JYU DISSERTATIONS 610

Dongdong Zhou

Automatic Sleep Stage Classification Based on Single-Channel EEG

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY

JYU DISSERTATIONS 610

Dongdong Zhou

Automatic Sleep Stage Classification Based on Single-Channel EEG

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 2
maaliskuun 17. päivänä 2023 kello 9.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 2, on March 17, 2023 at 9 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors
Marja-Leena Rantalainen
Faculty of Information Technology, University of Jyväskylä
Ville Korkiakangas
Open Science Centre, University of Jyväskylä

ISBN 978-951-39-9303-0 (PDF)
URN:ISBN:978-951-39-9303-0
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9303-0

ABSTRACT

Zhou, Dongdong
Automatic sleep stage classification based on single-channel EEG
Jyväskylä: University of Jyväskylä, 2023, 66 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 610)
ISBN 978-951-39-9303-0 (PDF)

Sleep issues are on the rise and have a negative impact on global population health, particularly during the COVID-19 outbreak. The most crucial stage is to correctly assess sleep quality and diagnose sleep disorders by categorizing the stages of sleep (also called sleep scoring). The most common tool for sleep scoring is the polysomnography (PSG) recording. However, this manual procedure is time-consuming and heavily reliant on clinic expertise. As a result, it is essential to develop automatic sleep stage classification methods to fulfill the growing unmet demands for sleep research. In this thesis, we focus on developing deep learning-based (DL-b) methods and solutions for the class imbalance problem (CIP) and model interpretability in automatic sleep scoring using single-channel EEG.

In *Article I*, we present an efficient one-dimensional Conventional Neural Network (1D-CNN) based model, namely SingleChannelNet (SCNet), for automatic sleep scoring with raw single-channel EEG. In *Article II*, we further seek to accelerate the training speed with the spectrogram input. In addition, our proposed LightSleepNet (LSNet) could achieve promising performance while requiring far fewer model parameters. To alleviate the CIP, we propose different balancing methods to balance the dataset samples and network connection with the Gaussian white noise addition (GWN), Generative adversarial network (GAN) and class weight redesign methods in *Articles III and IV*, respectively. In *Article V*, we provide an interpretable sleep stage classification scheme based on layer-wise relevance propagation (LRP), which can visually demonstrate the contribution of specific EEG patterns in each sleep stage to the final model prediction.

To conclude, this thesis proposes two DL-b methods for automatic sleep stage classification, which could obtain remarkable performance on public PSG datasets. In addition, we systematically analyze and present efficient solutions to the CIP and model interpretability in automatic sleep scoring. Ultimately, we expect this thesis to promote the practical application of DL-b automatic sleep scorning methods in the future.

Keywords: Sleep stage classification, single-channel EEG, deep neural network, class imbalance problem, model interpretability

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Zhou, Dongdong
Automaattinen univaiheen luokittelu yksikanavaisen EEG:n perusteella
Jyväskylä: University of Jyväskylä, 2023, 66 s. (+artikkelit)
(JYU Dissertations
ISSN 2489-9003; 610)
ISBN 978-951-39-9303-0 (PDF)

Uniongelmat lisääntyvät ja niillä on kielteinen vaikutus maailman väestön terveyteen, kuten COVID-19-pandemia osoitti. Uniongelmien analysoimisessa tärkein vaihe on arvioida oikein unen laatua ja diagnosoida unihäiriöt luokittelemalla unen vaiheet (kutsutaan myös unipisteytykseksi). Yleisin unen pisteytyksen työkalu on polysomnografiatallennus. Tämä toimenpide on kuitenkin aikaa vievä ja on tehtävä asiantuntevalla klinikalla. Tästä syystä tarvitaan automaattisia univaiheen luokittelumenetelmiä, jotka täyttävät unitutkimuksen kasvavat vaatimukset. Tässä väitöskirjassa keskitymme kehittämään syväoppimiseen perustuvia menetelmiä ja etsimään ratkaisuja luokkaepätasapaino-ongelmaan ja mallin tulkittavuuteen automaattisessa unen pisteytyksessä käyttäen yksikanavaista EEG:tä.

Artikkelissa I esittelemme tehokkaan yksiulotteisen konvoluutiohermoverkkopohjaisen mallin SingleChannelNet (SCNet). Se perustuu automaattiseen unen pisteytykseen yksikanavaisella EEG:llä. Artikkelissa II pyrimme parantamaan mallin optimointinopeutta spektrogrammin syötteen avulla. Ehdottamallamme mallilla LightSleepNet (LSNet) on lupaava suorituskyky ja se vaatii merkittävästi vähemmän malliparametreja. Luokkaepätasapaino-ongelman lieventämiseksi ehdotamme artikkeleissa III ja IV erilaisia menetelmiä tietojoukkonäytteiden tasapainottamiseksi Gaussin valkoisen kohinan lisäyksen, generatiivisen adversariaalisen verkon ja verkkoyhteyden avulla käyttäen luokkapainotuksen uudelleensuunnittelumenetelmiä. Artikkelissa V tarjoamme tulkittavan relevanssin kerroksittaiseen etenemiseen perustuvan univaiheen luokittelukaavion, joka voi visuaalisesti osoittaa kunkin univaiheen tiettyjen EEG-kuvioiden vaikutuksen lopulliseen mallin ennusteeseen.

Johtopäätöksenä tässä opinnäytetyössä ehdotetaan automaattiseen univaiheen luokitteluun kahta menetelmää, jotka voisivat saavuttaa huomattavan suorituskyvyn julkisissa polysomnografia-aineistoissa. Lisäksi analysoimme ja esittelemme systemaattisesti tehokkaita ratkaisuja luokkaepätasapaino-ongelmaan ja mallin tulkittavuuteen automaattisessa unipisteytyksessä. Odotamme tämän opinnäytetyön edistävän automaattisten unihalvausmenetelmien käytännön soveltamista tulevaisuudessa.

Avainsanat: Univaiheen luokittelu, yksikanavainen EEG, syvä hermoverkko, luokkaepätasapaino-ongelma, mallin tulkittavuus

Author

Dongdong Zhou
Faculty of Information Technology
University of Jyväskylä
Finland

Supervisors

Lauri Kettunen
Faculty of Information Technology
University of Jyväskylä
Finland

Fengyu Cong
Faculty of Information Technology
University of Jyväskylä
Finland

Zheng Chang
Faculty of Information Technology
University of Jyväskylä
Finland

Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers

Wei Chen
School of Information Science and Technology
Fudan University
China

Ari Visa
Faculty of Information Technology and
Communication Sciences
Tampere University
Finland

Opponent

Chengyu Liu
School of Instrument Science and Engineering
Southeast University
China

ACKNOWLEDGEMENTS

It seems that I need to say goodbye to my four-year doctoral study in the University of Jyväskylä. I would like to give all my best wishes and thanks to all persons who have provided me with great help and support in my daily life and doctoral research.

First of all, I would like to express my sincere thanks to my supervisor Prof. Tapani Ristaniemi. I will never forget the scene of the first remote interview in 2017. It started the mentor relationship between us. His bright smile always inspired me to face the difficulties in my research and daily life. Meanwhile, he always provided constructive suggestions for my doctoral research and generous assistance in adapting to the local life in Jyväskylä. This dissertation is to memorize Prof. Tapani Ristaniemi for his great help.

I would like to express my best thanks to my supervisor Prof. Fengyu Cong. He was a course teacher in my bachelor and master studies, and I was impressed by his professional expertise and humorous teaching approach. Prof. Cong guided my doctoral study and provided me with colossal assistance with my Ph.D. application. No matter what difficulties I encountered in study or life, he always provided timely and selfless help. In addition, his enthusiasm for scientific research motivates me to continue on the path of scientific research.

I would like to express my sincere gratitude to my supervisor Prof. Lauri Kettunen. Lauri is more than a supervisor, but a kindly friend. He never puts any pressure on me and constantly encourages me to work patiently on research problems. Lauri always shared his academic and daily life experiences to motivate and inspire me. Special thanks to Lauri for his valuable guidance on dissertation writing. In addition, I would like to thank Lauri for his generous support and help with the three months short-term grant application at the University of Jyväskylä for doctoral defense.

I would like to express my sincere gratitude to Prof. Zheng Chang. He helped me a lot not, only in my life and study. He is more like a warm-hearted older brother, he would give his suggestions and assistance timely whenever you asked for help from him. His humorous attitude to life also encouraged me to keep going optimistically.

I would like to express my profound appreciation to Dr. Qi Xu from Dalian University of Technology. We met each other in August 2020. He gave me a lot of help and instruction in my doctoral study. We often discussed my doctoral research together and he then gave me valuable feedback. I will never forget his encouragement after my work was rejected for the first time after a major revision. He also greatly assisted me with the paper drafting, submission and revision.

I am very grateful to Dr. Hongming Xu from the Dalian University of Technology, Dr. Guoqiang Hu from the Dalian Maritime University, Dr. Jiacheng Zhang from the Nanjing University of Information Science and Technology. Thanks for their valuable suggestions and support for my doctoral research.

I would like to express my gratitude to the China Scholarship Council for supporting me financially while I pursued my Ph.D. at the University of Jyväskyä (No. 201806060164). I am also very appreciative to the University of Jyväskyä for offering such an important training facility.

I would like to say special thanks to my reviewers: Prof. Wei Chen from Fudan University in China and Prof. Ari Visa from Tampere University in Finland and the opponent Prof. Chengyu Liu from Southeast University in China. Thanks to them for their time and efforts, and valuable comments.

I would also like to thank Deqing Wang, Yongjie Zhu, Jia Liu, Rui Yan, Xiulin Wang, Wenya Liu, Ville Isomöttönen, Nina Pekkala, Reza Mahini, Lili Tian, Guanghui Zhang, Huashuai Xu, Xiaoshuang Wang, Zhonghua Chen, Lina Sun, Liting Song, Jiaqi Zheng, Dong Tang, Xin Zuo, Dongying Zheng, Jiaqi Zhu and other lovely friends. I will never forget the fantastic times I had with you and your kind help. I would like to appreciate the time and effort that Marja-Leena Rantalainen has spent reviewing and editing my thesis.

Last but not least, I would like to give my sincerest thanks to my grandfather Huozhu Ai, grandmother Julan Zhou, father Mingquan Zhou, mother Haie Yan, and younger brother Weiwei Zhou and other family members for their unconditional love and support. I can imagine their surprised expressions that their names appear in an English doctoral dissertation. I would express special thanks to my girlfriend, Jian Wang, who always respects my every decision. Moreover, she gave me the best support for my doctoral research. We had an unforgettable and wonderful experience in Jyväskylä. Without your support and love, I could not complete my doctoral research without pressures. To myself, keep going.

Jyväskylä, Finland
February 28, 2023
Dongdong Zhou

LIST OF ACRONYMS

AASM	American Academy of Sleep Medicine
CIF	Class imbalance factor
CIP	Class imbalance problem
CNN	Conventional neural network
DA	Data augmentation
DL-b	Deep learning-based
DNN	Deep neural network
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyogram
EOG	Electrooculogram
GAN	Generative adversarial network
GWN	Gaussian white noise
LRP	Layer-wise relevance propagation
LSTM	Long short-term memory
ML-b	Machine learning-based
N1	Non-rapid eye movement stage 1
N2	Non-rapid eye movement stage 2
N3	Non-rapid eye movement stage 3
N4	Non-rapid eye movement stage 4
PSG	Polysomnography
R&K	Rechtschaffen and Kales
REM	Non-rapid eye movement
RNN	Recurrent neural network
SNR	Signal-to-noise ratio
STFT	Short time Fourier transform
W	Wake

LIST OF FIGURES

CONTENTS

LIST OF INCLUDED ARTICLES

PI **Dongdong Zhou**, Jian Wang, Guoqiang Hu, Jiacheng Zhang, Fan Li, Rui Yan, Lauri Kettunen, Zheng Chang, Qi Xu, and Fengyu Cong. SingleChannelNet: A model for automatic sleep stage classification with raw single-channel eeg. *Biomedical Signal Processing and Control, 75, 103592*, https://doi.org/10.1016/j.bspc.2022.103592, 2022.

PII **Dongdong Zhou**, Qi Xu, Jian Wang, Jiacheng Zhang, Guoqiang Hu, Lauri Kettunen, Zheng Chang, and Fengyu Cong. LightSleepNet: A Lightweight Deep Model for Rapid Sleep Stage Classification with Spectrograms. *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC2021), pp. 43-46*, https://doi.org/10.1109/EMBC46164.2021.9629878, 2021.

PIII Qi Xu, **Dongdong Zhou**, Jian Wang, Jiangrong Shen, Lauri Kettunen, and Fengyu Cong. Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance. *2022 International Joint Conference on Neural Networks (IJCNN2022), pp. 1-6*, https://doi.org/10.1109/IJCNN55064.2022.9892741, 2022.

PIV **Dongdong Zhou**, Qi Xu, Jian Wang, Hongming Xu, Lauri Kettunen, Zheng Chang, and Fengyu Cong. Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification. *IEEE Transactions on Instrumentation and Measurement, 75, 1-12*, https://doi.org/10.1109/TIM.2022.3191710, 2022.

PV **Dongdong Zhou**, Qi Xu, Jiacheng Zhang, Lei Wu, Lauri Kettunen, Zheng Chang, Hongming Xu, and Fengyu Cong. Interpretable Sleep Stage Classification Based on Layer-wise Relevance Propagation. *Submitted to IEEE Transactions on Cognitive and Developmental Systems, under revision*, 2023.

1 INTRODUCTION

This chapter begins by introducing the background and motivation for this thesis. After that, it gives a summary of the entire study before finally describing the thesis structure.

1.1 Research background

Sleep is a spontaneously recurring physiologic state that occupies around one-third of human life (Mesarwi et al., 2013; Kay and Dzierzewski, 2015; Oikonomou and Prober, 2017). Sufficient and high-quality sleep contributes to better brain function (Luyster et al., 2012). Additionally, there is also compelling proof that sleep aids in the development of long-term memory (Sawangjit et al., 2022). Unfortunately, there are numerous sleep-related diseases that affect millions of individuals worldwide. For instance, approximately 10-30% of the population worldwide exhibit insomnia symptoms (Wafford and Ebert, 2008; Bhaskar et al., 2016). Moreover, the prevalence of insomnia shows a noticeably incremental trend with the continued outbreaks of the COVID-19 pandemic (Morin et al., 2022). Additionally, abnormal sleep patterns can serve as a warning sign for several neurodegenerative illnesses (e.g., Parkinson's and Alzheimer's diseases) (Iranzo et al., 2006). Accurate sleep stage classification, also known as sleep scoring, is the first phase used in clinics and is crucial for determining sleep disorders and gauging the quality of one's sleep. (Sousa et al., 2015; Zhang and Wu, 2017, 2021).

Polysomnography (PSG) recordings are considered the golden tool for evaluating sleep quality and disorders (de Souza et al., 2003; Cook et al., 2017; Derbin et al., 2022). We illustrate an example of the PSG recording in Figure 1. As can be seen, PSG data typically include multi-channel signals, such as electroencephalogram (EEG) for recording the brain activities, electromyogram (EMG) for collecting muscle activation, electrocardiogram (ECG) for acquiring the heart electrical activities, electrooculogram (EOG) for measuring eye movements as well as other bio-signals for monitoring additional physiological information. The acquisition

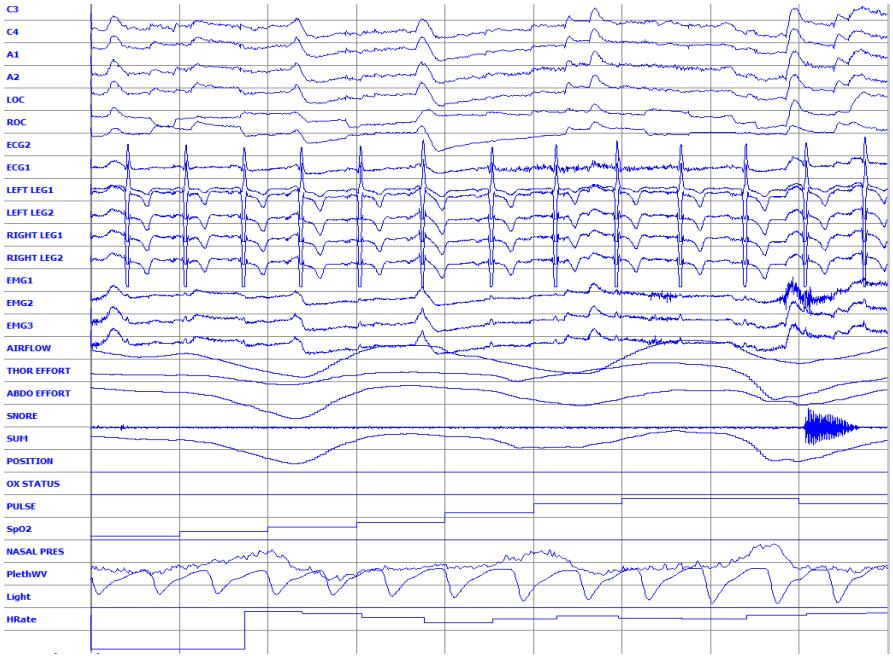


FIGURE 1 An illustration of PSG recordings.

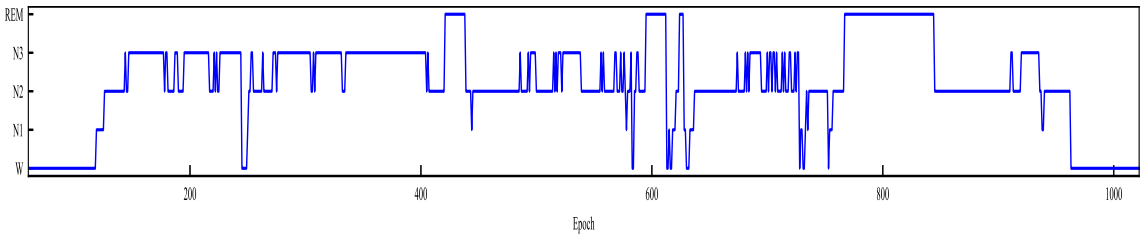


FIGURE 2 A healthy adult's hypnogram labeled by sleep experts.

of overnight PSG recordings is generally carried out in a specialized hospital or sleep laboratory. Typical overnight sleep structure can be characterized by five or six sleep stages, depending on different standards. The gathered PSG recordings are first segmented into different 30-second epochs, and with the Rechtschaffen and Kales (R&K) manual (Rechtschaffen, 1969), each 30-second epoch is labeled with Wake (W), Non-rapid eye movement (Non-REM: N1, N2, N3 and N4), and Rapid eye movement (REM). To abolish the stage 'movement time', the American Academy of Sleep Medicine (AASM) rule (Berry et al., 2012) merges stages N3 and N4 into stage N3. In this thesis, we adopt the five sleep stages scheme (i.e., W, N1, N2, N3, and REM). Sleep specialists will finally review the sleep scoring results to identify any potential patient sleep issues. In Figure 2, we demonstrate a typical whole-night sleep architecture of a healthy adult. The handmade labeling process takes a lot of time and is prone to human mistake (Phan et al., 2018b). Eight hours of whole-night PSG recording would take an experienced sleep specialist roughly two to four hours to annotate (Hassan and Bhuiyan, 2016) and the

overall agreement is around 80% (AASM: 82.0% and R&K: 80.6%) (Danker-hopfe et al., 2009). Consequently, a reliable and automatic tool is urgently needed to help lighten the workload of sleep physicians.

1.2 Research motivation

So far many methods have been proposed to accomplish automated sleep scroing assignments. In terms of the applied computational algorithms, we can categorize these methods into two groups: approaches based on conventional machine learning (ML-b) and deep learning (DL-b). The ML-b methods primarily concentrate on developing feature extraction approaches from the PSG recordings to identify the differentiable patterns of distinct sleep stages (Zhu et al., 2014; Karimzadeh et al., 2017; Satapathy et al., 2021; Fatimah et al., 2022). These pre-extracted temporal (Redmond and Heneghan, 2006; Šušmáková and Krakovská, 2008; Koley and Dey, 2012) or frequential (Sharma et al., 2018; Kouchaki et al., 2014; Tsinalis et al., 2016) features are subsequently loaded into traditional machine learning models to forecast sleep stages. However, each person's sleep structure varies physiologically depending on age, health, gender and other inherent factors. Model performance may vary depending on the types and quantity of features, and there is no standard rule about feature extraction and selection. Therefore, the capacity of manual preprocessing and feature extractions will always be constrained.

Deep neural networks (DNNs) have recently been successfully employed in a number of domains, including: computer vision (CV) (Simonyan and Zisserman, 2014; Krizhevsky et al., 2017; Esser et al., 2021), natural language processing (NLP) (Sutskever et al., 2014; Yang et al., 2016; Zhang et al., 2022b), robotics (Bai et al., 2018; Paolanti et al., 2019; Ding et al., 2022), biomedical and health informatics (Xu et al., 2018; Yan et al., 2021b; Zhou et al., 2022a), Etc. The DL-b method could effectively avoid the necessity for manual feature extraction with its excellent automatic feature learning ability. With regards to the signal modality, previous DL-b sleep scoring studies could be divided into multi-channel (Chambon et al., 2018; Zhang et al., 2019; Yan et al., 2021a) versus single-channel schemes (Supratak et al., 2017; Zhou et al., 2021; Goshtasbi et al., 2022) depending on the inputting channels. Although the multi-channel strategy could provide more informational references for sleep scoring, higher computational and acquisition costs can be expected. Given the home sleep monitoring environment, we can imagine the inconvenience and discomfort caused by multi-channel electrodes to the consumers. In contrast, the single-channel EEG scheme is considered a good alternative for practical real-time applications (e.g., portable sleep monitoring devices). Considering the enormous size (ups to millions of model parameters) of applied DNNs, the implementation of DL-b methods is also hampered by high costs in several areas,processing speed, storage capacity, network throughput, energy use, and hardware complexity (Zhou et al., 2021). In addition, most existing

PSG datasets experience an intrinsic class imbalance problem (CIP), where the amount of each sleep stage is wildly unbalanced due to the particular sleep structure. The DL-b models are skewed towards the majority sleep stages (i.e., with a high percentage) and the minority class suffers discrimination in recognition accuracy. Last but not least, Sleep experts were skeptical of DL-b models because of their unconvincing interpretability and black-box nature for the automatic sleep stage classification in real-world settings. Establishing trust among practitioners and properly articulating how the deep model makes decisions is a crucial and necessary step.

Here, we conclude four main crucial issues in the realm of automatical sleep scoring:

1. How to construct a trustworthy single-channel EEG model for automated sleep scoring without utilizing hand-crafted features?
2. How to expedite the training speed and build a compact yet efficient DL-b model for automated sleep scoring?
3. How to balance the dataset samples and the relationship between the applied DL-b model and the imbalanced dataset? How to improve the minority class identification accuracy without sacrificing the overall accuracy?
4. How to enhance the model explanation of the automatic sleep scoring model? Can the specific EEG patterns of each sleep stage be detected by the DL-b model for making the final decision? Is it consistent with the sleep scoring manuals?

This thesis seeks to investigate practical approaches to the four challenges mentioned above.

1.3 Introductory overview

This thesis explores the single-channel EEG-based automatic sleep scoring methods. The thesis addressed the four questions mentioned above in the automatic sleep scoring task. In each included article, the detailed object and solution are as follows:

Article I In order to use as few channel signals as possible, we introduce a one-demential convolutional neural network (1D-CNN) based model, named SingleChannelNet (SCNet) for automatic sleep scoring using raw single-channel EEG. This paper employs the many-to-one scheme, which imitates the manual sleep stage classification procedure conducted by sleep physicians. We then establish a multi-convolution (MC) block, which is composed of distinct sizes of filters, to learn different scale feature representations from long-time series input. Similarly, we combine the max-pooling and average-pooling layers as the max-average (M-Apooling) layer to replace the max-pooling for further enhancing the

feature extraction capacity of the proposed SCNet. We validate our model's efficiency and the many-to-one scheme's superiority on three public PSG datasets.

Article II To accelerate the training speed, we first adopt the short-time Fourier transform (STFT) to obtain the spectrograms from raw EEG signals, which are also considered higher-level feature representations of the source signals. We then present a lightweight two-dimensional CNN (2-D CNN) framework, LightSleepNet (LSNet), with far fewer model parameters to accommodate the spectrograms input. The training speed of our LSNet with the spectrograms input is verified by comparing it to the baseline model with time series input. Besides, we compare the model performance and the number of model parameters with other state-of-the-art methods using the same datasets.

Articles III and IV These two papers in this thesis seek the answers for the CIP in automatic sleep scoring tasks. In *Article III*, we propose two balancing methods, signal-driven and image-driven, to balance the dataset samples. To be specific, we expand the number of sleep N1 samples in the training set, which is regarded as the representative of the minority class. The signal-driven method adds the Gaussian white noise to the raw EEG signal of N1. The noisy EEG signal is then converted to the time-frequency image employing short time Fourier transform (STFT). While for the image-driven approach, we first transform the raw EEG signal of N1 to the time-frequency image and then add similar intensities of Gaussian white noise to the time-frequency image. In *Article IV*, we first introduce the class imbalance factor to statistically assess the severity of CIP. Then two novel balancing methods are presented. The first one is to generate new epochs of N1 from raw EGG signals through a proposed generative adversarial network (GAN) model and add different intensities of Gaussian white noise (i.e., 10, 5, 2, and 1 dB). Then, the efficiency of different times noise addition is investigated. Another approach is maintaining the current data distribution while balancing the network connection between the applied DL-b model and the unbalanced dataset. The class distribution and brain-inspired principle are used to allocate the class weight of each sleep stage.

