 18, p. 1174, 2018.
[49] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE Laurens," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.