Article V This paper proposes a new interpretable sleep stage classification system based on layer-wise relevance propagation (LRP). The first step is acquiring the time-frequency images carrying the EEG patterns information through the STFT method, allowing us to present the EEG patterns of different sleep stages visually. Next, the time-frequency images are utilized as the trained model input, and the DL-b model estimates the corresponding input. Finally, the LRP determines which crucial pixels (corresponding to frequency features) in the time-frequency image input are most relevant for the final layer. We seek to validate whether the proposed model can correctly identify specific EEG patterns in each stage when making the final decision.

1.4 Thesis structure

The following are brief descriptions of the rest of this thesis. Chapter 2 first describes four public PSG datasets employed in this thesis and then analyzes related research and theoretical basis. Chapter 3 summarizes the objective, method, result, discussion, and author contributions of five articles included in this thesis. Chapter 4 concludes the whole research work and discusses our work's limitations and potential future directions.

2 MATERIALS AND RELATED WORK

This chapter first introduces four open-access PSG datasets used in this dissertation. The related work and theoretical analysis are then provided. Given this, the novelties and contributions of this dissertation are finally discussed.

2.1 Experimental PSG datasets

There are four public PSG datasets employed in this thesis:

- Sleep Heart Health Study (SHHS) (Quan et al., 1997; Zhang et al., 2018),
- Cleveland Children's Sleep and Health Study (CCSHS) (Rosen et al., 2003; Zhang et al., 2018),
- Sleep-EDF-2013 (Sleep-EDF-V1), (Kemp et al., 2000) and
- Sleep-EDF-2018 (Sleep-EDF) (Kemp et al., 2000).

Thereinto, we adopt the CCSHS, Sleep-EDF-V1 and Sleep-EDF datasets in *Articles I, III, IV, and V* and the SHHS, Sleep-EDF-V1 and Sleep-EDF datasets in *Article II*. The detailed description is shown as follows:

SHHS The SHHS is a multi-center cohort research conducted by the National Heart Lung & Blood Institute to identify the effects of irregular breathing during sleep on the cardiovascular system and other factors. This study examines whether breathing during sleep is linked to an increased risk of high blood pressure, coronary heart disease, stroke, and all-cause death. There are two subsets in the SHHS dataset: initial SHHS (SHHS1) and second SHHS (SHHS2). We employ 100 subjects (over 40 years old) from the subset SHHS1 in *Article II* as the sampling rate (125 Hz) is the same for all recordings. There are two EEG channels (C4/A1 and C3/A2), two EGO channels, one ECG and EMG channel, two inductance plethysmography channels, one position sensor, one light sensor, one pulse oximeter, and an airflow sensor included in the SHHS1 dataset. Our study

adopts the single-channel EEG, C4/A1, in accordance with the AASM manual's guidance. (Data link: https://sleepdata.org/datasets/shhs).

CCSHS The CCSHS, which includes 515 kids aged 16 to 19 years, is one of the biggest population-based pediatric cohorts examined utilizing objective sleep studies. The CCSHS dataset mainly comprises two EEG channels (C3/A2 and C4/A1, sampled at 128 Hz), two EOC channels (sampled at 128 Hz), two ECG, and three EMG channels with a sampling rate of 256 Hz. Similarly, the single-channel EEG C4/A1 is employed in our study. (Data link: https://sleepdata.org/datasets/ccshs).

Sleep-EDF-V1 and Sleep-EDF There are two versions of the Sleep-EDF dataset, including two subsets (sleep-cassette (SC) and sleep-telemetry (ST)). We select the subset SC following the studies by Supratak et al. (2017) and Phan et al. (2018b). The first version was launched in 2013 (Sleep-EDF-V1), which comprises 39 overnight PSG records from 20 participants in the SC cohort, whose ages range from 25 to 34. The Sleep-EDF is an expanded edition released in 2018. The number of participants in the SC subgroup has grown to 78 (aged 25-101 years old), including 153 whole-night sleep recordings. The Fpz-Cz EEG sampled at 100 Hz is utilized in our study. (Data link: https://www.physionet.org/content/sleep-edfx/1.0.0/).

To validate the model generation of the proposed models in this research, we employ four PSG datasets with various age distributions and data properties.

2.2 Machine learning-based (ML-b) sleep stage classification

2.2.1 Overall procedure of the ML-b methods

The traditional ML-b methods generally contain four key steps:

1. Data preprocessing,
2. Feature extraction and optional selection,
3. Model selection, and
4. Classification.

An overview of the methods is shown in Figure 3. The PSG recordings frequently involve artifacts from several sources, notably EEG signals. Electrode shedding, perspiration, electrical interference (50 or 60 Hz), and additional electrophysiological signals, including eye activity, muscle movement, and heart impulses, are common sources of artifacts (Radüntz et al., 2015).

In Step 1, preprocessing is crucial to obtain cleaner sleep signals for subsequent analysis. For instance, the Butterworth notch filter is usually used to alleviate the influence of power line interference (Allen, 2009; Ţarălungă et al.,

FIGURE 3 The general schematic overview of the conventional machine learning-based sleep stage classification methods.

2014). Additionaly, the independent component analysis (ICA) technique could be applied to eliminate blink artifacts and the eye movement from EEG recordings (Jung et al., 1997; Mennes et al., 2010; Cong et al., 2015). In order to reduce individual variation and hasten algorithm convergence, signal normalization and feature standardization should also be taken into account.

Step 2 involves feature extraction and selection. Typically, these previously extracted traits can be categorized into three categories: time, frequency, and time-frequency domain features. The widely used time domain characteristics are mean, median, standard deviation, skewness, kurtosis, and percentile to highlight variations in the amplitude distribution and morphological characteristics of signals (Yan et al., 2019). Commonly-used techniques for frequency domain features extraction are fast Fourier transform (FFT) (Duhamel and Vetterli, 1990; Patanaik et al., 2018), discrete wavelet transform (DWT) (Shensa et al., 1992; Alickovic and Subasi, 2018), Etc. The different EEG patterns in sleep stages, such as Delta (<4 Hz), Theta (4-7.99 Hz), Alpha (8-13 Hz), and Beta (>13 Hz), could be visually demonstrated with frequency domain features. Regarding the time-frequency domain features, they can simultaneously offer information in both time and frequency domains, which are appropriate for nonstationary electrophysiological signal analysis. The most popular time-frequency analysis methods are the wavelet transform, the short-time Fourier transform (STFT), and the Hilbert-Huang transform (HHT) (Boostani et al., 2017). The feature selection is optional, which is mainly for determining the importance of different features to the classification performance.

Step 3 is the selection of classifiers, which can be categorized into unsupervised and supervised ones. Unsupervised classifiers (e.g., k-means clustering (Güneş et al., 2010; Shuyuan et al., 2015)) develop classification algorithms using unlabeled sleep recordings, which could significantly reduce the time and expense required to prepare labels for enormous volumes of sleep data. However, the classification performance is generally limited by the unsupervised training algorithm. By contrast, automated sleep scoring tends to use supervised classifiers more typically. The widely employed supervised classifiers are support vector machine (SVM) (Zhu et al., 2014; Lajnef et al., 2015; Sharma et al., 2018), random forest (RF) (Fraiwan et al., 2012; Xiao et al., 2013; Memar and Faradji, 2017), K-nearest neighbor (KNN) (Phan et al., 2013; Kayikcioglu et al., 2015; Boostani et al., 2017), Etc. These hand-crafted features acquired in the third step are acted as the input of classical classifiers to classify different sleep stages.

In Step 4, different sleep stages can be classified as

- four sleep stages: W, light sleep (N1, N2), deep sleep (N3, N4), and REM, or
- five sleep stages: W, N1, N2, N3, and REM, or
- six sleep stages: W, N1, N2, N3, N4, and REM.

In this thesis, we employ the five sleep stages scheme.

2.2.2 Performance metrics

Precision (*PR*), recall (*RE*), F1 score (*F*1), Overall accuracy (*ACC*), and Cohen's kappa coefficient (*K*) are commonly used performance evaluation metrics, they are defined as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}, \tag{1}$$

$$PR = \frac{TP}{TP + FP}, \tag{2}$$

$$RE = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = 2 \cdot \frac{RE \cdot PR}{RE + PR}, \tag{4}$$

$$K = \frac{\frac{\sum_{i=1}^{c} x_{ii}}{N} - \frac{\sum_{i=1}^{c}\left(\sum_{j=1}^{c} x_{ij}, \sum_{j=1}^{c} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{c}\left(\sum_{j=1}^{c} x_{ij} \sum_{j=1}^{c} x_{ji}\right)}{N^2}}, \tag{5}$$

where *TP*, *TN*, *FN*, and *FP* represent true positives, true negatives, false negatives, and false positives, respectively. N is the total numbers, c refers to the number of classes ($c = 5$ in here). The confusion matrix's diagonal value is denoted by x_{ii} ($1 \leq i \leq 5$). The model's *ACC* is expressed as a percentage of all correctly predicted outcomes. The proportion of successfully predicted positives to all positives is defined by *PR*. The ratio of true positives to all predictions in the actual class is known as *RE*. The weighted average of *RE* and *PR* is represented by *F*1. *K* represents the degree of agreement between actual and predicted labels.

2.3 Deep learning-based (DL-b) sleep stage classification

Although ML-b methods for automatic sleep scoring have considerably developed, the complicated procedure is the prevalent defect. In addition, the model generalization of ML-b methods is restricted due to the different properties of PSG datasets. The DL-b methods could efficiently avoid manual features and independently learn feature presentation from the model input. Regarding the channel number of the model input, these DL-b approaches could be categorized into multi-channel (Andreotti et al., 2018; Chambon et al., 2018; SM et al., 2019; Zhang et al., 2022a; Efe and Ozsen, 2023) and single-channel schemes (Malafeev

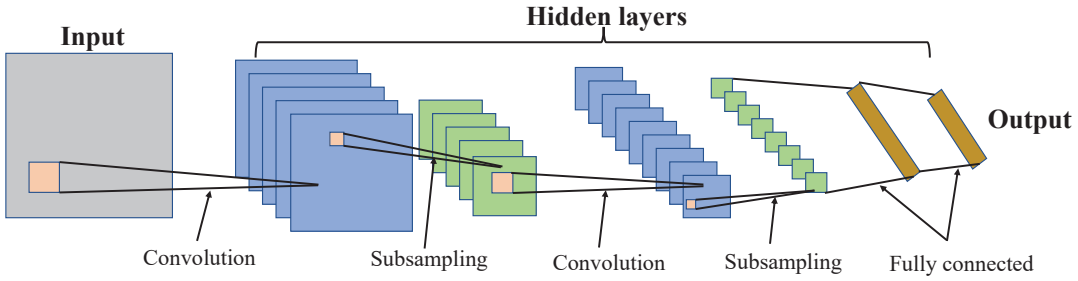


FIGURE 4 An illustration of a CNN model including three key components: convolutional layers, pooling layers, and fully connected layers.

et al., 2018; Humayun et al., 2019; Xu et al., 2022; Fu et al., 2021; Goshtasbi et al., 2022).

Chambon et al. (2018) proposed a two-dimensional CNN (2D-CNN) for automatic sleep scoring with multiple PSG signals, which achieved an accuracy of 80% on the MASS dataset (O'reilly et al., 2014). Zhang et al. (2019) provided a DL-b method combining the CNN and Long Short-Term Memory (LSTM) using multi-modal PSG data, the accuracy was 87% on the SHHS dataset. A prospective CNN-based method was introduced by Sors et al. (2018) using raw single-channel EEG. Although the architecture attained a high accuracy of 87%, the model has 12 convolutional layers, which increases its complexity. Besides, a multiple-input-one-output (also called many-to-one) scheme was investigated by Chambon et al. (2018), which imitated how sleep technologists do manual sleep scoring.

In addition to classification performance, model complexity is another aspect to consider. The model parameters quantities in Sors et al. (2018), Zhang et al. (2019), Supratak and Guo (2020); Supratak et al. (2017) and Mousavi et al. (2019) are around 2.2 million (**m**), 1.3 **m**, 1.3 **m**, 21 **m**, and 2.6 **m**, respectively. A simple yet efficient model using single-channel EEG is an optimal choice for a real-world application. *Article I* aims to build a reliable single-channel EEG framework for automated sleep stage classification without any preprocessing. *Article II* concentrates on developing a rapid and effective sleep scoring model with greatly reduced model parameters.

This thesis focuses on the application and development of CNN, LSTM, and the combination of CNN and LSTM based on single-channel EEG and many-to-one schemes in automatic sleep scoring. The proposed models in *Articles I, II, III, and V* are CNN-based. The presented framework in *Article IV* is CNN-LSTM based. Next, we will introduce the theoretical basis of CNN and LSTM.

2.3.1 Convolutional neural network (CNN)

In Figure 4, we present an illustration of a CNN model, which contains an input layer, multiple hidden layers, and an output layer (LeCun et al., 2015). The hidden layers generally comprise several convolutional, pooling, and fully connected layers. The convolutional layer extracts high-level feature presentations

from its input through the convolution operation. We can also determine the convolutional kernel's size and number to control the receptive field of convolution operation and the capacity of layer output, respectively. The convolution calculation is described as follows:

$$X^{(l)} = f^{l}\left(X^{(l-1)} * W^{(l)} + b^{(l)}\right), \tag{6}$$

where $X^{(l-1)}$ and $X^{(l)}$ are respectively the input and output of layer l, $W^{(l)}$ and $b^{(l)}$ denote the weight matrix and bias, respectively, $*$ refers the convolution calculation and $f^{(l)}$ stands for the activation function for adding the nonlinearization to convolution results. The activation function could improve the extraction capacity of high-level feature presentations by enhancing the model nonlinearity. The two most commonly used activation functions in the CNN are rectified linear unit (ReLU) (Nair and Hinton, 2010) and tanh (Lau and Lim, 2018). The ReLU and tanh functions are defined as follows:

$$\mathrm{ReLU}(x) = \max(0, x), \tag{7}$$

$$\tanh(x) = \left(e^{x} - e^{-x}\right) / \left(e^{x} + e^{-x}\right). \tag{8}$$

The feature maps obtained from the convolutional layer are down-sampled in space and orientation through the pooling layer. Each neuron in the pooling layer only accepts a few spatially adjacent elements of the same input feature map as input parameters and can map the input feature map to a specific position by taking the minimum, maximum, or average value. The average-pooling and max-pooling are two popular pooling methods in CNNs.

The output layer determines the final categorization using the softmax as the activation function, and the softmax is described as follows:

$$f(x_i) = e^{x_i} / \sum_{j=1}^{N} e^{x_j}, i = 1, \ldots, N, \tag{9}$$

where N is the class number. The probability of different classes can be calculated using the softmax function, with the class with the highest probability serving as the expected outcome.

In this thesis, we apply the 1D-CNN for the EEG time series input and 2D-CNN for the spectrogram and time-frequency image input.

2.3.2 Long Short-Term Memory neural network (LSTM)

Informally, the recurrent neural network (RNN) is a type of neural network that links all of the nodes in a chain while recursing in the direction of the sequence's progression(Dupond, 2019). This mechanism can allow the RNN to use previous events to infer subsequent events. However, conventional RNN struggles with Long-term dependencies and confronts the issue of vanishing and exploding gradients. The most well-known RNN variation, the LSTM, was presented by Hochreiter and Schmidhuber (1997) to cope with the problem of Long-term

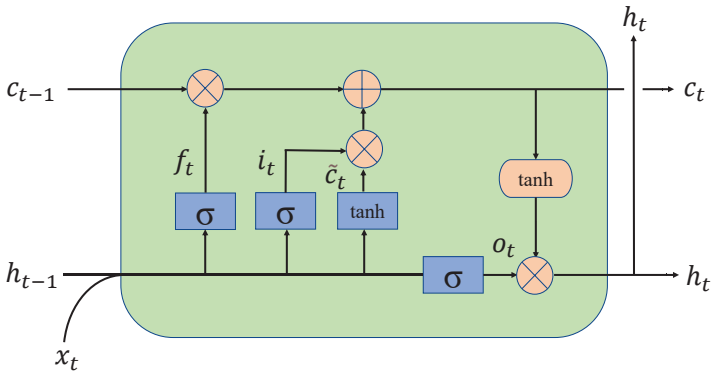


FIGURE 5 An illustration of the LSTM unit.

dependencies and vanishing and exploding gradients. During the learning process, LSTM can add useful information and discard irrelevant information (Zuo et al., 2022). Figure 5 illustrates the LSTM unit's fundamental architecture.

The LSTM unit can remove or add information to the cell state through three carefully designed gates (i.e., the input, forget and output gates). A gate is a is a mechanism for selectively permitting information to flow through, which comprises a pointwise multiplication operation and a sigmoid neural network layer. The sigmoid layer produces a number between 0 and 1 that indicates how much of each part can get through. A value of 0 denotes complete rejection and 1 complete acceptance. The definition of the sigmoid function is described as follows:

$$\sigma(x) = 1/(1 + \mathrm{e}^{-x}). \tag{10}$$

The forget gate is used in the LSTM's initial step to select which data should be removed from the cell state. The forget gate will receive the information of h_{t-1} and input x_t and output a value between 0 and 1 for cell state c_{t-1}, which is defined as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{11}$$

The second stage is to determine what new information is added to the cell state. The input gate determines what values to update through a sigmoid layer, which is described as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{12}$$

and a new candidate cell state $\tilde{C}_t$ is created with a tanh layer, which is defined as follows:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{13}$$

The third stage involves updating the old cell state C_{t-1} to the new cell state C_t, which is shown as follows:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t. \tag{14}$$

The final stage is determining what value to the output, which depends on the cell status C_t. The output layer selects which portion of the cell state to output

using the sigmoid layer, which is defined as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o). \tag{15}$$

The output of the LSTM unit can be acquired as follows:

$$h_t = o_t * \tanh(C_t). \tag{16}$$

2.4 Class imbalance problem (CIP) in automatic sleep stage classification

It is conceivable in the area of computer vision (CV) to ensure that some image datasets (e.g., CIFAR-10 database) include an equal quantity of each category. The sleep PSG database, however, suffers from significant CIP with an imbalanced class distribution due to the differences in sleep patterns between individuals' ages, genders, and physical conditions (Krishnan and Collop, 2006; Edwards et al., 2010). That is to say, certain sleep stages make up the majority, but other sleep stages fall into the minority. The proportion of each sleep stage in a regular night's sleep, excluding stage W, is shown in Figure 6. We can observe that the percentage of each stage is drastically unequal, and the class distribution is imbalanced. Specifically, stage W has the lowest proportion, barely a tenth of stage N2's proportion (i.e., 45%). Stages N3 and REM share the same percentage, which is 25%.

2.4.1 The definition of CIP in PSG datasets

To quantificationally evaluate the severity of class imbalance in PSG datasets, we first introduce a class imbalance factor in *Articles III and IV*, which is defined as

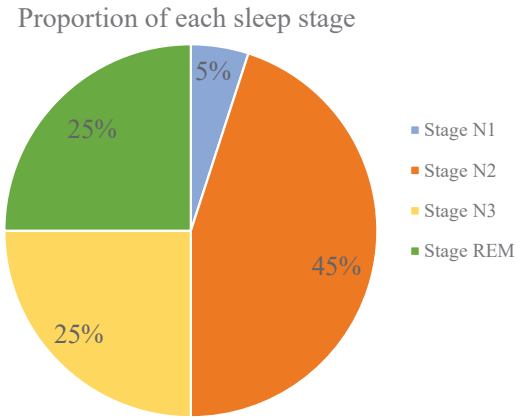


FIGURE 6 The proportion of each sleep stage in a regular night's sleep excluding stage W.

follows:

$$CIF = \frac{N}{2 \cdot c \cdot \min\{N_i\}} \quad i \in \{1, 2, \ldots, c\}, \tag{17}$$

where N is the total numbers, c denotes the number of the sleep class, and N_i refers to each sleep stage quantity. A value of $CIF = 0.5$ indicates that the PSG dataset could be considered balanced. A value of $CIF > 0.5$ in (17) implies that the PSG dataset experiences the CIP. Additionally, a greater CIF suggests that the PSG dataset is more imbalanced. The CIF of the CCSHS, Sleep-EDF-V1, and Sleep-EDF datasets are 3.6, 1.6, and 1.5, respectively, in which the CCSHS is the most imbalanced PSG dataset. The prediction model faces crucial challenges due to CIP since most machine learning or deep learning algorithms for classification were developed with the presumption that each group would have an equal amount of samples. Each class has an equal loss weight, which might result in prejudice towards the minority class and inadequate model training.

2.4.2 Balance the dataset samples

The most direct approach to tackle the CIP is to expand the proportion of the minority sleep stage to balance the dataset samples in the training set. Only a few research on sleep staging models have provided approaches to overcome CIP thus far. There are two methods to increase the percentage of the under-represented class: the under-sampling and the over-sampling. The under-sampling method discards large amounts of samples from the majority class that is not necessary to introduce new data. However, the assessment model may experience underfitting with a reduction in training samples. Tsinalis et al. (2016) introduced a class-balanced random sampling method to prevent biased performance on the majority sleep stages and substantially increase the stage N1 classification performance. Nevertheless, the overall accuracy, 78%, was not good enough. The class-balanced random sample lessened the significance of majority classes providing the principal contribution to classification performance, which is a significant factor.

By contrast, the over-sampling strategy directly increases the minority class quantity. The most used and easiest way is to randomly replicate samples from minority classes, which has been validated by Supratak et al. (2017) and Fan et al. (2020). The primary flaw is that simply repeating can not introduce new variabilities in the training process. In addition, The synthetic minority oversampling (SMOTE) approach allows us to create new samples for the minority class as well (Chawla et al., 2002). Fan et al. (2020) examined the effectiveness of five data argumentation (DA) methods for sleep EEG data, involving repeating minority classes, morphological change, signal segmentation and recombination, dataset-to-dataset transfer, and generative adversarial network (GAN). The morphological change of EEG signals was achieved by horizontal movements. The signal segmentation and recombination method first segments the 30 seconds of sleep epochs into smaller chunks. After that, random segment selection and recombination from the same class create new sleep epochs. The dataset-to-dataset

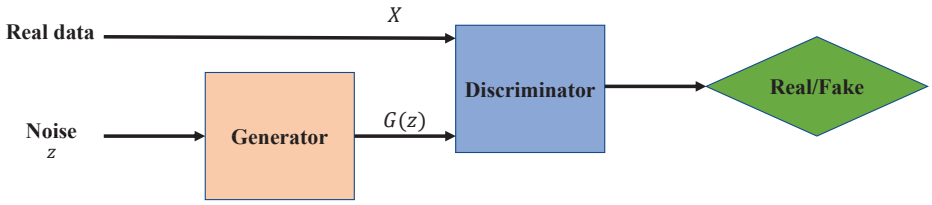


FIGURE 7 The original architecture of generative adversarial network (GAN) model.

transfer approach is transferring learning (Pan and Yang, 2010), which is capable of transferring sleep signals across datasets. The original architecture of GAN is demonstrated in Figure 7, which comprises two opposing networks (generator (G) and discriminator (D)) (Goodfellow et al., 2020). The purpose of G is to convert z, the noise variable, into G(z), the produced sample that learns p_{data}, the distribution of the real data x. The D is responsible for distinguishing whether a sample is real or artificial. G and D are optimized by function $V(D, G)$ as follows:

$$\min_G \max V(D, G) = E_{x \sim p_{\text{data}}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))], \quad (18)$$

where $D(x)$ refers to the probability of x sampled from the real samples p_{data}. $G(z)$ presents the fake signals created by the G. The overall performance is enhanced by the proposed five DA method. However, stage N1 accuracy indicated a slight decline. Sun et al. (2019) proposed a noise addition ranging from 8 to 14 dB method for EEG signals, the use of white noise's signal-to-noise ratio (SNR) might be expanded to study the effectiveness of various noise levels. In addition, the percentage of each group is targeted to be the same as in previous studies (e.g., Supratak et al., 2017; Sun et al., 2019; Fan et al., 2020). However, doing so drastically damages the fundamental sleep architecture. In *Articles III, IV*, we only increase the number of stage N1, as N1 is the representative of minority groups and is the most challenging to classify.

2.4.3 Balance the network connection

The CIP is imbalanced in the class distribution and network connection. Apart from balancing the dataset samples, how to balance the network connection between the trained model and the imbalanced PSG dataset is an alternative. It is more practically meaningful to boost performance without altering the distribution of classes, which can keep the overnight sleep structure intact. Each class's weight is equal by default for most DL-b methods. As a result, the dominating weight updating belongs to the majority class and has a longer gradient component. Additionally, the trained model unfairly influences how well the minority classes perform. For instance, the stage N1 accuracy is only around 50% (Li et al., 2017; Phan et al., 2018a; Seo et al., 2020), which is the most misclassified sleep stage.

The most straightforward way to eliminate the discrimination is to reassign the class weight of each sleep stage. The minority class can be set by a larger class

weight, while the majority class can be assigned to a smaller class weight. Wang and Wu (2018) determined the class weight lying on the ratio of the numbers of the majority class and each class, which is calculated as follows:

$$w_l = \frac{\max\{N_l\}_{l=1}^{L}}{N_l}, \tag{19}$$

where N_l is the number of class l samples, and L is the class number. While in Kwon et al. (2021), the values of each class weight (w_l) depend on the ratio of the numbers of the whole class and each class, which is defined as follows:

$$w_l = \frac{\sum_{i=1}^{L} N_l}{N_l}. \tag{20}$$

Zhou et al. (2020) introduced a balance coefficient pi adjusting by the grid search method was introduced to modify the class weight calculation as follows:

$$w_l = pi\frac{\sum_{i=1}^{L} N_l}{N_l}. \tag{21}$$

In Zhu et al. (2020), the number of samples did not entirely determine the class weights in each class. The class weight of majority (W), intermediate (N2, N3, and REM), and minority categories (N1) were set to 4, 2, and 1, respectively. Wang et al. (2022) proposed a more complicated strategy. The value of class weight was normalized to 1-5 as follows:

$$w_l = \min\left[5, \max\left(1, \ln\left(\frac{1}{p(class)}\right)\right)\right], \tag{22}$$

where $p(class)$ is the percentage of a particular class label to the entire label. In *Article IV*, we evaluate three class weight redistribution strategies to balance the network connection. The first is the approach used by Kwon et al. (2021), and the second is to obtain the natural logarithm of the class weight in the first method. The third is to set the class weight referring to the neuroscience principles (Zeng et al., 2017). More details can be found in *Article IV*.

2.5 Model interpretability in sleep stage classification

While deep learning (DL) is commonly employed in research on automatic sleep staging, most of these research do not provide the underlying mechanisms of their classifiers. It is still unknown how the DL-b methods make the right decision for sleep scoring, and their black-box nature hampers the deployment of DL-b models in clinical settings. The model explainability is one of the most crucial issues in automatic sleep stage classification that should be urgently addressed. The t-distributed stochastic neighbor embedding (t-SNE) method (Van der Maaten and Hinton, 2008) has been used successfully to visualize the output of the model

layer in many automatic sleep scoring studies (e.g., Jiang et al., 2019; Yang et al., 2021; Yan et al., 2021b; Zhao et al., 2021; He et al., 2022; Decat et al., 2022).

We will introduce the basics of the t-SNE approach in Section 2.5.1. The t-SNE technique is capable of converting high-dimensional data into two-dimensional data to provide feature visualization of each layer, which can relate the classification results of each sleep stage in each model layer. However, the characteristics learned by each layer of the applied model were not shown using the t-SNE method. Ellis et al. (2021) presented an ablation method presented to determine the significance of each modality signal to the applied CNN-based model. The significance of each modality can be evaluated by performance comparison before and after the ablation of each modality.

Nevertheless, the ablation strategy cannot effectively adapt to the single-channel EEG-based automatic sleep scoring models. Additionally, it is yet unclear what properties the model picks up from the input and whether or not these qualities are connected to different stages of sleep. A potential strategy, namely layer-wise relevance propagation (LRP) (Bach et al., 2015), is potentially applied to explain the DL-b method in automatic sleep scoring, which is also capable of the single-channel EEG-based scheme. In addition, the LRP-based method can determine whether the applied model can correctly identify specific EEG signals in each sleep state for making the correct decision. The basic theory of the LRP method is illustrated in Section 2.5.2. In *Article V*, we present an interpretable sleep scoring based on the LRP approach with single-channel EEG.

2.5.1 The t-distributed stochastic neighbor embedding (t-SNE)

The t-SNE method is an improved variation of Stochastic Neighbor Embedding (SNE) (Hinton and Roweis, 2002), which visualizes high-dimensional data by assigning a position to each datapoint on a two or three-dimensional map (Van der Maaten and Hinton, 2008). The SNE uses conditional probability ($p_{j|i}$) to describe the similarity between two data. Assuming that there are two points in a high-dimensional space, then the variance is constructed as a Gaussian distribution (σ_i) centered on the point x_i. For neighboring data points, $p_{j|i}$ is relatively higher than widely dispersed data points. The $p_{j|i}$ is given as follows:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}. \tag{23}$$

Then in the low-dimensional space, such conditional probability can also be used to define the distance between y_i and y_j. The σ_i is set to $1/\sqrt{2}$, then the conditional probability can be expressed as follows:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}. \tag{24}$$

The SNE employs a gradient descent method to minimize the total of Kullback-Leibler (KL) divergences across all datapoints, the cost function C is defined as

follows:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{25}$$

In t-SNE, the heavy-tailed distribution in the low-dimensional map is represented by a Student t-distribution with one degree of freedom. The joint probabilities can be defined as follows:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq I} \left(1 + \|y_k - y_l\|^2\right)^{-1}}. \tag{26}$$

The t-SNE method can efficiently solve the crowding problem in SNE, which is caused by the difference between the high-dimensional spatial distance distribution and the low-dimensional spatial distance distribution. More details can be found in Van der Maaten and Hinton (2008).

2.5.2 The Layer-wise relevance propagation

The LRP method is presented to analyze how each pixel of the input image x contributed to the prediction $f(x)$ (Bach et al., 2015). We illustrate the diagram of the LRP approach in Figure 8. The LRP method presupposes that a sum of terms of the various input dimensions x_d could describe the prediction $f(x)$:

$$f(x) = R_f = \sum_d R_d(x), \tag{27}$$

where R_f represents the relevance of prediction $f(x)$ and $R_d(x)$ refers to the relevance that results for the pixel x_d of input image x. It should be noted that the total importance of all nodes in each layer should be equal:

$$\sum_d R_d^{(1)} = \ldots = \sum_i R_i^{(l-1)} = \sum_j R_j^{(l)} = \ldots = R_f. \tag{28}$$

The relevance may be viewed as information flowing over the network connection, with the flow direction being from the output node to the input node. According to the concept of backpropagation, we may deconstruct the relevance along the sub-paths between nodes, as shown in Figure 8. Here, assuming that i stands for the sequence number of the lower layer neuron and j for the higher layer neuron:

$$R_{i \leftarrow j}^{(l-1,l)} = factor_{ij}^{(l-1,l)} \cdot R_j^{(l)}, \tag{29}$$

where the distribution factor, $factor_{ij}$, fits into the range of 0 to 1, is given by:

$$\sum_i factor_{ij}^{(l-1,l)} = 1. \tag{30}$$

For each higher layer neuron, $z_j^{(l)} = W_j^{(l)} a^{(l-1)}$ is the input and $a(l-1)$ represents the lower layer neuron's activation output vector. The relevance distribution factor between the lower layer neuron i and higher layer neuron j can

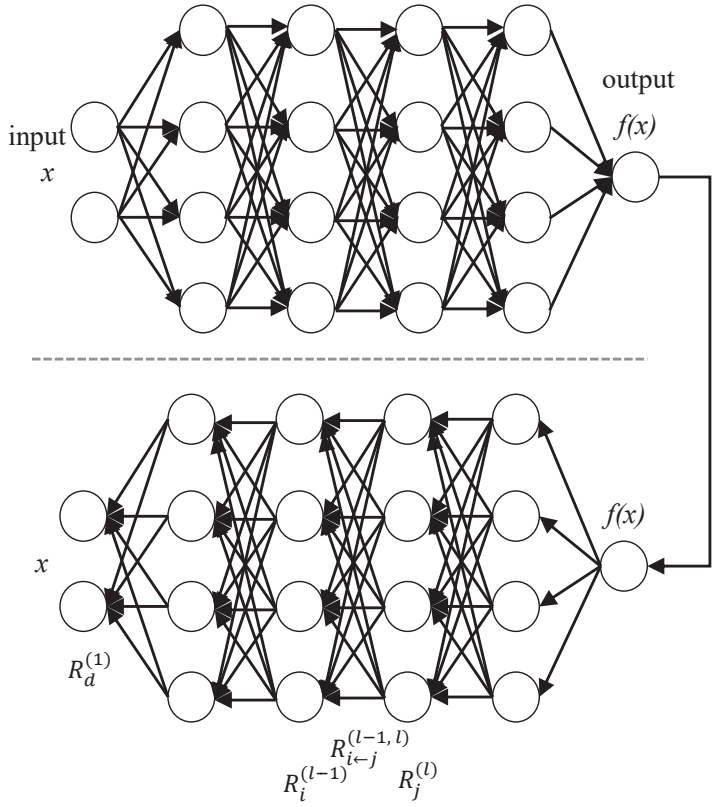


FIGURE 8 The diagram of the layer-wise relevance propagation method.

be expressed by each component $z_{ij}^{(l)}$ of $z_j^{(l)}$. To fulfill the restriction of (30), it is split as follows by the normalization factor $z_j^{(l)}$:

$$factor_{ij} = \frac{z_{ij}^{(l)}}{z_j^{(l)}} = \frac{w_{ij}^{(l)} a_i^{(l-1)}}{\sum_i w_{ij}^{(l)} a_i^{(l-1)}}. \tag{31}$$

Eq. (29) can be therefore rewritten by

$$R_{i\leftarrow j}^{(l-1,l)} = \frac{w_{ij}^{(l)} a_i^{(l-1)}}{\sum_i w_{ij}^{(l)} a_i^{(l-1)}} \cdot R_j^{(l)}. \tag{32}$$

For each input pixel x_d, the resultant relevance $R_d(x)$ may be converted to color space and then shown using a conventional heat mapping. In *Article V*, we first get the time-frequency image using the STFT method, and the frequency at every time point is represented by each pixel of the input time-frequency image. With the heat mapping, we can accurately determine if the EEG patterns related to a given sleep stage can be detected and are necessary for the applied model to distinguish this particular sleep stage.

3 SUMMARIES OF INCLUDED ARTICLES AND AUTHOR CONTRIBUTIONS

This chapter provides an overview of each article, including the objective, methods, results, conclusion, and discussion. The author contributions to the per article are also explicated.

3.1 *Article I*: SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG

Dongdong Zhou, Jian Wang, Guoqiang Hu, Jiacheng Zhang, Fan Li, Rui Yan, Lauri Kettunen, Zheng Chang, Qi Xu, and Fengyu Cong. (2022). SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG. Biomedical Signal Processing and Control, 75, 103592.

Objective

Manual sleep stage classification is tedious and time-consuming for sleep technicians (Phan et al., 2018b; Seo et al., 2020). Recently, many automatic sleep scoring methods have been successfully proposed, including the traditional machine learning-based and deep learning-based approaches. In general, the performance of machine learning-based methods is heavily reliant on the selection of handmade features, which is limited by poor generalization ability. By contrast, the deep learning-based approaches could efficiently learn features from the model input to avoid using the pre-extract features. In addition, the multi-channel signals strategy increases the computational cost and hinders the practical application due to the complicated data acquisition scheme. We propose an automated sleep scoring method based on single-channel EEG without employing feature extraction to fill these gaps.

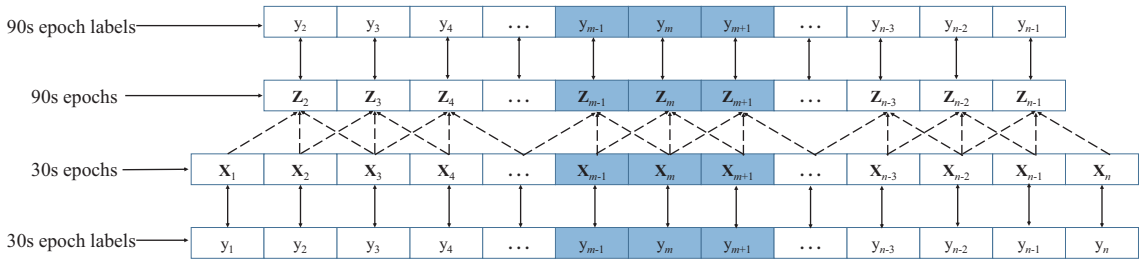


FIGURE 9 Illustration of 90s epochs and labels used in this paper. n denotes the number of 30s epochs for a subject, $\mathbf{Z}_m$ is comprised of X_{m-1}, X_m and X_{m+1}, $2 \leq m \leq n-1$.

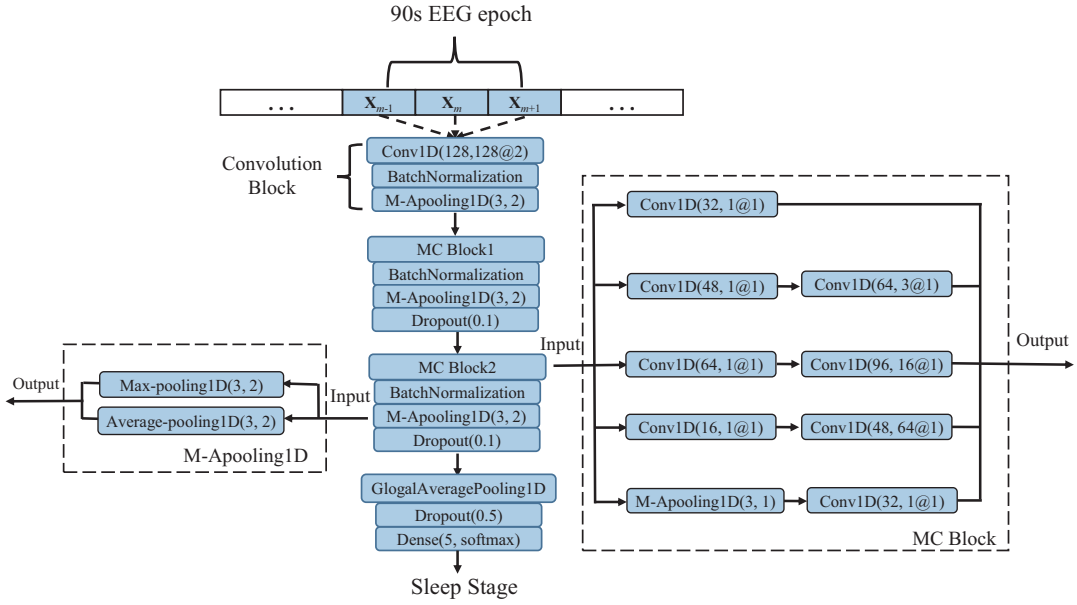


FIGURE 10 The overall architecture of the proposed SingleChannelNet (SCNet).

Methods

We adopt three public PSG datasets (CCSHS, Sleep-EDF-V1, Sleep-EDF, described in Section 2.1) in this study to evaluate the performance of the proposed SingleChannelNet.

In order to replicate how sleep clinicians perform manual sleep scoring (Chambon et al., 2018), we reconstruct a 90-second epoch as the contextual input. This many-to-one scheme is illustrated in Figure 9. Additionally, we only adopt 30-minute samples of stage W before and after other sleep stages, and other recordings of stage W are eliminated (Supratak et al., 2017; Qu et al., 2020). Inspired by the inception module (Szegedy et al., 2015), we design a multi-convolution (MC) block containing three different sizes of filters (i.e., small size: 3, medium size: 16 and big size: 64, Figure 10) for leaning multi-scale feature presentations from long time series input. Similarly, we concatenate the max-pooling

and average-pooling layers as the M-Apooling to further enhance the feature extraction capacity of the proposed SCNet model. In addition, we replace the traditional fully connected layer with the Global Average Pooling (GAP) layer without introducing trainable parameters (Lin et al., 2013).

Results

Regarding the model performance, *ACC* and *K* achieved by the proposed SCNet are as follows:

- CCSHS: $ACC = 90.2\%$, $K = 86.5\%$,
- Sleep-EDF: $ACC = 86.1\%$, $K = 80.5\%$,
- Sleep-EDF-V1: $ACC = 91.0\%$, $K = 87.8\%$.

We also make a performance comparison between the one-to-one (30-second input) and many-to-one (90-second input) schemes on CCSHS and Sleep-EDF datasets. Utilizing the many-to-one strategy, *ACC* could be improved by 1.1% and 4.1% on the CCSHS, and Sleep-EDF datasets, respectively. Likewise, *K* (Cohen, 1960) could be enhanced by 1.5% and 5.7%. Comparing with other state-of-the-art methods using the same dataset (Li et al., 2017; Phan et al., 2018a; Nakamura et al., 2019; Phan et al., 2019; Mousavi et al., 2019; Supratak and Guo, 2020), our model could obtain better performance.

Conclusion and discussion

In this work, we propose an effective CNN-based model, SCNet, for automatic sleep stage classification with raw single-channel EEG, combining feature extraction and classification abilities. The proposed SCNet could achieve promising performance on three PSG datasets with different characters, which shows robust model generalization. Furthermore, the single-channel strategy is suitable for portable household sleep monitoring devices with lower time delays and a more comfortable customer experience. However, *ACC* of stage W is much lower than that of sleep stages, mainly due to the class imbalance problem. This issue needs to be further explored.

Author contributions in *Article I*

Dongdong Zhou: Conceptualization, Methodology, Software, Writing original draft. **Jian Wang**: Writing – review & editing. **Guoqiang Hu**: Writing – review & editing. **Jiacheng Zhang**: Writing – review & editing. **Fan Li**: Review & editing. **Rui Yan**: Review & editing. **Lauri Kettunen**: Supervision. **Qu Xi**: Review & editing, Supervision. **Fengyu Cong**: Supervision.

3.2 *Article II*: LightSleepNet: A Lightweight Deep Model for Rapid Sleep Stage Classification with Spectrograms

Dongdong Zhou, Qi Xu, Jian Wang, Jiacheng Zhang, Guoqiang Hu, Lauri Kettunen, Zheng Chang, and Fengyu Cong. (2021). LightSleepNet: A Lightweight Deep Model for Rapid Sleep Stage Classification with Spectrograms. 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2021) (pp.43-46), IEEE.

Objective

The exceptional success that deep learning has made in classifying sleep stages is starting to open the door for possible real-world applications (Sun et al., 2020; Eldele et al., 2021; Sekkal et al., 2022; You et al., 2022). However, numerous deep neural network-based methods have overly complicated architectures with millions of model parameters resulting in potentially problematic overfitting and a great demand on computational resources. Therefore, a simple contact model would be required in future real-life implementation. In *Article I*, we also observe that training the lengthy series input is not sufficiently efficient with the CNN-based model. How to enhance the training efficiency is another factor worthy of being considered.

Methods

PSG datasets SHHS-100, Sleep-EDF, and Sleep-EDF-V1 are employed in this study. More details can be found in Section 2.1. We randomly chose 100 subjects (i.e., SHHS-100) from the subset (SHHS1) with the criteria: the Respiratory Disturbance Index 3 Percent (RDI3P) was less than 15, and there were no reports of high pressure. Small-scale SHHS-100 aims to validate the few-shot learning ability of the proposed lightweight model (LightSleepNet, LSNet) for sleep scoring sleep assessments. The LSNet is a simplified version (2D CNN) of the proposed SCNet in *Article I*, which is designed for two-dimensional input (shown in Figure 11). The filter sizes are 1×1, 3×3, 7×7 in LSNet. We transfer the raw EEG signal into spectrograms to accelerate the training speed through the short-time Fourier transform (STFT). The size of the spectrogram input is $F \times T$, where $F = 61$, $T = 89$. To demonstrate the efficiency of the spectrogram input, we construct a baseline model whose structure and training setup are consistent with the LSNet except for the filter size. The filter size of $N \times N$ in LSNet is replaced by the filer size of N in the baseline model, and the input of the baseline model is long-time series EEG.

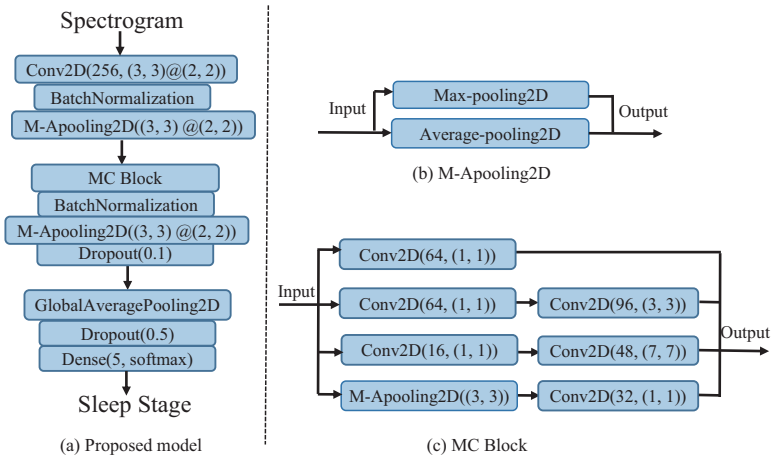


FIGURE 11 The overall architecture of proposed LightSleepNet (LSNet) model.

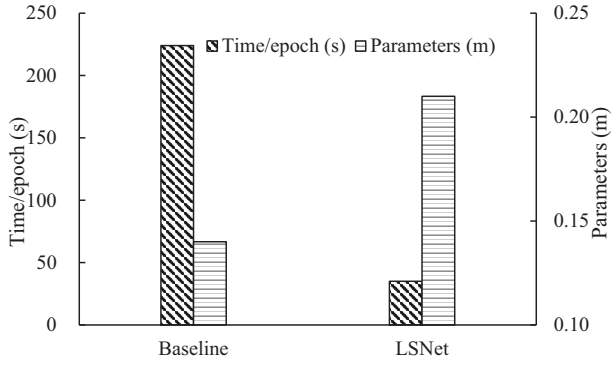


FIGURE 12 The comparison between the baseline and LSNet in terms of the model parameters and computational cost for each training epoch. The black diagonal stripe represents the training time of each iteration, the gray cross stripe denotes the number of model parameters.

Results

We first compare the number of model parameters and the computational cost for each training epoch between the LSNet and baseline model. As demonstrated in Figure 12, we can see that the baseline model has around 0.14 million (**m**) parameters with an iteration time cost of 224 seconds. Regarding the LSNet, even though LSNet contains more parameters (roughly 0.21 **m**), the time consumption of each training epoch is only 35 seconds, which is less than a sixth of the baseline model. Furthermore, we compare the model performance and the number of model parameters with other state-of-the-art methods on the same dataset. There are around 2.2 **m**, 1.3 **m**, 1.3 **m**, 21 **m**, and 2.6 **m** model parameters in Sors et al. (2018), Zhang et al. (2019), Supratak and Guo (2020), Supratak et al. (2017) and Mousavi et al. (2019), respectively. In contrast, our proposed LSNet is much simpler. The number of model parameters is at least six times smaller than men-

tioned studies. Additionally, our LSNet could attain comparable (near or better) performance on three datasets (SHHS100: 86.7%-81.3%, Sleep-EDF: 83.7%-77.5%, Sleep-EDF-V1: 88.3%-84.5%). The goal of decreasing the model parameters is achieved without sacrificing model performance.

Conclusion and discussion

This work presents LSNet, a simple yet powerful CNN-based model for quick and accurate sleep stage classification. We verify the efficiency of spectrogram inputs, which could speed up the computational training cost immensely. More importantly, the lightweight characteristic enables the proposed LSNet's potential practical application in real life. In addition, more energy-efficient brain-inspired models, such as spiking neural networks (Yu et al., 2018; Xu et al., 2020, 2021), could be investigated in the future.

Author contributions in *Article II*

Dongdong Zhou: Conceptualization, Methodology, Software, Writing original draft. **Qu Xi**: Writing – review & editing, Supervision. **Jian Wang**: Writing – review & editing. **Guoqiang Hu**: Writing – review & editing. **Jiacheng Zhang**: Writing – review & editing. **Lauri Kettunen**: Writing – review & editing, Supervision. **Fengyu Cong**: Writing – review & editing, Supervision.

3.3 *Article III*: Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance

Qi Xu, **Dongdong Zhou**, Jian Wang, Jiangrong Shen, Lauri Kettunen, and Fengyu Cong. (2022). Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance. 2022 International Joint Conference on Neural Networks, (IJCNN 2022) (pp.1-6), IEEE.

Objective

Due to unique sleep patterns, the quantity of each sleep stage in most of the available PSG datasets is significantly uneven (i.e., class imbalance problem, CIP). The prediction model faces critical challenges due to CIP since most machine learning or deep learning algorithms for classification were developed with the presumption that there would be an equal number of samples in each class. Herein, stage N1, as a presentative of the minority class, generally accounts for 2-5% (Altevogt and Colten, 2006; Fan et al., 2020; Zhou et al., 2022b). Besides, the recognition accuracy of N1 is always the lowest among the five sleep stages. The performance of the minority category suffers discrimination from the applied model. Additionally, the minority class's poor accuracy serves as a restriction on the overall accuracy.

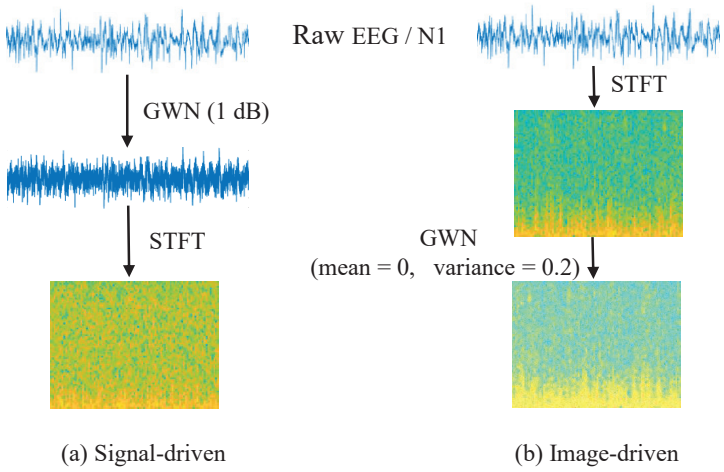


FIGURE 13 The proposed two balancing methods: signal-driven and image-driven. STFT: short-time Fourier transform, GWN: Gaussian white noise

Methods

The most direct approach is to increase the number of the minority class. In Dong et al. (2017), Sun et al. (2019), and Fan et al. (2020), the percentage of each group is set to be the same. By doing this, the initial sleep structure is severely damaged. This study only raises the stage N1 quantity in the training set. We first introduce a class imbalance factor (CIF) to define the severity of the CIP in PSG datasets quantitatively.

We propose two balancing methods, signal-driven and image-driven, to balance the dataset samples, which is shown in Figure 13. Unlike duplicating samples from the minority class (Supratak et al., 2017), we generate new samples of N1 by adding different intensities of Gaussian white noise (GWN). Three intensities are defined in two balancing methods: signal-driven (low: 10 dB, moderate: 5 dB, and high: 1 dB) and image-driven (the mean (M) is 0, the variances (V) are 0.05, 0.1, 0.2 respectively). The signal-driven method initially adds the Gaussian white noise to the raw EEG signal and then converts the noisy EEG signal to the time-frequency image with STFT. While in the picture-driven method, the raw EEG signal is first converted to the time-frequency image, to which Gaussian white noise is then added. We validate the efficiency of the proposed two balancing methods on the CCSHS, Sleep-EDF, and Sleep-EDF-V1 datasets.

Results

We first compare the performance between the baseline model without and with the proposed balancing methods. A 1 dB Gaussian white addition could achieve the most notable ACC and K improvement for the signal-driven method in three datasets:

- Sleep-EDF-V1: ACC +0.8%, K +1.2%.
- Sleep-EDF: ACC +0.2%, K +0.3%.

– CCSHS: *ACC* +0.5%, *K* +0.5%.

The accuracies of the N1 stage are improved as follows:

– Sleep-EDF-V1: $(38.9 + 3.8)\%$ with 5 dB,
– Sleep-EDF: $(24.6 + 1.5)\%$ with 10 dB,
– CCSHS: $(22.9 + 4.4)\%$ with 1 dB.

On the Sleep-EDF-V1 and Sleep-EDF datasets, *ACC* improvement is the same for the low and moderate intensities (V = 0.05 and 0.1) of the image-driven DA. However, on the Sleep-EDF and CCSHS datasets, only the high intensity (V = 0.2) experiences a mild increase (1.3% and 0.1%) of N1 accuracy. Compared to other state-of-the-art approaches utilizing the same dataset (Phan et al., 2018a; Fan et al., 2020; Zhou et al., 2021; Phan et al., 2021), our proposed baseline model can also achieve overall performance.

Conclusion and discussion

The PSG datasets' inherent CIP has significantly hampered the use of automatic sleep scoring algorithms in real-world settings. This research investigates possible CIP solutions for automatic sleep scoring methods. In order to measure the degree of imbalance in three widely used PSG datasets, we first define the CIF. Although there are no discernible differences between the two proposed balancing methods in terms of model performance improvement, different intensities could enhance the overall and N1 stage categorization rates.

Author contributions in *Article III*

Qu Xi: Conceptualization, Methodology, Software, Writing original draft. **Dongdong Zhou**: Conceptualization, Methodology, Software, Writing original draft, Co-first author. **Jian Wang**: Writing – review & editing. **Jiangrong Shen**: Writing – review & editing. **Lauri Kettunen**: Writing – review & editing, Supervision. **Fengyu Cong**: Writing – review & editing, Supervision.

3.4 *Article IV*: Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification

Dongdong Zhou, Qi Xu, Jian Wang, Hongming Xu, Lauri Kettunen, Zheng Chang, and Fengyu Cong. (2022). Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification. IEEE Transactions on Instrumentation and Measurement, 75, 1-12.

Objective

In *Article III*, we investigate two different methods for adding gaussian white noise to the time-frequency image to raise the stage N1 numbers. However,

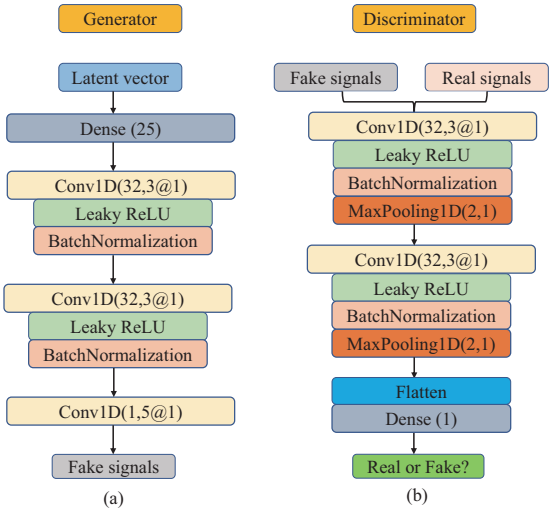


FIGURE 14 The framework of the GAN model. Demonstrate the structure: (a) Generator, (b) Discriminator.

other data augmentation methods (DA), such as generative adversarial network (GAN), and different GWN addition methods, have not been investigated. How to balance the deep neural network is another factor worth considering in addition to balancing the samples.

Methods

In this study, we first propose a GAN model to create fictitious signals of stage N1 (see Figure 14). The discriminator is intended to distinguish fake signals created by the generator. We further explore the effectiveness of different intensities and times GWN in addition to the raw EEG signal. This balancing method can be performed on raw EEG data while maintaining EEG characteristics. We add four intensities (1, 2, 5, and 10 dB) GWN to the raw EGG signal to boost the N1 numbers in the training set. In terms of the amplitude, we illustrate this DA method using various intensities and the DA with the GAN model in Figure 15. The efficiency of different times GWN addition is then evaluated. When the ideal intensity x dB has been determined, the intensities of GWN addition are specified as follows:

- three times: $x - 0.2$, x, $x + 0.2$ (dB),
- five times: $x - 0.2$, $x - 0.1$, x, $x + 0.1$, $x + 0.2$ (dB).

Finally, we seek to balance the relationship between the imbalanced dataset and the trained model without modifying the data distribution. The class weight (CW) is rearranged based on the ratio of the numbers of all samples to each class and the brain-inspired rule (Zeng et al., 2017). The W_i, W_j using CW (Ratio), CW

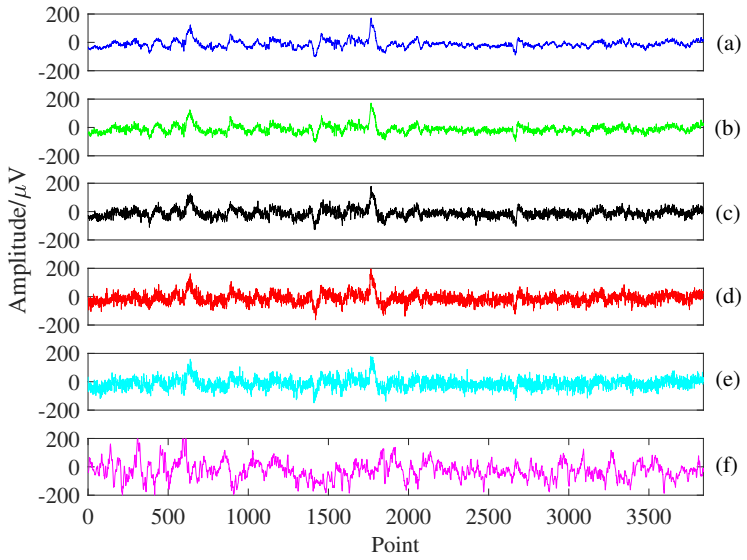


FIGURE 15 Raw EEG signal (N1 stage) and Gaussian white noise addition with four SNR. (a) Raw EEG. (b) Gaussian white noise addition with 10 dB. (c) Gaussian white noise addition with 5 dB. (d) Gaussian white noise addition with 2 dB. (e) Gaussian white noise addition with 1 dB. (f) Artificial signal by the proposed GAN model.

(Log_R) methods are defined as follows:

$$W_i = \frac{N}{N_i} \qquad i \in \{1, 2, \ldots, 5\}, \tag{33}$$

$$W_j = \ln \frac{N}{N_j} \qquad j \in \{1, 2, \ldots, 5\}, \tag{34}$$

where N, N_i and N_j represent the numbers of all classes, class i and j samples, respectively. Considering how neurons are distributed in the human brain's information-processing processes (Heeger and Ress, 2002; Zeng et al., 2017), we put the weight of the N1 stage at 8.5 and the weight of the remaining stages at 1.5 (CW(E_I)). We build a CNN-LSTM model as a Baseline model to validate the performance of proposed balancing methods on CCSHS, Sleep-EDF, and Sleep-EDF-V1 datasets.

Results

The proposed GAN model can increase overall accuracy compared to the Baseline model. However, it exhibits a slight decline in N1 accuracy on the experimental datasets. On three datasets, the GWN approaches have improved *ACC*, *K*, and N1 accuracy to varying degrees. In particular, the enhancement of N1 accuracy is as follows:

- CCSHS: 9.7% with Baseline + GWN (1 dB),
- Sleep-EDF: 16.2% with Baseline + GWN (10 dB),

– Sleep-EDF-V1: 12.0% with Baseline + GWN (1 dB).

Besides, more N1 stage samples (three and five times GWN addition) could not concurrently produce superior overall and N1 accuracy than the Baseline + GWN (1 dB). All CW methods result in a significant increase in the N1 accuracy, with values on the CCSHS dataset of 52.0% for CW (Ration), 28.3% for CW (Log_R), and 44.7% for CW (E_I). However, *ACC* and *K* marginally decline in place. In comparison, *ACC* and *K* only slightly improve on the Sleep-EDF and Sleep-EDF-V1 datasets, except for the CW (Log_R) and CW (E_I) approaches on the Sleep-EDF dataset. In addition, N1 accuracy improvements are not as significant as those on the CCSHS dataset.

Conclusion and discussion

In this study, we introduce two balancing methods to alleviate the CIP in automatic sleep scoring tasks. One method is to balance the dataset samples by increasing the number of the minority class (i.e., N1). Another solution is to redesign the class weight to balance the network connection. According to the obtained results, the proposed approaches could enhance biased performance. This work provides new avenues for further addressing the automatic sleep scoring class imbalance problem.

Author contributions in *Article IV*

Dongdong Zhou: Conceptualization, Methodology, Software, Writing original draft. **Qu Xi**: Writing – review & editing, Supervision. **Jian Wang**: Writing – review & editing. **Hongming Xu**: Writing – review & editing. **Lauri Kettunen**: Writing – review & editing, Supervision. **Zheng Chang**: Writing – review & editing, Supervision. **Fengyu Cong**: Writing – review & editing, Supervision.

3.5 *Article V*: Interpretable Sleep Stage Classification Based on Layer-wise Relevance Propagation

Dongdong Zhou, Qi Xu, Jiacheng Zhang, Lei Wu, Lauri Kettunen, Zheng Chang, Hongming Xu, and Fengyu Cong. Interpretable Sleep Stage Classification Based on Layer-wise Relevance Propagation. Under review.

Objective

Recently, numerous deep learning-based (DL-b) methods for automatic sleep scoring objectives have been successfully established and have made substantial development (Korkalainen et al., 2020; Jia et al., 2021; Guillot and Thorey, 2021; Zarei et al., 2022). However, it is still a long way from practical clinical application. One of the most crucial factors is the inadequate model explanation of the

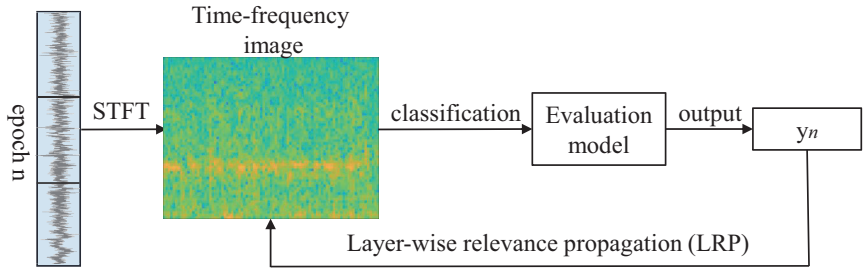


FIGURE 16 The overall schematic diagram of the interpretable sleep stage classification with layer-wise relevance propagation.

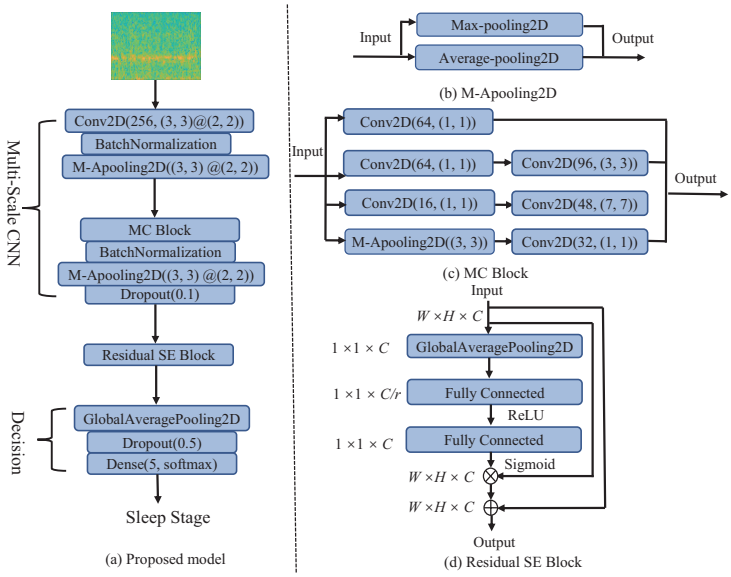


FIGURE 17 The structure of the proposed MSSENet: a) an overall of the proposed framework; b) the M-Apooling, which combines the Max-pooling and Average-pooling; c) the structure of MC Block, which includes different filter sizes; d) the SE Block with the shortcut connection strategy.

DL-b methods, and the DL-b methods' application in real life is met with skepticism by sleep specialists. Establishing trust among practitioners and properly articulating how the deep model makes choices is a crucial step. The t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) is a widely used tool for data exploration and model visualization in the automatic sleep scoring (e.g., Van Leeuwen et al., 2019; Yang et al., 2021; Huang et al., 2022), however, the characteristics learned by each layer of the applied model were not depicted. The ablation method is another popular method to validate the importance of each modality to the model decision (e.g., Jia et al., 2020; Neng et al., 2021; Ellis et al., 2021). Nevertheless, the ablation strategy cannot effectively adapt to the single-channel EEG-based automatic sleep scoring models. It is also unclear what features the model picks up from the input and whether learned features are associated with sleep stages.

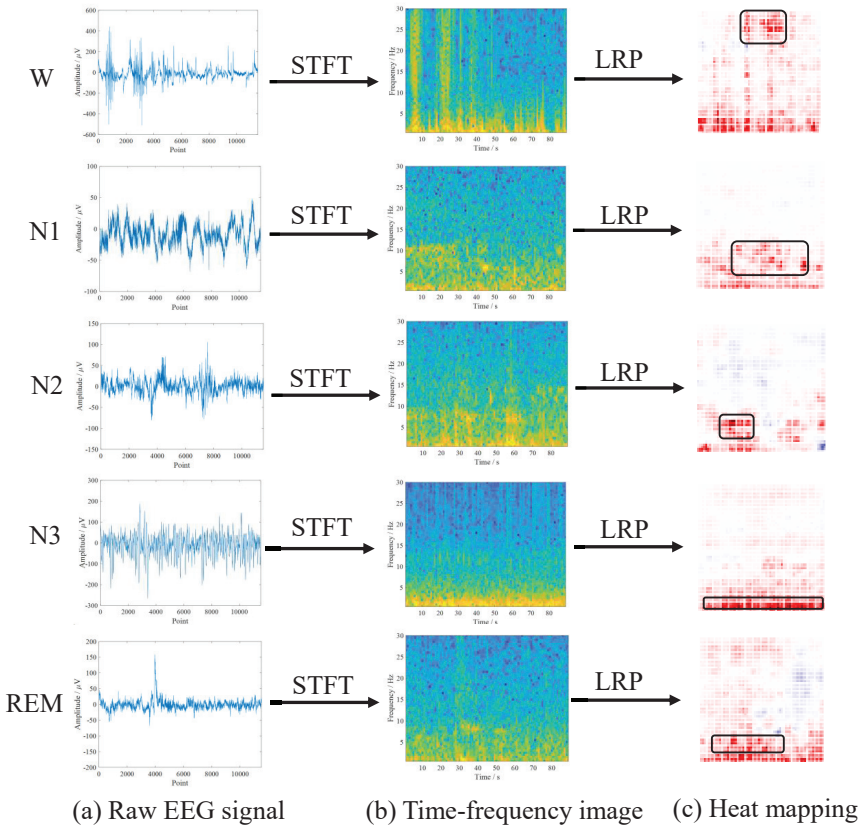


FIGURE 18 The explainability results for each sleep stage implementing the LRP method.

Methods

We propose an explainable scheme to investigate the internal relationship between the applied model's input and prediction, shown in Figure 16. We first get time-frequency images with information on the EEG patterns to attain the model input using the STFT. We also present a novel CNN-based model for automatic sleep scoring that assembles with Multi-Scale CNN (MSCNN) and Residual Squeeze-and-Excitation block (R-SE) (see Figure 17). Finally, layer-wise relevance propagation (LRP), a conservative relevance redistribution approach, is used to recognize effective pixels (corresponding to frequency patterns) in the time-frequency image input that contribute the most to the final layer and benefit most from it. We seek to see whether the proposed model can correctly identify specific EEG patterns in each sleep stage while making a final decision.

Results

We first conduct the ablation study to validate the efficiency of each module (MSCNN and R-SE blocks) on three experimental datasets (CCSHS, Sleep-EDF, and Sleep-EDF-V1). Although MSCNN + R-SE is unable to enhance the performance on the Sleep-EDF dataset compared to MSCNN alone, with the addition

of R-SE block, *ACC* is improved by 0.7% on Sleep-EDF-V1 and 0.3% on CCSHS. Likewise, the proposed MSSENet improves *K* on the Sleep-EDF-V1 and CCSHS datasets by 0.9% and 0.5%, respectively.

In Figure 18, we show the explainability results for each sleep stage using the LRP approach. Part (a) depicts the raw EEG signals of five sleep phases, whereas part (b) depicts the corresponding time-frequency images generated by the short-time Fourier Transform, which comprises the EEG pattern information. The part (c) depicts the heat mapping of each sleep state using the LRP. We can see that stage W's heat mapping has a high importance in a high-frequency band (i.e., Beta waves), which is compatible with stage W's major EEG patterns. The slowest EEG rhythms (i.e., Delta waves), indicative of profound sleep in stage N3, also play a significant role in stage W prediction. The LRP-based results are in accordance with the recommendations for the sleep score manual.

Conclusion and discussion

In *Article V*, we provide a new, understandable solution for automatic sleep scoring. This solution is assembled in our proposed MSSENet model. Our MSSENet could outperform other state-of-the-art methods using the same PSG dataset. Furthermore, the model prediction could be visually interpreted using the LRP-based heat mapping in Figure 18. The EEG patterns (Delta, Theta, Alpha, and Beta waves) of different sleep stages exhibit great relevance to the correct prediction, which aligns with the sleep scoring standard.

Author contributions in *Article V*

Dongdong Zhou: Conceptualization, Methodology, Software, Writing original draft. **Qu Xi**: Writing – review & editing, Supervision. **Jiacheng Zhang**: Writing – review & editing. **Lei Wu**: Writing – review & editing. **Lauri Kettunen**: Writing – review & editing, Supervision. **Zheng Chang**: Writing – review & editing, Supervision. **Hongming Xu**: Writing – review & editing. **Fengyu Cong**: Writing – review & editing, Supervision.

4 CONCLUSION AND DISCUSSION

The overview of this thesis is presented in this chapter initially. The limitations of all studies are then covered. Finally, several prospective research directions are discussed.

4.1 Summary of the thesis

This dissertation focuses on deep learning-based algorithms for automatic sleep stage classification using single-channel EEG. *Articles I and II* propose two novel CNN-based sleep scoring frameworks (SCNet and LSNet) based on raw EEG signals and spectrograms, respectively. *Articles III* and *IV* aim to investigate the solutions for the class imbalance problem in automatic sleep scoring tasks. *Article V* presents an interpretable sleep stage classification system with layer-wise relevance propagation for the model explanation.

Specifically, *Article I* develops an end-to-end 1D-CNN-based model (SCNet) that combines the capability of multi-scale feature learning and classification. We then construct a contextual epoch employing the many-to-one scheme as the model input. The proposed SCNet with raw single-channel EEG could achieve promising performance on public PSG datasets. In *Article II*, we first obtain spectrograms from raw EEG signals using the STFT method. The spectrograms are then fed into a lightweight yet effective 2D-CNN-based framework (LSNet). Our LSNet model could realize rapid sleep scoring with spectrograms and attain comparable performance compared with other state-of-the-art methods with much fewer model parameters. To tackle the class imbalance problem, *Article III* presents two balancing methods to increase the minority class quantity based on the time-frequency images of raw EEG signals. While in *Article IV*, we introduce the data augmentation method with different intensities and times GWN and a GAN model. In addition, we explore the solution to balancing the network connection while unchanging the dataset samples. Proposed balancing methods in *Articles III* and *IV* can enhance the overall and N1 accuracies to differ-

ent extents. We provide an LRP-based explainable model with CNNs (MSSENet) to demonstrate the contribution of specific EEG patterns in each sleep stage to the model prediction visually in *Article V*. The MSSENet model could gain promising performance on three PSG datasets, and the LRP-based results show that EEG patterns of certain sleep stages (i.e., Delta, Theta, Alpha, and Beta waves) indicate strong relevance to the correct prediction, which is compatible with the sleep scoring criteria.

In conclusion, this thesis systematically investigates DL-b methods for automatic sleep stage classification and solutions for the CIP and model explanation in automatic sleep scoring with single-channel EEG. In a larger sense, we expect sleep monitoring in clinical or daily care to be made more accessible by the proposed methods in this thesis.

4.2 Research limitations and future directions

Despite the fact that our results are favorable, there are still several limitations in this study. The first one is the restriction of the datasets employed. In this dissertation, we validate our proposed methods on four public PSG datasets with different age groups. However, the different data attributes may influence the proposed approaches' efficiency. Subjects from these datasets are primarily healthy individuals. The performance of the proposed models should be further validated on patients with sleep disorders to enhance the model generalization ability. In addition, larger and more high-quality clinic PSG datasets are required for potential clinic use of presented models.

Secondly, we only investigate the performance of single-channel EEG. Although the single-channel EEG scheme can effectively reduce the computational cost and simplify the data acquisition procedure, the beneficial contribution of channel increase to the model performance has not yet been investigated (Yan et al., 2019). For example, three main features of REM sleep need to be captured by EEG, EOG, and EMG collectively (Berry et al., 2012). The single-channel EEG-based model may hinder the recognition accuracy of REM.

Thirdly, it is challenging to perform well when training a model on dataset A but testing it on another dataset B, which has different data attributes. In *Article I*, we train our SCNet on the CCSHS dataset and then test the trained SCNet on the Sleep-EDF dataset, the accuracy is 65.9%, and the proposed SCNet obtains an accuracy of 70.2% using the inverse training strategy. By doing this, the overall accuracy of our proposed SCNet decreases drastically, a typical flaw of most deep learning methods that must be further overcome.

Fourthly, the proposed GAN model in *Article IV* for balancing the dataset samples could improve the overall accuracy. However, the N1 accuracy exhibits a slight decrease in experimental PSG datasets. More balancing methods customized explicitly for EEG signals could be further explored to enhance the overall and N1 accuracy simultaneously. Last but not least, the proposed interpretable

scheme in *Article V* fails to detect other EEG patterns associated with specific sleep stages (e.g., K-complexes in stage N2, Saw tooth waves in stage REM), which can improve sleep scoring performance. Additionally, the multi-modal PSG data would enhance the model interpretation of DL-b methods for automatic sleep scoring.

Based on the study's limitations, as stated above, we briefly summarize the following future directions in automatic sleep stage classification tasks:

1. More clinic PSG datasets from hospitals, especially individuals with sleep disorders (e.g., apnea, insomnia, narcolepsy), are valuable future practical applications. Collaboration with specialized hospitals or sleep laboratories is essential for achieving this aim.
2. The impact of different numbers of channels on model performance needs to be further explored using a channel selection strategy. The model performance and computational cost with different channels input are worth a comprehensive comparison.
3. This thesis focus on the two most popular deep learning methods: CNN and LSTM (as well as the combination of CNN and LSTM). More deep learning approaches (e.g., Transformer (Vaswani et al., 2017), Graph neural network (Scarselli et al., 2008), Spiking neural network (Ghosh-Dastidar and Adeli, 2009)) should be investigated and discussed in terms of the training efficiency and model performance in future studies.
4. More efficient dataset sample balancing methods should be examined and tailored specifically for EEG signals, such as the variational auto-Encoding network (Kingma and Welling, 2013). The accuracy enhancement of the minority class should not sacrifice the overall accuracy. The continuous exploration of CIP is one of the main directions in my future work.
5. Future studies will also focus on multi-modal data for improved model interpretability in automatic sleep scoring tasks. In addition, it is also valuable to determine the role of other EEG patterns (e.g., K-complexes, Sleep spindles, Vertex waves, Saw tooth waves) in model prediction.

YHTEENVETO (SUMMARY IN FINNISH)

Tämä opinnäytetyö keskittyy syväoppimiseen perustuviin algoritmeihin automaattisessa univaiheen luokittelussa käyttämällä yksikanavaista EEG:tä. Artikkeleissa I ja II ehdotamme kahta uutta konvoluutiohermoverkkopohjaista unen pisteytyskehystä SCNet ja LSNet, jotka perustuvat raaka-EEG-signaaleihin ja spektrogrammeihin. Artikkeleissa III ja IV pyritään tutkimaan ratkaisuja luokkaepätasapaino-ongelmaan automaattisissa unen pisteytystehtävissä. Artikkelissa V esitellään tulkittavissa oleva unen pisteytysjärjestelmä, jossa on kerroksittainen relevanssin eteneminen mallin selittämiseksi.

Artikkelissa I kehitetään yksiulotteinen konvoluutiohermoverkkopohjainen malli SCNet, joka yhdistää ominaisuuksien oppimisen ja luokittelun. Rakennamme sitten kontekstuaalisen kierroksen (engl., epoch) käyttämällä monesta-yhteen -kaaviota mallin syötteenä. Ehdotettu SCNet, jossa on raaka yksikanavainen EEG, voisi saavuttaa lupaavan suorituskyvyn julkisissa polysomnografiatietosarjoissa. Artikkelissa II saamme ensin spektrogrammit raaka-EEG-signaaleista STFT-menetelmällä. Spektrogrammit syötetään sitten kevyeen mutta tehokkaaseen kaksiulotteiseen konvoluutiohermoverkkopohjaiseen kehykseen (LSNet). LSNet- verkkomme ei pystynyt ainoastaan toteuttamaan nopeaa unipisteytystä spektrogrammeilla, vaan myös saavuttamaan vertailukelpoisen suorituskyvyn vähemmillä malliparametreilla muihin alan vakiintuneisiin menetelmiin verrattuna. Luokkaepätasapaino-ongelman ratkaisemiseksi artikkelissa III esitetään kaksi tasapainotusmenetelmää vähemmistöluokan määrän lisäämiseksi raaka-EEG-signaalien aika-taajuuskuvien perusteella. Artikkelissa IV esittelemme datan lisäysmenetelmiä, jotka käyttävät Gaussin valkoisen kohinan lisäyksen eri intensiteettejä ja aikoja sekä generatiivisen adversariaalisen verkon mallia. Lisäksi tutkimme ratkaisua verkkoyhteyden tasapainottamiseen muuttamatta tietojoukon näytteitä. Artikkeleissa III ja IV ehdotetut tasapainotusmenetelmät voivat parantaa yleistä ja univaiheen N1 tarkkuutta. Artikkelissa V esittelemme konvoluutiohermoverkkopohjaisen relevanssin kerroksittaiseen etenemiseen perustuvan selitettävän mallin MSSENet, joka osoittaa visuaalisesti kunkin univaiheen tiettyjen EEG-kuvioiden vaikutuksen mallin ennustamiseen. MSSENet-malli voisi saada lupaavan suorituskyvyn kolmella polysomnografiatietojoukolla. Relevanssin kerroksittaiseen etenemiseen perustuvat tulokset osoittavat, että tiettyjen univaiheiden EEG-kuvioilla (eli delta-, theta-, alfa- ja beeta-aallot) on suuri merkitys oikean ennustamisen kannalta, mikä on yhteensopivaa unen pisteytyskriteerien kanssa.

Lopuksi tässä opinnäytetyössä tutkitaan systemaattisesti menetelmiä automaattisessa univaiheen luokittelussa sekä etsitään ratkaisuja luokkaepätasapaino-ongelmaan ja automaattisen unipisteytysmallin selittämiseen käyttäen yksikanavaista EEG:tä. Laajemmassa mielessä odotamme tässä opinnäytetyössä ehdotettujen menetelmien helpottavan unen seurantaa kliinisessä tai päivittäisessä hoidossa.

REFERENCES

Alickovic, E. and Subasi, A. (2018). Ensemble SVM method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement*, 67(6):1258–1265.

Allen, D. P. (2009). A frequency domain hampel filter for blind rejection of sinusoidal interference from electromyograms. *Journal of Neuroscience Methods*, 177(2):303–310.

Altevogt, B. M. and Colten, H. R., editors (2006). *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press.

Andreotti, F., Phan, H., Cooray, N., Lo, C., Hu, M. T., and De Vos, M. (2018). Multichannel sleep stage classification and transfer learning using convolutional neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 171–174. IEEE.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One*, 10(7):e0130140.

Bai, J., Lian, S., Liu, Z., Wang, K., and Liu, D. (2018). Deep learning based robot for automatically picking up garbage on the grass. *IEEE Transactions on Consumer Electronics*, 64(3):382–389.

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., et al. (2012). Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine*, 8(5):597–619.

Bhaskar, S., Hemavathy, D., and Prasad, S. (2016). Prevalence of chronic insomnia in adult patients and its correlation with medical comorbidities. *Journal of Family Medicine and Primary Care*, 5(4):780.

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*, 140:77–91.

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. (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):758–769.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

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

Cong, F., Ristaniemi, T., and Lyytinen, H. (2015). *Advanced signal processing on brain event-related potentials: filtering ERPs in time, frequency and space domains sequentially and simultaneously*, volume 13. World Scientific.

Cook, J. D., Prairie, M. L., and Plante, D. T. (2017). Utility of the fitbit flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, 217:299–305.

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

de Souza, L., Benedito-Silva, A. A., Pires, M. L. N., Poyares, D., Tufik, S., and Calil, H. M. (2003). Further validation of actigraphy for sleep studies. *Sleep*, 26(1):81–85.

Decat, N., Walter, J., Koh, Z. H., Sribanditmongkol, P., Fulcher, B. D., Windt, J. M., Andrillon, T., and Tsuchiya, N. (2022). Beyond traditional sleep scoring: Massive feature extraction and data-driven clustering of sleep time series. *Sleep Medicine*, 98:39–52.

Derbin, M., McKenna, L., Chin, D., Coffman, B., and Bloch-Salisbury, E. (2022). Actigraphy: Metrics reveal it is not a valid tool for determining sleep in neonates. *Journal of Sleep Research*, 31(1):e13444.

Ding, Y., Hua, L., and Li, S. (2022). Research on computer vision enhancement in intelligent robot based on machine learning and deep learning. *Neural Computing and Applications*, 34(4):2623–2635.

Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P. M., and Guo, Y. (2017). Mixed neural network approach for temporal sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):324–333.

Duhamel, P. and Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art. *Signal Processing*, 19(4):259–299.

Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control*, 14:200–230.

Edwards, B. A., O'Driscoll, D. M., Ali, A., Jordan, A. S., Trinder, J., and Malhotra, A. (2010). Aging and sleep: physiology and pathophysiology. In *Seminars in Respiratory and Critical Care Medicine*, volume 31, pages 618–633.

Efe, E. and Ozsen, S. (2023). CoSleepNet: Automated sleep staging using a hybrid CNN-LSTM network on imbalanced EEG-EOG datasets. *Biomedical Signal Processing and Control*, 80:104299.

Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C.-K., Li, X., and Guan, C. (2021). An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818.

Ellis, C. A., Zhang, R., Carbajal, D. A., Miller, R. L., Calhoun, V. D., and Wang, M. D. (2021). Explainable sleep stage classification with multimodal electrophysiology time-series. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2363–2366. IEEE.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883.

Fan, J., Sun, C., Chen, C., Jiang, X., Liu, X., Zhao, X., Meng, L., Dai, C., and Chen, W. (2020). EEG data augmentation: towards class imbalance problem in sleep staging tasks. *Journal of Neural Engineering*, 17(5):056017.

Fatimah, B., Singhal, A., and Singh, P. (2022). A multi-modal assessment of sleep stages using adaptive Fourier decomposition and machine learning. *Computers in Biology and Medicine*, 148:105877.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Fu, M., Wang, Y., Chen, Z., Li, J., Xu, F., Liu, X., and Hou, F. (2021). Deep learning in automatic sleep staging with a single channel electroencephalography. *Frontiers in Physiology*, 12:628502.

Ghosh-Dastidar, S. and Adeli, H. (2009). Spiking neural networks. *International Journal of Neural Systems*, 19(04):295–308.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Goshtasbi, N., Boostani, R., and Sanei, S. (2022). Sleep fcn: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2088–2096.

Guillot, A. and Thorey, V. (2021). RobustSleepNet: Transfer learning for automated sleep staging at scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1441–1451.

Güneş, S., Polat, K., and Yosunkaya, Ş. (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, 37(12):7922–7928.

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*, 271:107–118.

He, Z., Du, L., Wang, P., Xia, P., Liu, Z., Song, Y., Chen, X., and Fang, Z. (2022). Single-channel EEG sleep staging based on data augmentation and cross-subject discrepancy alleviation. *Computers in Biology and Medicine*, 149:106044.

Heeger, D. J. and Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, 3(2):142–151.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, 833-840:106044.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, J., Ren, L., Zhou, X., and Yan, K. (2022). An improved neural network based on senet for sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*.

Humayun, A. I., Sushmit, A. S., Hasan, T., and Bhuiyan, M. I. H. (2019). End-to-end sleep staging with raw single channel EEG using deep residual convnets. In *2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE.

Iranzo, A., Molinuevo, J. L., Santamaría, J., Serradell, M., Martí, M. J., Valldeoriola, F., and Tolosa, E. (2006). Rapid-eye-movement sleep behaviour disorder as an early marker for a neurodegenerative disorder: a descriptive study. *The Lancet Neurology*, 5(7):572–577.

Jia, Z., Cai, X., Zheng, G., Wang, J., and Lin, Y. (2020). SleepPrintNet: a multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Transactions on Artificial Intelligence*, 1(3):248–257.

Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., Zhou, Y., and Li-wei, H. L. (2021). Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986.

Jiang, D., Yu, M., and Yuanyuan, W. (2019). Sleep stage classification using covariance features of multi-channel physiological signals on Riemannian manifolds. *Computer Methods and Programs in Biomedicine*, 178:19–30.

Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M., Iragui, V., and Sejnowski, T. J. (1997). Extended ICA removes artifacts from electroencephalographic recordings. *Advances in Neural Information Processing Systems*, 10.

Karimzadeh, F., Boostani, R., Seraj, E., and Sameni, R. (2017). A distributed classification procedure for automatic sleep stage scoring based on instantaneous electroencephalogram phase and envelope features. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):362–370.

Kay, D. B. and Dzierzewski, J. M. (2015). Sleep in the context of healthy aging and psychiatric syndromes. *Sleep Medicine Clinics*, 10(1):11–15.

Kayikcioglu, T., Maleki, M., and Eroglu, K. (2015). Fast and accurate pls-based classification of EEG sleep using single channel data. *Expert Systems with Applications*, 42(21):7825–7830.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A., and Oberye, J. J. (2000). Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koley, B. and Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*, 42(12):1186–1195.

Korkalainen, H., Aakko, J., Duce, B., Kainulainen, S., Leino, A., Nikkonen, S., Afara, I. O., Myllymaa, S., Töyräs, J., and Leppänen, T. (2020). Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep*, 43(11):zsaa098.

Kouchaki, S., Sanei, S., Arbon, E. L., and Dijk, D.-J. (2014). Tensor based singular spectrum analysis for automatic scoring of sleep EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(1):1–9.

Krishnan, V. and Collop, N. A. (2006). Gender differences in sleep disorders. *Current Opinion in Pulmonary Medicine*, 12(6):383–389.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kwon, H. B., Choi, S. H., Lee, D., Son, D., Yoon, H., Lee, M. H., Lee, Y. J., and Park, K. S. (2021). Attention-based LSTM for non-contact sleep stage classification using IR-UWB radar. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3844–3853.

Lajnef, T., Chaibi, S., Ruby, P., Aguera, P.-E., Eichenlaub, J.-B., Samet, M., Kachouri, A., and Jerbi, K. (2015). Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of Neuroscience Methods*, 250:94–105.

Lau, M. M. and Lim, K. H. (2018). Review of adaptive activation function in deep neural network. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 686–690. IEEE.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

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):375–385.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Luyster, F. S., Strollo, P. J., Zee, P. C., and Walsh, J. K. (2012). Sleep: a health imperative. *Sleep*, 35(6):727–734.

Malafeev, A., Laptev, D., Bauer, S., Omlin, X., Wierzbicka, A., Wichniak, A., Jernajczyk, W., Riener, R., Buhmann, J., and Achermann, P. (2018). Automatic human sleep stage scoring using deep neural networks. *Frontiers in Neuroscience*, 12:781.

Memar, P. and Faradji, F. (2017). A novel multi-class EEG-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1):84–95.

Mennes, M., Wouters, H., Vanrumste, B., Lagae, L., and Stiers, P. (2010). Validation of ICA as a tool to remove eye movement artifacts from EEG/ERP. *Psychophysiology*, 47(6):1142–1150.

Mesarwi, O., Polak, J., Jun, J., and Polotsky, V. Y. (2013). Sleep disorders and the development of insulin resistance and obesity. *Endocrinology and Metabolism Clinics*, 42(3):617–634.

Morin, C. M., Vézina-Im, L.-A., Ivers, H., Micoulaud-Franchi, J.-A., Philip, P., Lamy, M., and Savard, J. (2022). Prevalent, incident, and persistent insomnia in a population-based cohort tested before (2018) and during the first-wave of covid-19 pandemic (2020). *Sleep*, 45(1):zsab258.

Mousavi, S., Afghah, F., and Acharya, U. R. (2019). SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS One*, 14(5):e0216456.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the International Conference on Machine Learning*.

Nakamura, T., Davies, H. J., and Mandic, D. P. (2019). Scalable automatic sleep staging in the era of big data. In *2019 41st Annual International Conference of*

the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2265–2268. IEEE.

Neng, W., Lu, J., and Xu, L. (2021). CCRRSleepNet: A hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel EEG. *Brain Sciences*, 11(4):456.

Oikonomou, G. and Prober, D. A. (2017). Attacking sleep from a new angle: contributions from zebrafish. *Current Opinion in Neurobiology*, 44:80–88.

O'reilly, C., Gosselin, N., Carrier, J., and Nielsen, T. (2014). Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, 23(6):628–635.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Paolanti, M., Romeo, L., Martini, M., Mancini, A., Frontoni, E., and Zingaretti, P. (2019). Robotic retail surveying by deep learning visual and textual data. *Robotics and Autonomous Systems*, 118:179–188.

Patanaik, A., Ong, J. L., Gooley, J. J., Ancoli-Israel, S., and Chee, M. W. (2018). An end-to-end framework for real-time automatic sleep stage classification. *Sleep*, 41(5):zsy041.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. (2018a). DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification. In *2018 40th Annual International Conference of the IEEE engineering in Medicine and Biology Society (EMBC)*, pages 453–456. IEEE.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. (2018b). Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296.

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*, 27(3):400–410.

Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., and De Vos, M. (2021). XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915.

Phan, H., Do, Q., Do, T.-L., and Vu, D.-L. (2013). Metric learning for automatic sleep stage classification. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5025–5028. IEEE.

Qu, W., Wang, Z., Hong, H., Chi, Z., Feng, D. D., Grunstein, R., and Gordon, C. (2020). A residual based attention model for EEG based sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2833–2843.

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., et al. (1997). The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085.

Radüntz, T., Scouten, J., Hochmuth, O., and Meffert, B. (2015). EEG artifact elimination by extraction of ICA-component features using image processing algorithms. *Journal of Neuroscience Methods*, 243:84–93.

Rechtschaffen, A. (1969). A manual of standardized terminology and scoring system for sleep stages of human subjects. *Electroencephalography and Clinical Neurophysiology*, 26(6):644.

Redmond, S. J. and Heneghan, C. (2006). Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *IEEE Transactions on Biomedical Engineering*, 53(3):485–496.

Rosen, C. L., Larkin, E. K., Kirchner, H. L., Emancipator, J. L., Bivins, S. F., Surovec, S. A., Martin, R. J., and Redline, S. (2003). Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *The Journal of Pediatrics*, 142(4):383–389.

Satapathy, S. K., Bhoi, A. K., Loganathan, D., Khandelwal, B., and Barsocchi, P. (2021). Machine learning with ensemble stacking model for automated sleep staging using dual-channel EEG signal. *Biomedical Signal Processing and Control*, 69:102898.

Sawangjit, A., Harkotte, M., Oyanedel, C. N., Niethard, N., Born, J., and Inostroza, M. (2022). Two distinct ways to form long-term object recognition memory during sleep and wakefulness. *Proceedings of the National Academy of Sciences*, 119(34):e2203165119.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Sekkal, R. N., Bereksi-Reguig, F., Ruiz-Fernandez, D., Dib, N., and Sekkal, S. (2022). Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77:103751.

Seo, H., Back, S., Lee, S., Park, D., Kim, T., and Lee, K. (2020). Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomedical Signal Processing and Control*, 61:102037.

Sharma, M., Goyal, D., Achuth, P., and Acharya, U. R. (2018). An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank. *Computers in Biology and Medicine*, 98:58–75.

Shensa, M. J. et al. (1992). The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on Signal Processing*, 40(10):2464–2482.

Shuyuan, X., Bei, W., Jian, Z., Qunfeng, Z., Junzhong, Z., and Nakamura, M. (2015). Notice of removal: An improved k-means clustering algorithm for sleep stages classification. In *2015 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 1222–1227. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

SM, I. N., Zhu, X., Chen, Y., and Chen, W. (2019). Sleep stage classification based on EEG, EOG, and CNN-GRU deep learning model. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–7. IEEE.

Sors, A., Bonnet, S., Mirek, S., Vercueil, L., and Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114.

Sousa, T., Cruz, A., Khalighi, S., Pires, G., and Nunes, U. (2015). A two-step automatic sleep stage classification method with dubious range detection. *Computers in Biology and Medicine*, 59:42–53.

Sun, C., Fan, J., Chen, C., Li, W., and Chen, W. (2019). A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation. *IEEE Access*, 7:109386–109397.

Sun, H., Ganglberger, W., Panneerselvam, E., Leone, M. J., Quadri, S. A., Goparaju, B., Tesh, R. A., Akeju, O., Thomas, R. J., and Westover, M. B. (2020). Sleep staging from electrocardiography and respiration with deep learning. *Sleep*, 43(7):zsz306.

Supratak, A., Dong, H., Wu, C., and Guo, Y. (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.

Supratak, A. and Guo, Y. (2020). TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 641–644. IEEE.

Šušmáková, K. and Krakovská, A. (2008). Discrimination ability of individual measures used in sleep stages classification. *Artificial Intelligence in Medicine*, 44(3):261–277.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Ţarălungă, D. D., Ungureanu, G. M., Hurezeanu, B., Strungaru, R., Gussi, I., and Wolf, W. (2014). Abdominal signals processing: Power line interference removing by applying notch filters. In *2014 International Conference and Exposition on Electrical and Power Engineering (EPE)*, pages 158–161. IEEE.

Tsinalis, O., Matthews, P. M., and Guo, Y. (2016). Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of Biomedical Engineering*, 44(5):1587–1597.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Van Leeuwen, K., Sun, H., Tabaeizadeh, M., Struck, A., Van Putten, M., and Westover, M. (2019). Detecting abnormal electroencephalograms using deep convolutional networks. *Clinical Neurophysiology*, 130(1):77–84.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wafford, K. A. and Ebert, B. (2008). Emerging anti-insomnia drugs: tackling sleeplessness and the quality of wake time. *Nature Reviews Drug Discovery*, 7(6):530–540.

Wang, H., Lu, C., Zhang, Q., Hu, Z., Yuan, X., Zhang, P., and Liu, W. (2022). A novel sleep staging network based on multi-scale dual attention. *Biomedical Signal Processing and Control*, 74:103486.

Wang, Y. and Wu, D. (2018). Deep learning for sleep stage classification. In *2018 Chinese Automation Congress (CAC)*, pages 3833–3838. IEEE.

Xiao, M., Yan, H., Song, J., Yang, Y., and Yang, X. (2013). Sleep stages classification based on heart rate variability and random forest. *Biomedical Signal Processing and Control*, 8(6):624–633.

Xu, Q., Peng, J., Shen, J., Tang, H., and Pan, G. (2020). Deep CovDenseSNN: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment. *Neural Networks*, 121:512–519.

Xu, Q., Shen, J., Ran, X., Tang, H., Pan, G., and Liu, J. K. (2021). Robust transcoding sensory information with neural spikes. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):1935–1946.

Xu, Q., Zhou, D., Wang, J., Shen, J., Kettunen, L., and Cong, F. (2022). Convolutional neural network based sleep stage classification with class imbalance. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Xu, S. S., Mak, M.-W., and Cheung, C.-C. (2018). Towards end-to-end ECG classification with raw signal extraction and deep neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1574–1584.

Yan, R., Li, F., Zhou, D., Ristaniemi, T., and Cong, F. (2021a). A deep learning model for automatic sleep scoring using multimodality time series. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1090–1094. IEEE.

Yan, R., Li, F., Zhou, D. D., Ristaniemi, T., and Cong, F. (2021b). Automatic sleep scoring: A deep learning architecture for multi-modality time series. *Journal of Neuroscience Methods*, 348:108971.

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*, 49:14–23.

Yang, B., Zhu, X., Liu, Y., and Liu, H. (2021). A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. *Biomedical Signal Processing and Control*, 68:102581.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

You, Y., Zhong, X., Liu, G., and Yang, Z. (2022). Automatic sleep stage classification: A light and efficient deep neural network model based on time, frequency and fractional Fourier transform domain features. *Artificial Intelligence in Medicine*, 127:102279.

Yu, Z., Guo, S., Deng, F., Yan, Q., Huang, K., Liu, J. K., and Chen, F. (2018). Emergent inference of hidden Markov models in spiking neural networks through winner-take-all. *IEEE Transactions on Cybernetics*, 50(3):1347–1354.

Zarei, A., Beheshti, H., and Asl, B. M. (2022). Detection of sleep apnea using deep neural networks and single-lead ECG signals. *Biomedical Signal Processing and Control*, 71:103125.

Zeng, Y., Zhang, T., and Xu, B. (2017). Improving multi-layer spiking neural networks by incorporating brain-inspired rules. *Science China Information Sciences*, 60(5):1–11.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018). The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358.

Zhang, H., Wang, X., Li, H., Mehendale, S., and Guan, Y. (2022a). Auto-annotating sleep stages based on polysomnographic data. *Patterns*, 3(1):100371.

Zhang, J. and Wu, Y. (2017). A new method for automatic sleep stage classification. *IEEE Transactions on Biomedical Circuits and Systems*, 11(5):1097–1110.

Zhang, J. and Wu, Y. (2021). Competition convolutional neural network for sleep stage classification. *Biomedical Signal Processing and Control*, 64:102318.

Zhang, L., Fabbri, D., Upender, R., and Kent, D. (2019). Automated sleep stage scoring of the sleep heart health study using deep neural networks. *Sleep*, 42(11):zsz159.

Zhang, T., Ye, W., Yang, B., Zhang, L., Ren, X., Liu, D., Sun, J., Zhang, S., Zhang, H., and Zhao, W. (2022b). Frequency-aware contrastive learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11712–11720.

Zhao, R., Xia, Y., and Zhang, Y. (2021). Unsupervised sleep staging system based on domain adaptation. *Biomedical Signal Processing and Control*, 69:102937.

Zhou, D., Wang, J., Hu, G., Zhang, J., Li, F., Yan, R., Kettunen, L., Chang, Z., Xu, Q., and Cong, F. (2022a). SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG. *Biomedical Signal Processing and Control*, 75:103592.

Zhou, D., Xu, Q., Wang, J., Xu, H., Kettunen, L., Chang, Z., and Cong, F. (2022b). Alleviating class imbalance problem in automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12.

Zhou, D., Xu, Q., Wang, J., Zhang, J., Hu, G., Kettunen, L., Chang, Z., and Cong, F. (2021). LightSleepNet: A lightweight deep model for rapid sleep stage classification with spectrograms. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 43–46. IEEE.

Zhou, J., Wang, G., Liu, J., Wu, D., Xu, W., Wang, Z., Ye, J., Xia, M., Hu, Y., and Tian, Y. (2020). Automatic sleep stage classification with single channel EEG signal based on two-layer stacked ensemble model. *IEEE Access*, 8:57283–57297.

Zhu, G., Li, Y., and Wen, P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1813–1821.

Zhu, T., Luo, W., and Yu, F. (2020). Convolution-and attention-based neural network for automated sleep stage classification. *International Journal of Environmental Research and Public Health*, 17(11):4152.

Zuo, X., Zhang, C., Cong, F., Zhao, J., and Hämäläinen, T. (2022). Driver distraction detection using bidirectional long short-term network based on multiscale entropy of EEG. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19309–19322.

ORIGINAL PAPERS

PI

SINGLECHANNELNET: A MODEL FOR AUTOMATIC SLEEP STAGE CLASSIFICATION WITH RAW SINGLE-CHANNEL EEG

by

Dongdong Zhou, JianWang, Guoqiang Hu, Jiacheng Zhang, Fan Li, Rui Yan, Lauri Kettunen, Zheng Chang, Qi Xu, and Fengyu Cong 2022

Biomedical Signal Processing and Control, 75, 103592,
https://doi.org/10.1016/j.bspc.2022.103592

Biomedical Signal Processing and Control 75 (2022) 103592

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG

Dongdong Zhou [a,b], Jian Wang [a,b], Guoqiang Hu [a], Jiacheng Zhang [c], Fan Li [a], Rui Yan [a,b], Lauri Kettunen [b], Zheng Chang [b], Qi Xu [d,*], Fengyu Cong [a,b,d,e,*]

[a] School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, Dalian, China
[b] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
[c] School of Information and Communication Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, Dalian, China
[d] School of Artifical Intelligence, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, Dalian, China
[e] Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province, Dalian University of Technology, Dalian, China

ARTICLE INFO

Keywords:
Sleep stage classification
Raw single-channel EEG
Contextual input
Convolutional neural network

ABSTRACT

In diagnosing sleep disorders, sleep stage classification is a very essential yet time-consuming process. Various existing state-of-the-art approaches rely on hand-crafted features and multi-modality polysomnography (PSG) data, where prior knowledge is compulsory and high computation cost can be expected. Besides, it is a big challenge to handle the task with raw single-channel electroencephalogram (EEG). To overcome these shortcomings, this paper proposes an end-to-end framework with a deep neural network, namely SingleChannelNet, for automatic sleep stage classification based on raw single-channel EEG. The proposed model utilizes a 90s epoch as the textual input and employs two multi-convolution (MC) blocks and several max-average pooling (M-Apooling) layers to learn different scales of feature representations. To demonstrate the efficiency of the proposed model, we evaluate our model using different raw single-channel EEGs (C4/A1 and Fpz-Cz) on two public PSG datasets (Cleveland children's sleep and health study: CCSHS and Sleep-EDF database expanded: Sleep-EDF). Experimental results show that the proposed architecture can achieve better overall accuracy and Cohen's kappa (CCSHS: 90.2%–86.5%, Sleep-EDF: 86.1%–80.5%) compared with state-of-the-art approaches. Additionally, the proposed model can learn features automatically for sleep stage classification using different single-channel EEGs with distinct sampling rates and without using any hand-engineered features.

1. Introduction

Sleep occupies one-third of human life, which plays a vitally important role in restoring body and mind [1]. Whereas roughly 33% of the population in the world suffers from insomnia disorder [2]. Correctly identifying sleep stage using whole-night PSG data is essential to diagnose and treat sleep-related disorders [3–6]. The PSG recordings comprise of the EEG, electrocardiogram (ECG), electrooculogram (EOG), electromyogram (EMG) and other respiration signals [7].

According to the guidelines of the Rechtschaffen and Kales (R&K) [8] or American Academy of Sleep Medicine (AASM) [9], the PSG data should be first segmented into 30s epochs typically, then these sequential epochs are defined as different stages. Some sleep-related disorders have particular sleep structure, it is therefore beneficial to diagnose them with accurate sleep stage classification. Traditionally, the sleep stage classification task is conducted by experts manually following the R&K or AASM rule which is often time-consuming, labor-intensive and prone to subjective mistakes [6]. Hence, there is an urgent need for automatic sleep stage classification approach to assist the clinician's work and achieve reliable results.

Some methods based on machine learning have been proposed to identify the sleep stage. These approaches generally extract either time-domain features [3,10,11] or frequency-domain features [12–16] from the PSG signals and these pre-extracted features are then fed into the conventional classifier, such as support vector machine (SVM) [4,14,17,18], *k*-nearest neighbors (KNN) [16,19,20], random forest [21–24] and so on. The performance tremendously relies on the categories and the number of features, which are extracted based on the

* Corresponding authors at: School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, Dalian, China (F. Cong).
E-mail addresses: xuqi@dlut.edu.cn (Q. Xu), cong@dlut.edu.cn (F. Cong).

https://doi.org/10.1016/j.bspc.2022.103592
Received 17 September 2021; Received in revised form 27 December 2021; Accepted 20 February 2022
Available online 2 March 2022

characteristics of experimental datasets. Therefore, these approaches may not be robust enough to be generalized to different datasets because of the distinct properties between datasets.

In recent years, the deep networks show great capacity for automatic features learning from data, and it can avoid the reliance on hand-engineered features. Meanwhile, a series of deep learning methods are applied to the sleep stage classification task. Here, we categorize these approaches into multi-channel [6,25–29] versus single-channel schemes [30–35] based on the number of input channels. Following the multi-channel scheme, Phan et al. [6] first transformed the raw signals into the time-frequency image through the short-time Fourier transform as the input of the proposed convolutional neural network (CNN). The overall accuracy achieved was equal to 82.3%, in which there is room for improvement. Besides, the time-frequency image relies much on many preprocessing steps, it would be time-consuming and in need of prior knowledge of signal processing. Aiming at this, Chambon et al. [27] proposed a novel network architecture of low computational cost adopting multivariate and multimodal time series from EEG, EMG and EOG, but the classification performance is not good enough with the accuracy of 80% compared to state-of-the-art methodologies [7,30]. One important reason is that the convolutional layers with fixed filter size were stacked sequentially, which can not learn multiscale features simultaneously. A promising approach was proposed by Zhang et al. [29], who employed the CNN and recurrent neural network (RNN) to capture temporal and spatial information simultaneously from the PSG data. The architecture attained an accuracy of 87%. Although the combination of CNNs and RNNs can enhance the model performance to some extent, the high computational cost of RNNs should be taken into consideration. To the best of our knowledge, the training speed of CNNs would be dozens of times faster than that of RNNs under the same GPU acceleration when implementing long time-series input. To sum up, despite the fact that multi-channel PSG data can provide additional referenced information compared to single-channel EEG, there is also some irrelevant information being introduced. Furthermore, multi-channel recordings can limit the practical application on account of more complex operation and equipment costs.

Compared to the multi-channel scheme, the single-channel scheme can reduce the related cost and be much easier for data acquisition. Under the single-channel scheme, Supratak et al. [30] introduced a deep learning model called DeepSleepNet. DeepSleepNet utilizes the capacity of deep learning to extract time-invariant features automatically, the proposed model can be adapted to different datasets. However, the accuracy obtained from DeepSleepNet was 82%, which can not outperform the state-of-the-art approaches. A promising CNN model was proposed by Sors et al. [31], who used raw single-channel EEG to classify the sleep stage without any preprocessing. The architecture attained an accuracy of 87%, whereas the model complexity is high with 12 convolutional layers. Furthermore, the filer size was chosen among 7, 5, 3, the performance of larger size filters should be compared considering the long length of input.

To tackle these problems, this paper proposes the SingleChannelNet (SCNet), a model for automatic sleep stage classification based on raw single-channel EEG, which can learn different scale features simultaneously. We aim to automate the sleep stage classification completely by utilizing the capabilities of the proposed model. The main contributions of this work are summarized as follows:

- We propose a new deep learning model with low model complexity for sleep stage classification using 90s raw single-channel EEG.
- We implement two multi-convolution (MC) blocks with different filter sizes in our model. In addition, the max-average (M-Apooling) layer is applied to take place of the conventional max-pooling layer. Two strategies are used for capturing more feature representations from different scales to enhance the capacity of the feature extraction.
- The results demonstrate that our model can obtain promising performance on different raw single-channel EEGs (C4/A1, Fpz-Cz) from CCSHS and Sleep-EDF datasets, without modifying the architecture and hyper-parameters of model and training algorithm. Moreover, all features are learned by the proposed model automatically.

2. Experimental datasets

Two public datasets are employed to evaluate the performance of the proposed framework in this work, namely Cleveland Children's Sleep and Health Study (CCSHS) [36,37] and Sleep-EDF Database Expanded (Sleep-EDF, 2018 version) [38]. It should be noted that all hypnograms of experimental datasets are manually scored according to the R&K manual rather than the AASM rule.

2.1. Cleveland children's sleep and health study (CCSHS)

The CCSHS dataset comprises of overnight PSG recordings from 515 subjects aged 8–11 years, which is one of the largest population-based pediatric cohorts studied with objective sleep studies. Each 30s epoch is manually divided by experts into several stages: Wake (W), Rapid Eye Movement (REM), Non-REM1 (N1), Non-REM2 (N2), and Non-REM3 (N3). In this work, single-channel EEG C4/A1 sampled at 128 Hz is selected.

2.2. Sleep-EDF database expanded (Sleep-EDF)

The Sleep-EDF dataset consists of two subsets: sleep-cassette (SC) contains 78 healthy Caucasians aged from 25 to 101 years and sleep-telemetry (ST) comprises 22 Caucasians receiving temazepam treatment. In this study, we use subjects from SC for evaluating the model performance. Each participant was recorded two subsequent night PSG data except the subject 13, subject 36 and subject 52, from the SC subset who had only a one-night record. Each epoch of recordings is manually labelled by clinicians according to the R&K rule into W, N1, N2, N3, N4, REM, MOVEMENT and UNKNOWN stages respectively. In addition, MOVEMENT and UNKNOWN are discarded, as they do not belong to the six stages. The PSG data include two-channel EEGs (Fpz-Cz and Pz-Oz), single-channel EOG, single-channel EMG and the event marker (sampled at 1 Hz). The sampling rate f_s of EEG, EOG, and EMG is 100 Hz. Single-channel EEG Fpz-Cz is adopted in our experiment. For the Sleep-EDF dataset, stages N3 and N4 are merged into stage N3 which is consistent with the AASM manual. Additionally, resampling operation is not applied to C4/A1 and Fpz-Cz EEGs. We also found that most previous studies use the Sleep-EDF dataset of the first 20 subjects (Sleep-EDF-v1). For a fairer comparison, we also experiment with the Sleep-EDF-v1 dataset. We employ 30 min samples of stage W before and after other sleep stages as the recommendation of [30,33].

2.3. Contextual input

In previous works, most schemes use a single 30s epoch as the classifier input [7,35,31] and then produce a single output label. Although being straightforward, this classification method ignores the existing correlation and dependency between surrounding epochs. It is considered that the sleep stage classification depends not only on the local epoch, but also on the prior and following temporal features [6,9]. For this reason, an extension of single 30s epoch input is conducted by combining it with its neighboring epochs to make a contextual input. Furthermore, we employ 90s epoch ($\mathbf{Z}_m$) as contextual input of the proposed model, and it contains three sequential epochs: prior 30s epoch ($\mathbf{X}_{m-1}$), 30s epoch ($\mathbf{X}_m$) and subsequent 30s epoch ($\mathbf{X}_{m+1}$). The ground truth label of $\mathbf{Z}_m$ is y_m which also denotes $\mathbf{X}_m$'s label. As in

$$\mathbf{Z}_m = (\mathbf{X}_{m-1}, \mathbf{X}_m, \mathbf{X}_{m+1}) \mapsto y_m. \quad (1)$$

Details are illustrated in Fig. 1. As shown in Table 1, we summarize

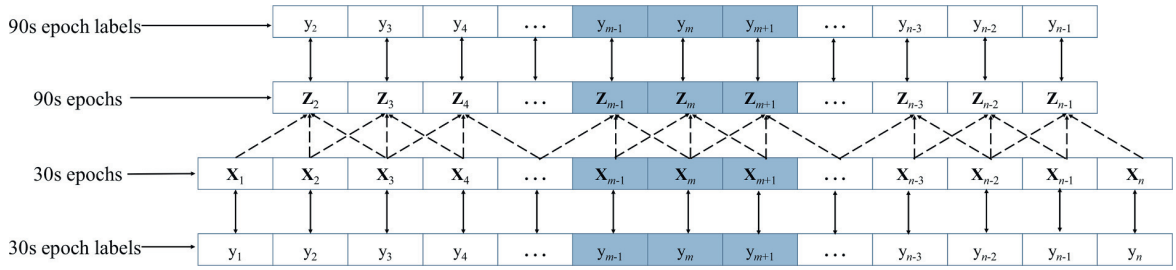


Fig. 1. Illustration of 90s epochs and labels used in this paper, n denotes the number of 30s epochs for a subject, $\mathbf{Z}_m$ is comprised of X_{m-1}, X_m and X_{m+1}, $2 \leqslant m \leqslant n-1$.

Table 1
Number of 90s Epochs for Each Sleep Stage from Experimental Datasets.

Dataset	W	N1	N2	N3	REM	Total
CCSHS	211030	19221	249681	110188	100252	690372
Sleep-EDF	69518	21522	69132	13039	25835	199046
Sleep-EDF-v1	10197	2804	17799	5703	7717	44220

the number of 90s epochs for each sleep stage from CCSHS, Sleep-EDF and Sleep-EDF-v1 datasets in our experiments. The distribution of the number of five stages is imbalanced. For all datasets, W and N2 stages account for more than 60% of all 90s epochs. By contrast, the proportion of stages N1 and N3 is the smallest.

3. Proposed SCNet

Fig. 2 shows the overall architecture of the SCNet. The convolution block performs three operations sequentially: one-dimensional convolutional layer (Conv1D), batch normalization and M-Apooling1D. Similarly, each MC Block is followed by batch normalization, M-Apooling1D and Dropout layer in sequence. In our model, we employ the concatenation of max-pooling and average-pooling to take place of the max-pooling for capturing more representable features. Similar to the inception module [39], the MC block contains different sizes of convolutional filters to capture the corresponding information. Besides, we use the Global Average Pooling (GAP) layer to replace the traditional fully connected layer, and it is proved to be more robust spatial translations of the input without parameter optimization[40].

3.1. Model specification

In Table 2, we relate detailed parameters of the proposed model. The size of the model's input is $(90 \times f_s, 1)$, where f_s is the sampling rate. To be specific, the f_s of EEG C4 and Fpz-Cz is 128 Hz and 100 Hz, respectively. Here, the SCNet does not restrict the length of input which can be applied to different datasets.

The first convolutional layer with 128 filters of size 128 and a stride of 2 is applied to obtain the feature map from raw single-channel EEG. The activation function of this layer is rectified linear unit (ReLU) which is defined as the positive part of its argument:

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

where x is the input of a neuron. To normalize the prior layer output, we apply the batch normalization technique. Besides, the M-Apooling layer can get the combination of maximum and average values from each of a cluster of neurons at the previous layer.

We implement two MC blocks in our model, and the filter sizes are selected among 1, 3, 16 and 64 to obtain multiscale representative features. More specifically, the small filter is prone to learn temporal information (i.e., when certain EEG patterns appear for a specific sleep

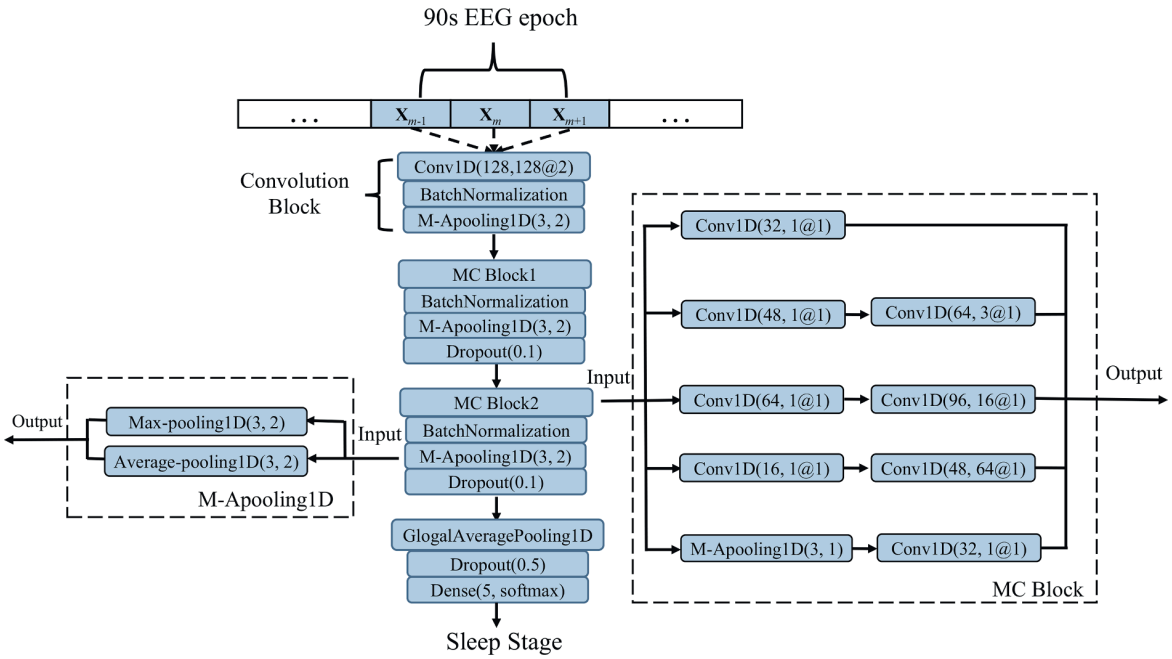


Fig. 2. An overall architecture of the proposed SCNet.

Table 2
Parameters of the Proposed Model.

Layer	Layer Type	Filters	Size	Stride	Activation	Output dimension
1	Input	–	–	–	–	$(90 \times f_s, 1)$
2	Conv1D	128	128	2	Relu	$(45 \times f_s, 128)$
3	M-Apooling1D	–	3	2	–	$(\lceil 45 \times f_s/2 \rceil, 256)$
4	MC Block1	–	–	1	Relu	$(\lceil 45 \times f_s/2 \rceil, 272)$
5	M-Apooling1D	–	3	2	–	$(\lceil 45 \times f_s/4 \rceil, 544)$
6	Dropout(0.1)	–	–	–	–	$(\lceil 45 \times f_s/4 \rceil, 544)$
7	MC Block2	–	–	1	Relu	$(\lceil 45 \times f_s/4 \rceil, 272)$
8	M-Apooling1D	–	3	2	–	$(\lceil 45 \times f_s/8 \rceil, 544)$
9	Dropout(0.1)	–	–	–	–	$(\lceil 45 \times f_s/8 \rceil, 544)$
10	GAP	–	–	–	–	544
11	Dropout(0.5)	–	–	–	–	544
12	Dense	–	–	–	Softmax	5

stage), while the large filter is better to capture frequency information [30]. Considering the long length of input ($128 \times 90, 100 \times 90$), we optimize the filter sizes from the small sizes (3, 5 and 7), medium sizes (16 and 32) and big sizes (64, 128 and 256). The filter size of 1 is to improve the nonlinearity of the network and reduce the dimension of previous layer output. It would not reduce the size of the feature map but can enhance the nonlinearity of the network through the nonlinear activation function. The filter sizes are chosen with 1, 3, 16 and 64 based on the optimized results. Furthermore, after concatenating the output of all convolutional layers, the dimension of the MC block1 output is ($\lceil 45 \times f_s/2 \rceil$, 272). The following M-Apooling layer can get ($\lceil 45 \times f_s/4 \rceil$, 544) dimension feature map. Each MC block is followed by a batch normalization layer, a M-Apooling layer with size of 3 and a dropout layer with the probability of 0.1. To find appropriate strides, we test 4 strides: 1, 2, 3 and 5. The stride of two MC blocks is set to 1, while the stride of the M-Apooling layer and the first convolutional layer is 2. The GAP layer is applied to flat the previous output before the final decision layer. Through a drop layer with drop rate of 0.5, the dense layer using softmax as the activation function makes the final decision. Softmax function can calculate the probabilities of five stages, the stage with maximum probability is as the consequence of the predicted sleep stage.

3.2. Regularization

We adopt two regularization approaches to help prevent the overfitting problem. The first technique is L2 regularization that adds squared magnitude of coefficient as penalty term to the loss function. It is important to choose a proper regularization rate (lambda), if lambda is very large, it would add too much weight causing an underfitting issue. By contrast, a very small lambda would make the model more complex, then the model would learn too much about the particularities of the training data, L2 regularization therefore has little effect on avoiding overfitting. Hence, we test four lambda values: $10^{-1}, 10^{-2}$, 10^{-3} and 10^{-4}, the results show that 10^{-3} achieves the best performance. The L2 regularization is applied to all convolutional layers, including the MC block.

Another regularization method is dropout, which randomly drops units from the model during training with a specific probability from 0 to 1. We evaluate two dropout rates (0.1 and 0.5) in the process of hyper-parameters optimization. Dropout layers with a probability of 0.1 and 0.5 are employed for the MC block and GAP layer, respectively.

3.3. Training setup

We select Adam as the network optimizer whose parameters ((learning rate) $lr, beta1$ and $beta2$) are set to 10^{-3}, 0.9 and 0.999 respectively. Moreover, ReduceLROnPlateau of Callback in Keras is implemented to reduce the lr. Specifically, when the model monitors the validation accuracy showing no improvement within 3 epochs, the lr would drop to half of it. The minimum lr is set to 10^{-7}. To find out appropriate batch size of mini-batch, sizes of 32, 64, 128, and 256 are evaluated, we select 64 as the size of mini-batch finally. The categorical cross entropy is chosen as the loss function of the model which is always used for classifying multi-class tasks. The model converges to the optimal solution within 40 iterations, hence the number of iteration is set to 40.

There are two types of methods to split the training and test sets [30]. One is the subject-wise scheme which splits the training and test datasets based on the subjects. Another one is the epoch-wise method in which the split is conducted by epochs rather than subjects. In the epoch-wise scheme, We use 20% of whole data set as the test set and the remaining 80% epochs as the training set. As for the subject-wise approach, 80% subjects are selected as the training set, the other 20% subjects are used as the test set. Furthermore, we use the 5-fold cross-validation (80% training set for training, 20% training set for validation) scheme to train and evaluate our model for both datasets. In addition, only 90s epochs from the CCSHS dataset are used to determine the hyper-parameters of the proposed model. Once achieving optimal hyper-parameters, they would be used in all experiments. To be specific, when the model is applied to another dataset, there would be no need to modify the architecture and hyper-parameters of the model except for the input length which should adapt to the f_s of EEG from different datasets. To show the effect of using the contextual input, we also conduct the experiments employing the 30s epochs with the same training setup.

Graphic card Nvidia Tesla P100 with 16 Gbytes memory is used for model training. The implementation is written in Keras [41] with the Tensorflow backend [42].

4. Experimental results

4.1. Performance metrics

We evaluate the model performance (epoch-wise) using overall accuracy (*ACC*), precision (*PR*), recall (*RE*), F1 score (*F*1), and Cohen's kappa coefficient (*K*). *ACC* is the proportion of correct predictions made by the model to the total predications. *PR* calculates the ratio of correctly predicted positives to all positives. *RE* means the fraction between true positives and all predications in the actual class. *F*1 represents the weighted average of *PR* and *RE*. *K* measures the agreement between true labels and predicted labels. A large value of *K* can indicate good performance of the model.They are calculated as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{3}$$

$$PR = \frac{TP}{TP + FP}. \tag{4}$$

$$RE = \frac{TP}{TP + FN}. \tag{5}$$

$$F1 = 2 \cdot \frac{RE \cdot PR}{RE + PR}. \tag{6}$$

$$K = \frac{\sum_{i=1}^{n} \frac{x_{ii}}{N} - \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji} \right)}{N^2}}{1 - \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji} \right)}{N^2}}. \tag{7}$$

where *TP*, *TN*, *FN* and *FP* stand for the true positives, true negatives, false negatives and false positives, respectively. *N* is the number of 90s epochs of the test set, *n* represents the number of classes. In this work, *n* equals 5, x_{ii} ($1 \leqslant i \leqslant 5$) represents the diagonal value of the confusion matrix.

To show the performance of each fold cross-validation from the CCSHS and Sleep-EDF datasets, we present the normalized confusion matrices (CM) in Fig. 3. Firstly, we use single-channel EEG C4/A1 (90s epochs) from the CCSHS dataset to tune the hyper-parameters. Once getting the best performance, the hyper-parameters and model architecture are fixed for all experiments. Table 3 provides the mean CM of 5-fold cross-validation from the CCSHS dataset, we can see that the overall accuracy and *K* are respectively 90.2% and 86.5%. The proposed model shows the best ability to detect the W stage with the *PR* of 94.7%. By contrast, the performance of stage N1 classification is the worst which is consistent with the results of existing works. To be specific, there are 33.0% of N1 90s epochs being recognized correctly. In addition, 27.4% of N1 samples are misclassified as W, 19.5% as N2 and 20.1% as REM. Stages N2, N3 and REM have similar classification results in terms of the *PR* corresponding to 88.8%, 91.0% and 87.9% respectively.

To demonstrate the generalization capability of the proposed architecture, we also conduct the 5-fold cross-validation using the same model determined by the CCSHS dataset (i.e., without any hyper-parameters modification except for the input length) on the Sleep-EDF dataset. As can be seen from Table 1, the distribution of the numbers of five stages is different. Stage W has the biggest proportion and the number of N3 is the smallest in Sleep-EDF dataset, whereas the largest percentage is stage N2 in the CCSHS dataset. Besides, the EEG channel used in two datasets is also distinct, C4/A1 for the CCSHS dataset and Fpz-Cz for the Sleep-EDF dataset. It is worthy to note that despite the EEG channel and the size of the input length ($90 \times f_s$, 1) are quite different, the proposed model can obtain promising performance on two different datasets by comparing Tables 3 and 4. The performance of the conventional 30s input is illustrated in Table 5 and Table 6. Employing with the contextual input, the *ACC* achieves an enhancement of 1.1% and 4.1% respectively on the CCSHS, Sleep-EDF datasets compared to the *ACC* of 30s input length. Likewise, the *K* could be improved respectively by 1.5% and 5.7%. We further reveal the hypnogram comparison labeled by experts and the model's prediction for one subject of CCSHS and Sleep-EDF datasets in Fig. 4.

4.2. Performance comparison

We make a comparison between the proposed model (epoch-wise and subject-wise) with some existing works using the same datasets in terms of the *ACC* and *K* in Tables 7 and 8. Table 7 reveals that the proposed framework can achieve higher *ACC* and *K* using raw single-channel C4/A1 EEG compared to approaches using multi-channel PSG data [43] or the single-channel EEG [44] on the CCSHS dataset. For the Sleep-EDF and Sleep-EDF-v1 databases, the proposed model also achieves comparable performance compared to state-of-the-art methods. Some studies [25,34] extract features manually or some methods adopt single-channel EEG [33,45,46]. Considering results of the comparison, the proposed framework can achieve promising performance on CCSHS, Sleep-EDF and Sleep-EDF-v1 datasets.

5. Discussion and conclusion

In this paper, we propose an end-to-end framework with CNNs, namely SCNet, which combines the feature learning ability and classification capacity. The proposed model is applied to classify sleep stages automatically from raw single-channel EEG without using any hand-engineered features and any other preprocessing (e.g., signal filtering and resample implementation). There are two main advantages that we train and evaluate the model with raw single-channel EEG. Comparing with those methods with hand-crafted features [4,12,47], where extracting hand-engineered features is conducted with priori knowledge and not in a data-driven way, and it is time-consuming for the researchers. Moreover, the selection of types and number of features would result in different model performance, there is no gold standard about the extraction of hand-crafted features. The second advantage is that it is much easier and more comfortable to record single-channel EEG data compared to the multi-channel scheme [6,28] either at the hospital or home. Moreover, multi-channel PSG data used as input can increase the computational cost. Considering practical applications, the use of raw single-channel EEG can simplify the measurement scheme and reduce the related cost.

Comparing with the conventional deep neural network based on CNNs, where the convolutional layers with the fixed filter size are assembled in sequence. In such a case, it is not capable of capturing features representation from different scales. To address this issue, our model employs two MC blocks, which are the concatenation of several

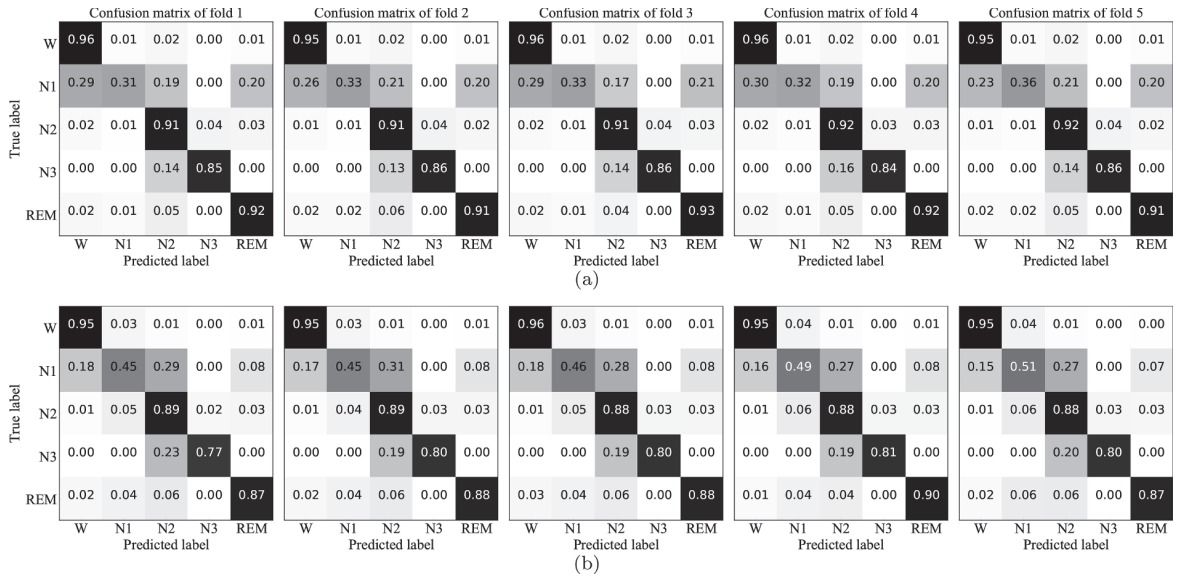


Fig. 3. The normalized confusion matrices of each fold cross-validation. (a) CCSHS dataset and (b) Sleep-EDF dataset.

Table 3
Mean Confusion Matrix of 5-Fold Cross-validation on Raw Single-channel EEG C4/A1 from the CCSHS Dataset.

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	*PR*(%)	*RE*(%)	*F*1(%)	*ACC*(%)	*K*(%)
W	40450	440	934	124	392	**94.7**	**95.5**	**95.1**		
N1	1039	1253	740	1	765	52.8	33.0	40.6		
N2	766	382	45679	1739	1359	88.8	91.5	90.1	90.2	86.5
N3	60	0	3126	18791	8	91.0	85.5	88.1		
REM	384	299	971	6	18358	87.9	91.7	89.8		

Table 4
Mean Confusion Matrix of 5-Fold Cross-validation on Raw Single-channel EEG Fpz-Cz from the Sleep-EDF Dataset.

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	*PR*(%)	*RE*(%)	*F*1(%)	*ACC*(%)	*K*(%)
W	14650	498	142	8	82	**94.1**	**95.3**	**94.7**		
N1	712	2020	1214	11	330	58.4	47.1	52.1		
N2	109	705	12255	385	383	84.9	88.6	86.7	86.1	80.5
N3	4	5	532	2127	4	84.0	79.6	81.7		
REM	99	233	299	1	4573	85.1	87.9	86.5		

Table 5
Mean Confusion Matrix of 5-Fold Cross-validation on Raw Single-channel EEG Fpz-Cz from the CCSHS Dataset with the 30s Input Length.

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	*PR*(%)	*RE*(%)	*F*1(%)	*ACC*(%)	*K*(%)
W	40760	417	606	85	472	**95.1**	**96.3**	**95.7**		
N1	742	1068	636	1	1351	46.0	28.1	34.9		
N2	746	430	44952	1586	2211	88.2	90.0	89.2	89.1	85.0
N3	72	1	3665	18235	12	91.6	82.9	87.0		
REM	520	407	1062	9	18020	81.7	90.0	85.6		

Table 6
Mean Confusion Matrix of 5-Fold Cross-validation on Raw Single-channel EEG Fpz-Cz from the Sleep-EDF Dataset with the 30s Input Length.

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	*PR*(%)	*RE*(%)	*F*1(%)	*ACC*(%)	*K*(%)
W	14522	443	181	11	223	**91.8**	**94.4**	**93.1**		
N1	840	1370	1338	19	720	48.2	32.0	38.4		
N2	218	644	11831	376	768	82.3	85.5	83.9	82.0	74.8
N3	8	7	508	2140	9	83.9	80.1	81.9		
REM	225	380	511	5	4086	70.4	78.5	74.2		

convolutional layers with four distinct filer sizes, to extract different scale features. Instead of using the traditional max-pooling layer, we adopt the M-Apooling layer to add average feature representation with maximum features simultaneously, which further improve the proposed model's ability of feature learning. In addition, the SCNet model is quite simple and compact with a total 5×10^5 parameters compared to the methods in [45] which has 2.1×10^7 parameters and [30] in which the number of parameters of the representation learning and sequence residual learning parts has up to 6×10^5 and 2×10^7 respectively. Moreover, the proposed SCNet model can achieve the comparable performance with less computing resources occupied. Concerning online and realtime applications (e.g., sleep monitoring), our model with raw single-channel EEG is more reasonable to reduce the time latency and obtain reliable results.

As the use of textual input, it is considered that the sleep stage classification depends not only on the local epoch, but also on prior and following temporal features [9]. For instance, the beginning of stage N2 is decided by the occurrence of K-complex or beta-frequency spindle activity in the early or last half of the prior 30 epoch [7]. Inspired by this, we choose the 90s epoch as textual input of proposed model despite the higher computable cost. The performance comparisons in Table 5 and Table 6 also demonstrate the advantage of contextual input. To validate the generalization of the proposed architecture, different single-channel EEGs from two datasets are adopted. The length of input is not restricted to a fixed number, our model can be adapted to different length of input relating to the f_s of EEG efficiently. Experimental results show that the proposed model can obtain promising performance on two datasets (CCSHS: *ACC*-90.2%, *K*-86.5%; Sleep-EDF: *ACC*-86.1%, *K*-80.5%), which indicate the desirable generalization of the SCNet model.

It is challenging to train on dataset A and test on B, not only for the proposed SCNet but also for typical CNNs. CNNs are running in a data-driven way which means the model must learn some crucial features from the training samples. Otherwise, it cannot perform well on an unfamiliar dataset. This is also the biggest difference (generalization ability) between machine and human, human beings are good at deducing and inducing. To further show the generalization ability of the proposed model, we perform two additional experiments. Firstly, we train our model with the CCSHS database, the obtained model then is tested on the Sleep-EDF dataset without any training, the accuracy is 65.9%. In reverse, The proposed model is trained on the Sleep-EDF dataset and tested with the CCSHS database, the accuracy achieved is 70.2%. In our future work, we will try to construct a more brain-inspired model with some cognitive neural dynamic from neuroscience [48] to

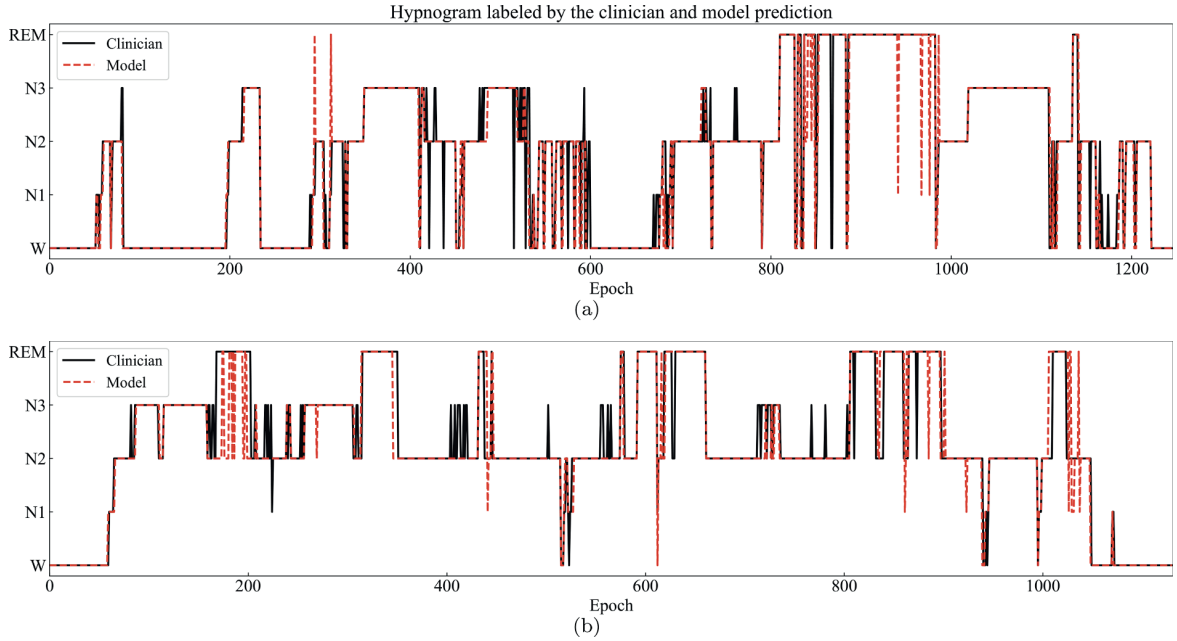


Fig. 4. The comparison between hypnogram labeled by the clinician and the model's prediction. The solid black line is the ground truth, the dotted red line donates the hypnogram labeled by the prediction of the proposed model. (a) CCSHS dataset (ccshs-trec-1800905) and (b) Sleep-EDF dataset (SC4091).

Table 7
Performance Comparison between The Proposed Method and Previous Methods on the CCSHS Dataset.

Study	Database	Method	Input channel	Preprocessing	Input type	Subjects	*ACC*(%)	*K*(%)
Nakamura et al. [43]	CCSHS	HMM	C4/A1 + C3/A2	Yes	Spectrogram	515	-	73
Li et al. [44]	CCSHS	Random Forest	C4/A1	Yes	Features	116	86.0	80.5
Proposed (subject-wise)	**CCSHS**	**Deep CNN**	**C4/A1**	**No**	**Time series**	**515**	**88.2**	**83.8**
Proposed (epoch-wise)	**CCSHS**	**Deep CNN**	**C4/A1**	**No**	**Time series**	**515**	**90.2**	**86.5**

Table 8
Performance Comparison between The Proposed Method and Previous Methods on the Sleep-EDF and Sleep-EDF-v1 Datasets.

Study	Database	Method	Input channel	Preprocessing	Input type	Subjects	*ACC*(%)	*K*(%)
Phan et al. [25]	Sleep-EDF	RNN	Fpz-Cz	Yes	Time-frequency image	78	82.6	76
Mousavi et al. [45]	Sleep-EDF	CNN + LSTM	Fpz-Cz	No	Time series	78	80.0	73
Supratak et al. [46]	Sleep-EDF	CNN + LSTM	Fpz-Cz	No	Time series	78	83.1	77
Proposed (subject-wise)	**Sleep-EDF**	**Deep CNN**	**Fpz-Cz**	**No**	**Time series**	**78**	**83.9**	**77.8**
Proposed (epoch-wise)	**Sleep-EDF**	**Deep CNN**	**Fpz-Cz**	**No**	**Time series**	**78**	**86.1**	**80.5**
Supratak et al. [30]	Sleep-EDF-v1	CNN + LSTM	Fpz-Cz	No	Time series	20	82.0	76
Seo et al. [32]	Sleep-EDF-v1	CNN + LSTM	Fpz-Cz	No	Time series	20	83.9	78
Wei et al. [33]	Sleep-EDF-v1	Deep CNN	Fpz-Cz	Yes	Time series	20	84.3	78
Phan et al. [34]	Sleep-EDF-v1	1-max CNN	Fpz-Cz	Yes	Time-frequency image	20	82.6	76
Proposed (subject-wise)	**Sleep-EDF-v1**	**Deep CNN**	**Fpz-Cz**	**No**	**Time series**	**20**	**86.2**	**81.1**
Proposed (epoch-wise)	**Sleep-EDF-v1**	**Deep CNN**	**Fpz-Cz**	**No**	**Time series**	**20**	**91.0**	**87.8**

increase the generalization ability for the sleep stage classification task. On the other hand, the class distribution of PSG datasets is highly imbalanced, this class imbalance problem has not been solved well in this work. As a representation of the monitory class, the recognition rate of N1 is still much lower than that of other stages. It is worthy of investigating appropriate data argumentation methods to balance the samples of PSG datasets. Also, it is valuable to adopt clinic datasets that have rarely been explored in previous studies.

CRediT authorship contribution statement

Dongdong Zhou: Conceptualization, Methodology, Software, Writing – original draft. **Jian Wang:** Writing – review & editing. **Guoqiang Hu:** Writing – review & editing. **Jiacheng Zhang:** Writing – review & editing. **Fan Li:** Writing – review & editing. **Rui Yan:** Writing – review & editing. **Lauri Kettunen:** Supervision. **Zheng Chang:** Supervision. **Qi Xu:** Supervision. **Fengyu Cong:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 91748105), National Foundation in China (No.

JCKY2019110B009, 2020-JCJQ-JJ-252) and the Fundamental Research Funds for the Central Universities [DUT20LAB303, DUT20LAB308, DUT21RC(3)091] in Dalian University of Technology in China and the Scholarships from China Scholarship Council (No. 201806060164, No. 202006060226) and CAAI-Huawei Mindspore Open Fund (No. CAAIXSJLJJ-2021-003A). This study is also to memorize Prof. Tapani Ristaniemi from University of Jyväskylä for his great help to the authors.

References

[1] R. Ferri, M. Manconi, G. Plazzi, O. Bruni, S. Vandi, P. Montagna, L. Ferini-Strambi, M. Zucconi, A quantitative statistical analysis of the submentalis muscle emg amplitude during sleep in normal controls and patients with rem sleep behavior disorder, J. Sleep Res. 17 (1) (2008) 89–100.
[2] C. Kuo, S. Liang, Automatic stage scoring of single-channel sleep eeg based on multiscale permutation entropy, IEEE Biomed. Circuits Syst. Conf (BioCAS) (2011) 448–451.
[3] S.J. Redmond, C. Heneghan, Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea, IEEE Trans. Biomed. Eng 53 (3) (2006) 485–496.
[4] G. Zhu, Y. Li, P. Wen, Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal, IEEE J. Biomed. Health. Inf. 18 (6) (2014) 1813–1821.
[5] H.G. Jo, J.Y. Park, C.K. Lee, S.K. An, S.K. Yoo, Genetic fuzzy classifier for sleep stage identification, Comput. Biol. Med 40 (7) (2010) 629–634.
[6] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, Joint classification and prediction cnn framework for automatic sleep stage classification, IEEE Trans. Biomed. Eng. 66 (5) (2018) 1285–1296.
[7] H. Dong, A. Supratak, W. Pan, C. Wu, P.M. Matthews, Y. Guo, Mixed neural network approach for temporal sleep stage classification, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2) (2017) 324–333.
[8] A. Rechtschaffen, A manual of standardized terminology and scoring system for sleep stages of human subjects, Electroencephalogr. Clin. Neurophysiol. 26 (6) (1969) 644.
[9] R.B. Berry, R. Budhiraja, D.J. Gottlieb, D. Gozal, C. Iber, V.K. Kapur, C.L. Marcus, R. Mehra, S. Parthasarathy, S.F. Quan, et al., Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine, J. Clin. Sleep Med. 8 (5) (2012) 597–619.
[10] B. Koley, D. Dey, An ensemble system for automatic sleep stage classification using single channel eeg signal, Comput. Biol. Med. 42 (12) (2012) 1186–1195.
[11] K. Šušmáková, A. Krakovská, Discrimination ability of individual measures used in sleep stages classification, Artif. Intell. Med. 44 (3) (2008) 261–277.
[12] F. Karimzadeh, R. Boostani, E. Seraj, R. Sameni, A distributed classification procedure for automatic sleep stage scoring based on instantaneous electroencephalogram phase and envelope features, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2) (2017) 362–370.
[13] O. Tsinalis, P.M. Matthews, Y. Guo, Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders, Ann. Biomed. Eng. 44 (5) (2016) 1587–1597.
[14] M. Sharma, D. Goyal, P. Achuth, U.R. Acharya, An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank, Comput. Biol. Med. 98 (2018) 58–75.
[15] S. Kouchaki, S. Sanei, E.L. Arbon, D.-J. Dijk, Tensor based singular spectrum analysis for automatic scoring of sleep eeg, IEEE Trans. Neural Syst. Rehabil. Eng. 23 (1) (2014) 1–9.
[16] H. Phan, Q. Do, T.-L. Do, D.-L. Vu, Metric learning for automatic sleep stage classification, IEEE Eng. Med. Biol. Soc (EMBC) (2013) 5025–5028.
[17] E. Alickovic, A. Subasi, Ensemble svm method for automatic sleep stage classification, IEEE Trans. Instrum. Meas. 67 (6) (2018) 1258–1265.
[18] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines, J. Neurosci. Methods 250 (2015) 94–105.
[19] 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.
[20] S. Güneş, K. Polat, Ş. Yosunkaya, Efficient sleep stage recognition system based on eeg signal using k-means clustering based feature weighting, Expert Syst. Appl. 37 (12) (2010) 7922–7928.
[21] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier, Comput. Methods Programs Biomed. 108 (1) (2012) 10–19.
[22] M. Xiao, H. Yan, J. Song, Y. Yang, X. Yang, Sleep stages classification based on heart rate variability and random forest, Biomed. Signal Process. Control 8 (6) (2013) 624–633.
[23] T.L. da Silveira, A.J. Kozakevicius, C.R. Rodrigues, Single-channel eeg sleep stage classification based on a streamlined set of statistical features in wavelet domain, Med. Biol. Eng. Comput. 55 (2) (2017) 343–352.
[24] P. Memar, F. Faradji, A novel multi-class eeg-based sleep stage classification system, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (1) (2017) 84–95.
[25] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (3) (2019) 400–410.
[26] R. Yan, F. Li, D.D. Zhou, T. Ristaniemi, F. Cong, Automatic sleep scoring: A deep learning architecture for multi-modality time series, J. Neurosci. Methods 348 (2021), 108971.
[27] S. Chambon, M.N. Galtier, P.J. Arnal, G. Wainrib, A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (4) (2018) 758–769.
[28] K. Mikkelsen, M. De Vos, Personalizing deep learning models for automatic sleep staging, arXiv Preprint. arXiv:1801.02645,.
[29] L. Zhang, D. Fabbri, R. Upender, D. Kent, Automated sleep stage scoring of the sleep heart health study using deep neural networks, Sleep 42 (11) (2019) zsz159.
[30] A. Supratak, H. Dong, C. Wu, Y. Guo, DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel eeg, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (11) (2017) 1998–2008.
[31] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, J.-F. Payen, A convolutional neural network for sleep stage scoring from raw single-channel eeg, Biomed. Signal Process. Control 42 (2018) 107–114.
[32] H. Seo, S. Back, S. Lee, D. Park, T. Kim, K. Lee, Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg, Biomed. Signal Process. Control 61 (2020), 102037.
[33] W. Qu, Z. Wang, H. Hong, Z. Chi, D.D. Feng, R. Grunstein, C. Gordon, A residual based attention model for eeg based sleep staging, IEEE J. Biomed. Health Inf.
[34] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification, IEEE Eng. Med. Biol. Soc (EMBC) (2018) 453–456.
[35] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks, IEEE Eng. Med. Biol. Soc (EMBC) (2018) 1452–1455.
[36] G.-Q. hang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, S. Redline, The national sleep research resource: towards a sleep data commons, J. Am. Med. Inform. Assoc. 25(10) (2018) 1351–1358.
[37] C.L. Rosen, E.K. Larkin, H.L. Kirchner, J.L. Emancipator, S.F. Bivins, S.A. Surovec, R.J. Martin, S. Redline, Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity, J. Pediatr. 142 (4) (2003) 383–389.
[38] B. Kemp, A.H. Zwinderman, B. Tuk, H.A. Kamphuisen, J.J. Oberye, Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg, IEEE Trans. Biomed. Eng. 47 (9) (2000) 1185–1194.
[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR), IEEE (2015) 1–9.
[40] M. Lin, Q. Chen, S. Yan, Network in network, arXiv Preprint. arXiv:1312.4400.
[41] F. Chollet, et al., Keras: Deep learning library for theano and tensorflow. URL: https://keras.io/k, 7(8) (2015) T1.
[42] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv Preprint. arXiv:1603.04467,.
[43] T. Nakamura, H.J. Davies, D.P. Mandic, Scalable automatic sleep staging in the era of big data, IEEE Eng. Med. Biol. Soc (EMBC) (2019) 2265–2268.
[44] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, G.-Q. Zhang, Hyclasss: a hybrid classifier for automatic sleep stage scoring, IEEE J. Biomed. Health Inf. 22 (2) (2017) 375–385.
[45] S. Mousavi, F. Afghah, U.R. Acharya, Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach, PLOS ONE 14 (5) (2019) 1–15.
[46] A. Supratak, Y. Guo, TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg, IEEE Eng. Med. Biol. Soc (EMBC) (2020) 641–644.
[47] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, C.-Y. Hsu, Automatic sleep stage recurrent neural classifier using energy features of eeg signals, Neurocomputing 104 (2013) 105–114.
[48] Q. Xu, J. Peng, J. Shen, H. Tang, G. Pan, Deep CovDenseSNN: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment, Neural Netw. 121 (2020) 512–519.

PII

LIGHTSLEEPNET: A LIGHTWEIGHT DEEP MODEL FOR RAPID SLEEP STAGE CLASSIFICATION WITH SPECTROGRAMS

by

Dongdong Zhou, Qi Xu, JianWang, Jiacheng Zhang, Guoqiang Hu, Lauri Kettunen, Zheng Chang, and Fengyu Cong 2021
43rd Annual International Conference of the IEEE Engineering in Medicine and

Biology Society (EMBC2021), pp. 43-46,
https://doi.org/10.1109/EMBC46164.2021.9629878

LightSleepNet: A Lightweight Deep Model for Rapid Sleep Stage Classification with Spectrograms

Dongdong Zhou[1,2], Qi Xu[3,4,*], Jian Wang[1,2], Jiacheng Zhang[5], Guoqiang Hu[1], Lauri Kettunen[2], Zheng Chang[2], *Senior Member, IEEE* and Fengyu Cong[1,2,3,6,*], *Senior Member, IEEE*

***Abstract*— Deep learning has achieved unprecedented success in sleep stage classification tasks, which starts to pave the way for potential real-world applications. However, due to its enormous size, deployment of deep neural networks is hindered by high cost at various aspects, such as computation power, storage, network bandwidth, power consumption, and hardware complexity. For further practical applications (e.g., wearable sleep monitoring devices), there is a need for simple and compact models. In this paper, we propose a lightweight model, namely LightSleepNet, for rapid sleep stage classification based on spectrograms. Our model is assembled by a much fewer number of model parameters compared to existing ones. Furthermore, we convert the raw EEG data into spectrograms to speed up the training process. We evaluate the model performance on several public sleep datasets with different characteristics. Experimental results show that our lightweight model using spectrogram as input can achieve comparable overall accuracy and Cohen's kappa (SHHS100: 86.7%-81.3%, Sleep-EDF: 83.7%-77.5%, Sleep-EDF-v1: 88.3%-84.5%) compared to the state-of-the-art methods on experimental datasets.**

I. INTRODUCTION

High quality sleep plays an important role in humans' health. It has a significant influence on diagnosing and treating sleep-related disorders (e.g., insomnia) through correct sleep stage classification [1]–[4]. In order to accomplish the sleep scoring task, overnight polysomnography (PSG) data need to be recorded by several sensors attaching to different parts of the body. The PSG recordings mainly comprise electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG), electrooculogram (EOG) and so on [5]. In clinical practice, the PSG data are usually split into 30s segments sequentially. Each 30s epoch is further classified into different sleep stages by experienced clinicians manually according to sleep manuals. Specifically, sleep stages include six stages: Wake (W), Rapid Eye Movement (REM), Non-REM1 (N1), Non-REM2 (N2), Non-REM3 (N3) and Non-REM4 (N4) based on the Rechtschaffen and Kales (R&K) standard [6].

Nevertheless, the manual sleep stage classification is not only prone to subjective error but also time-consuming. Therefore, there is an urgent need for an effective sleep scoring method to release the workload of clinicians and obtain reliable performance. Recently, deep learning has been applied to automatic sleep stage classification successfully due to its powerful learning ability of feature extraction in a data-driven way. Whereas, various approaches based on convolutional neural network (CNN) have too much complex structures with millions of parameters [1] leading to probable overfitting issue and the high demand for computing resources. This drawback hinders those methods from further practical application (e.g., portable sleep monitor devices), which require light but efficient methods on resource-constrained devices, the time costs of model training should be also considered. Compared to previous studies using PSG data time series ($30 \times f_s$, channel) as input, CNNs are more efficient to process static imagery or matrix structure, which means CNNs are good at processing static images through their powerful feature extraction structure.

To solve mentioned problems, we propose the LightSleepNet (LSNet), a lightweight model for automatic sleep stage classification based on single-channel EEG. Transforming the raw time-series EEG signals to static spectrograms as the input, this model could be implemented and trained in a more efficient way which is suitable for rapid sleep stage classification tasks. The main contributions of this work are as follows:

i) We propose a light but efficient model with much fewer model parameters for automatic sleep stage classification.
ii) To speed up the training process, we utilize the spectrograms through short-time Fourier transform as the model input rather than the long time series EEG signals.
iii) The results demonstrate that our model can achieve comparable performance on different single-channel EEGs (C4/A1, Fpz-Cz) on experimental datasets with different characteristics.

This work was support by National Natural Science Foundation of China (Grant No.91748105), National Foundation in China (No. JCKY2019110B009, 2020-JCJQ-JJ-252), Fundamental Research Funds for Central Universities [DUT2019, DUT20LAB303] in Dalian University of Technology in China and the scholarships from China Scholarship Council (No.201806060164, No.202006060226), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2020-024A).

[1]School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China (*corresponding authors: xuqi123@zju.edu.cn, cong@dlut.edu.cn).

[2]Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland.

[3]School of Artificial Intelligence, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China.

[4]3College of Computer Science and Technology, Zhejiang University, 310027, Hangzhou, China.

[5]School of Information and Communication Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China.

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

TABLE I
THE DISTRIBUTION OF EACH SLEEP STAGE OF EACH DATASET

Dataset	W	N1	N2	N3	REM	Total
SHHS-100	23708	3010	41207	14306	14989	97220
Sleep-EDF	69518	21522	69132	13039	25835	199046
Sleep-EDF-v1	10917	2804	17799	5703	7717	44220

II. MATERIALS AND METHODS

A. Data Description

We evaluate the performance of proposed model employing three public PSG datasets: Sleep Heart Health Study (SHHS), Sleep-EDF Database (Sleep-EDF-v1, version 2013) and Sleep-EDF Database Expanded (Sleep-EDF, version 2018). The corresponding hypnograms of three datasets were scored by the well-trained clinicians following the R&K rule. For all employed datasets, we adopt single-channel EEG which can benefit to further reduce the computational cost and simply the scheme of data acquisition.

The SHHS dataset includes two subsets: initial PSG (SHHS1) and second PSG (SHHS2). Unlike the computer vision research, it is difficult to acquire abundant PSG samples to train the model. In order to better show the few-shot learning ability of the proposed model in sleep stage classification tasks, we adopt 100 near-normal subjects from the SHHS1 (i.e., SHHS-100) with the standard of the respiratory disturbance index 3 percent (RDI3P) $<$ 15 and no reported high pressure, cardiopathy or stroke. Single-channel EEG C4 sampled at 125 Hz is utilized for evaluating the proposed model as the suggestion of AASM manual. Detailed information of the SHHS dataset can be found in [7].

In the Sleep-EDF database, a total of 78 subjects with 153 whole-night PSG recordings from the sleep-cassette (SC) subset are selected. In addition, we also conduct the experiments on the first version of the Sleep-EDF dataset (Sleep-EDF-v1) before the expansion to make a fair comparison with the existing methods. We employ the single-channel EEG Fpz-Cz with 100 Hz in our experiments. To keep the same f_s, the EEG Fpz-Cz is resampled at 125 Hz. The [8] presents the detailed description of Sleep-EDF dataset.

For both datasets, we merge the N3 and N4 stages into stage N3 according to the latest AASM manual [9]. Hence each 30s epoch is labeled as one of five sleep stages (i.e., W, N1, N2, N3 and REM). In this paper, we use three successive EEG epochs (90s epoch) rather than the conventional 30s epoch as the contextual input of model. It is considered that experts classify the sleep stage depend not only on the current epoch but also the preceding and succeeding epochs. Also, the contextual input can enhance the model's learning ability of the transition information between epochs. The corresponding label of the 90s epoch is the label of current 30s epoch. As shown in Table I, we reveal the distribution of each stage from experimental datasets.

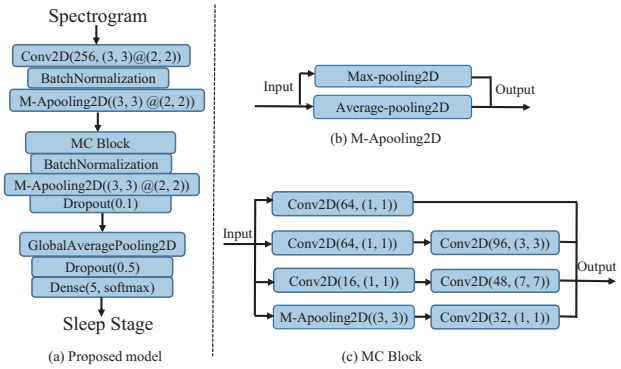


Fig. 1. The overall architecture of proposed model.

B. Data Preprocessing

The raw EEG data are filtered by a notch filter, a high-pass filter and a low-pass filter to eliminate the effect of noise and artifacts. To get the power spectrum of each 90s epoch, we adopt the short-time Fourier transform (STFT) with a window size of two seconds and 50% overlap. Moreover, Hamming window and 256 points Fast Fourier Transform (FFT) are conducted. The effective frequency band is set to 0.5-30 Hz and the obtained power spectrum is then converted to the log-power spectrum of size of $F \times T$, where F = 61, T = 89.

C. The Proposed Model

Different from our prior work [10] in which the SCNet is trained with the raw EEG data, the proposed model here, namely LSNet, is designed for handling with the spectrogram input. We show in Fig. 1 the overall architecture of proposed model. The first two-dimensional convolutional (Conv2D) layer with 256 filters of size 3×3 and the stride of 2 points is used to attain the feature map from the spectrum input ($61 \times 89 \times 1$) and the activation function is rectified linear unit (ReLU). Additionally, we apply the batch normalization to normalize the output of the first Conv2D layer. The M-Apooling2D layer is the concatenation of max-pooling2D and average-pooling2D layers that can learn feature representation from two scales.

We construct a multi-convolution (MC) block containing three different sizes of filters (1×1, 3×3, 7×7) to obtain multiscale features simultaneously. To be specific, the small filter is better to learn the temporal context, while the large filter is prone to capture the frequency information. Similar to the first Conv2D layer, the MC block is followed by the batch normalization and M-Aplooing2D layers. Besides, a dropout layer with the probability of 0.1 is applied to avoid the overfitting problem. It should be noticed that the size of stride in the MC block is set as 1×1.

Aiming to flat the previous output, we implement the global average pooling layer before the decision layer. A dropout layer with a drop rate of 0.5 can further prevent the overfitting issue. Except for the dropout method, another solution for overfitting applying in this work is the L2

regularization, which adds a squared magnitude of coefficient as penalty term to the loss function. The regularization rate, lambda, is chosen as 10^{-3} based on the experimental results of four lambda values (10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}). The final output is achieved by a dense layer whose activation function is the softmax to determine the probability of each stage, the stage with maximum probability is considered as the predicted sleep stage. The detailed parameters of proposed model are illustrated in Table II.

We also assemble a baseline model, in which time series as the input, for making a comparison with the proposed model in terms of the training computational cost for each iteration. The baseline is consistent with the structure and training setup except for the filter sizes. To be specific, the filter size of $N \times N$ is replaced by the filer size of N. For instance, 3×3 in the LSNet model should be converted into size of 3 in the baseline model.

D. Training Setup

We use 5-fold cross-validation to assess our model performance on all databases. The whole subjects are divided into training and test sets with a ratio of 4 to 1 using the subject-wise and epoch-wise schemes independently. For the subject-wise approach, the division of training and test sets is based on subjects. Nevertheless, we divide the whole epochs into training and test sets following the epoch-wise scheme. In each fold, we further employ 20% of the training set as the validation set to validate the training model. The model with the best overall accuracy is kept for evaluation on the test set. The model is trained by the Adam optimizer in 30 epochs, where the learning rate (lr), $beta1$ and $beta2$ are set as 10^{-3}, 0.9 and 0.999 respectively. Moreover, ReduceLROnPlateau of Callback in Keras is implemented to tune the lr dynamically. The lr would be reduced to half of it when the validation accuracy does not increase within 3 epochs. We choose the categorical cross-entropy as the loss function of the model. In addition, batch sizes of 32, 64, 128 and 256 are tested to determine the final batch size of 64 for training. Furthermore, we evaluate the performance of proposed model using the accuracy (ACC) and Cohen's kappa coefficient (K), which are defined as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{1}$$

$$K = \frac{\frac{\sum_{i=1}^{n} x_{ii}}{N} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}. \tag{2}$$

Where TP, FP, TN and FN represent the true positive, false positive, true negative and false negative respectively. The N is the number of 90s epoch in the test set, n is the number of classes. x_{ii} represents the diagonal value of the confusion matrix. It is noteworthy that we optimize the hyper-parameters of our model on the SHHS-100 database. Once obtaining the optimal model, there is no need to tune the architecture and hyper-parameters on the Sleep-EDF and Sleep-EDF-v1 datasets.

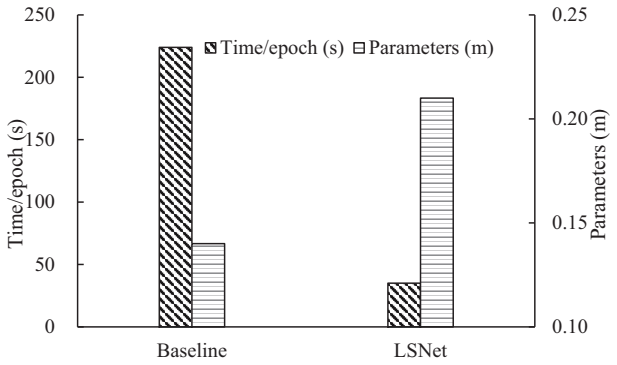


Fig. 2. The comparison between the baseline and LSNet in terms of the model parameters and computational cost for each epoch. The black diagonal stripe represents the training time of each iteration, the gray cross stripe denotes the number of model parameters.

III. EXPERIMENTAL RESULTS

To show the efficiency of spectrogram input, Fig.2 demonstrates the number of parameters and training computational cost for each iteration in the baseline and LSNet models. The number of parameters of the baseline is about 0.14 million (m) and time cost in each iteration ups to 224 s. By contrast, even if the LSNet has more parameters (around 0.21 m), the training speed is more than 6 times faster (35 s) for each epoch.

In Table III, we further make the performance comparison between our framework (epoch-wise and subject-wise) and other state-of-the-art methods using the same dataset across the ACC and K. The values of ACC are more than 83% on all datasets, which show the proposed LSNet model can achieve comparable performance compared to the existing ones. More importantly, the number of parameters of LSNet is much less than that of compared models. Besides, the K demonstrates that our model can reach perfect (0.81 to 1) and substantial (0.61 to 0.8) agreement with the sleep experts.

IV. DISCUSSION

In this paper, we propose a lightweight but effective CNN based model for rapid automatic sleep stage classification, named LSNet. Different with taking time series EEG signals as input, the proposed LSNet transforms those dynamic data to static spectrograms. The proposed LSNet also employs a light structure with few parameters compared to over-parameterized CNN models which lead to the proposed model could be trained in a more efficient way. As the result shows that even though there are roughly 1.5 times as many model parameters of the LSNet as of the baseline model, our model can realize rapid sleep stage classification with more than 6 times speed promotion.

On the other hand, among the previous studies that we compare, the numbers of model parameters in Sors *et al* [11], Zhang *et al* [12], Supratak *et al* [13], [14] and Mousavi *et al* [15] are about 2.2 m, 1.3 m, 1.3 m, 21 m and 2.6 m respectively, which are at least 6 times larger than our model

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD (LSNET) AND PREVIOUS METHODS ON THE SHHS, SLEEP-EDF AND SLEEP-EDF-V1 DATASETS

Study	Dataset	Method	Input channel	Input type	Parameters ($\times 10^6$)	Subjects	$ACC(\%)$	$K(\%)$
Proposed (epoch-wise)	SHHS-100	Deep CNN	C4-A1	Spectrogram	0.2	100	86.7	81.3
Proposed (subject-wise)	SHHS-100	Deep CNN	C4-A1	Spectrogram	0.2	100	85.6	79.4
Sors *et al*. [11]	SHHS	Deep CNN	C4-A1	Time series	2.2	5728	**87**	81
Seo *et al*. [16]	SHHS	CNN + LSTM	C4-A1	Time series	-	5791	86.7	79.8
Zhang *et al*. [12]	SHHS	CNN + LSTM	2EEG + 2EOG + EMG	Spectrogram	1.3	5793	**87**	**82**
Proposed (epoch-wise)	Sleep-EDF	Deep CNN	Fpz-Cz	Spectrogram	0.2	78	**83.7**	**77.5**
Proposed (subject-wise)	Sleep-EDF	Deep CNN	Fpz-Cz	Spectrogram	0.2	78	83.4	76.7
Supratak *et al*. [13]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	1.3	78	83.1	77
Mousavi *et al*. [15]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	21	78	80.0	73
Proposed (epoch-wise)	Sleep-EDF-v1	Deep CNN	Fpz-Cz	Spectrogram	0.2	20	**88.3**	**84.5**
Proposed (subject-wise)	Sleep-EDF-v1	Deep CNN	Fpz-Cz	Spectrogram	0.2	20	86.1	81.0
Supratak *et al*. [14]	Sleep-EDF-v1	CNN + LSTM	Fpz-Cz	Time series	21	20	82.0	76
Supratak *et al*. [13]	Sleep-EDF-v1	CNN + LSTM	Fpz-Cz	Time series	1.3	20	85.4	80
Mousavi *et al*. [15]	Sleep-EDF-v1	CNN + LSTM	Fpz-Cz	Time series	2.6	20	84.3	79

(0.21 m). We do not list the number of model parameters in [16] as it cannot be calculated from the literature. Besides, we do not sacrifice the model performance to achieve the purpose of reducing the model parameters. Experimental results show that our model can attain similar performance compared to the state-of-the-art methods on adopted datasets, which indicate the desirable generalization of the LSNet model. Considering the computing resources and time delay for real-time application, the lightweight model we propose for rapid sleep stage classification maybe more easily adaptable to clinical or wearable devices applications. In future works, we will explore more brain-inspired models (e.g., spiking neural networks) [17], [18] to realize energy-efficient implementation on sleeping scoring tasks.

ACKNOWLEDGMENT

This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä for his great help.

REFERENCES

[1] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[2] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal," *IEEE J. Biomed. Health. Inf.*, vol. 18, no. 6, pp. 1813–1821, 2014.

[3] R. Yan *et al.*, "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods.*, vol. 348, p. 108971, 2021.

[4] H. Phan *et al.*, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018.

[5] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2017.

[6] A. Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.

[7] G. Zhang *et al.*, "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[8] B. Kemp *et al.*, "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.

[9] C. Iber *et al.*, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, vol. 1. Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[10] D. Zhou *et al.*, "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *bioRxiv*, 2021.

[11] A. Sors *et al.*, "A convolutional neural network for sleep stage scoring from raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 42, pp. 107–114, 2018.

[12] L. Zhang *et al.*, "Automated sleep stage scoring of the sleep heart health study using deep neural networks," *Sleep.*, vol. 42, no. 11, p. zsz159, 2019.

[13] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 641–644, 2020.

[14] A. Supratak *et al.*, "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.

[15] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE.*, vol. 14, no. 5, pp. 1–15, 2019.

[16] H. Seo *et al.*, "Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 61, p. 102037, 2020.

[17] Q. Xu *et al.*, "Deep CovDenseSNN: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment," *Neural Netw.*, vol. 121, pp. 512–519, 2020.

[18] Z. Yu *et al.*, "Emergent inference of hidden markov models in spiking neural networks through winner-take-all," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1347–1354, 2018.

PIII

CONVOLUTIONAL NEURAL NETWORK BASED SLEEP STAGE CLASSIFICATION WITH CLASS IMBALANCE

by

Qi Xu, **Dongdong Zhou**, JianWang, Jiangrong Shen, Lauri Kettunen, and Fengyu Cong 2022

2022 International Joint Conference on Neural Networks (IJCNN2022), pp. 1-6, https://doi.org/10.1109/IJCNN55064.2022.9892741

Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance

Qi Xu[a,b,1,*], Dongdong Zhou[c,d,1], Jian Wang[c,d], Jiangrong Shen[e], Lauri Kettunen[d], Fengyu Cong[a,c,d]

[a]*School of Artificial Intelligence, Dalian University of Technology, Dalian, China*
[b]*Guangdong Laboratory of Artificial Intelligence and Digital Economy, Shenzhen, China*
[c]*School of Biomedical Engineering, Dalian University of Technology, Dalian, China*
[d]*Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland*
[e]*College of Computer Science and Technology, Zhejiang University, Hangzhou, China*
Email: xuqi@dlut.edu.cn, dongdong.w.zhou@student.jyu.fi, wangjian009@mail.dlut.edu.cn
jrshen@zju.edu.cn, lauri.y.o.kettunen@jyu.fi, cong@dlut.edu.cn

***Abstract*—Accurate sleep stage classification is vital to assess sleep quality and diagnose sleep disorders. Numerous deep learning based models have been designed for accomplishing this labor automatically. However, the class imbalance problem existing in polysomnography (PSG) datasets has been barely investigated in previous studies, which is one of the most challenging obstacles for the real-world sleep staging application. To address this issue, this paper proposes novel methods with signal-driven and image-driven ways of noise addition to balance the imbalanced relationship in the training dataset samples. We evaluate the effectiveness of the proposed methods which are integrated into a convolutional neural network (CNN) based model. Experimental results evaluated on Sleep-EDF-V1, Sleep-EDF and CCSHS databases demonstrate that the proposed balancing approaches with specific tensity Gaussian white noise could enhance the overall or stage N1 recognition to some degree, especially the combination of two types of Data augmentation (DA) strategies shows the superiority of overall accuracy improvement.**

***Index Terms*—Sleep stage classification, Class imbalance problem, Data augmentation, Time-frequency image**

I. Introduction

Sleep is one of the most important human activities, which makes great contributions to one's mental and physical health and recovery [1], [2]. However, millions of people around the world suffer from different degrees and types of sleep-related issues [3]. It is a time-consuming and labor-intensive procedure to diagnose and treat them, thereinto correct sleep stage classification is an essential step. Clinically, whole-night sleep PSG data, including electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG), electrooculogram (EOG), etc, are divided into 30s epochs with labels of Wake (W), Rapid Eye movement (REM), Non-REM1 (N1), Non-REM2 (N2) and Non-REM3 (N3) by hands [4]. Although large amounts of deep learning methods have been proposed to handle this task automatically [5]–[12], it seems that there is still a gap from real-world implementation, one of possibilities is that the class imbalance problem (CIP) of PSG datasets which has not been paid enough attention and solved well.

In simple terms, the CIP in sleep scoring refers to the duration of each sleep stage is not equal because of the special sleep structure. For instance, stage W and N2 occupy the dominant proportion of samples (more than 60%). By contrast, the N1 stage usually accounts for 2%-5% of overnight sleep time. It is not fair for minority classes when training a deep neural network model with the imbalanced dataset. In such a way, the major categories contribute the leading weight updating, while the contribution of minority ones is biased during the back-propagation. Whether the overall accuracy or the recognition rate is limited by the CIP, which is worth further exploration. As the representation of minority groups, N1 stage suffers from heavy discrimination with the highest misclassification rate. Only a few of works have focused on the solutions for CIP in the sleep scoring. Supratak *et al*. [6] duplicated the minority sleep stages in the training set in which each sleep stage is equally shown. Similarly, Dong *et al*. [13] used oversampling to generate new samples to keep the same percentage of all sleep stages. However, if increasing the number of minority classes in a mechanical way to reach a state that all sleep stags have an equal number of samples, the initial sleep structure was totally destroyed. Fan *et al*. [14] applied five DA methods to assess the enhancement of overall accuracy and N1 classification rate, although the overall performance was improved, the N1 accuracy showed a slight drop sadly.

To remedy the CIP in the sleep scoring task with deep learning based models, we aim to balance the dataset samples by only increasing the number of N1 stage in the original training set with Gaussian white noise addition, this way could retain the original sleep architecture as much as possible. Additionally, we further investigate two categories of Gaussian white noise addition to the EEG signal. One is to add Gaussian white noise to the raw EEG signals (signal-driven) and then transform the noisy EEG signals to time-frequency images. Another one is to convert the EEG signals to time-frequency images then add the noise to the images (image-driven). These two balancing methods are embedded into a CNN based model to show the effectiveness of the relatively balanced state

[1] Equal contribution to this work, * Corresponding author: xuqi@dlut.edu.cn

TABLE I
THE SCHEME OF SIGNAL-DRIVEN AND IMAGE-DRIVEN APPROACHES

Intensity	signal-driven	image-driven
low	10 dB	mean = 0, variance = 0.05
moderate	5 dB	mean = 0, variance = 0.1
high	1 dB	mean = 0, variance = 0.2

between the imbalanced training data and model. Both signal-driven and image-driven balancing methods could improve overall accuracy or N1 accuracy to varying degrees.

The rest of this paper is organized as follows: The Sec.II describes the class imbalance problem and defines the class imbalance factor of PSG datasets. We present the experiments and experimental results in Sec. III and IV, respectively. The final conclusion and discussion are included in Sec.V.

II. CLASS IMBALANCE PROBLEM

Class imbalance is a common yet easily overlooked issue in the sleep stage classification task, the class distribution of the PSG dataset not only depends on the physical or mental conditions but also depends on the ages and genders. When the number of each category is severely unequal, we can say the dataset suffers from the CIP. Here, we define a class imbalance factor (CIF) to quantify the degree of CIP as follow:

$$CIF = \frac{N}{2 \cdot c \cdot \min\{N_i\}} \quad i \in \{1, 2, \ldots, c\} \tag{1}$$

Where the N is the total samples, c refers to the number of sleep stages, and N_i represents the number of each stage. If the $CIF = 0.5$, it means the dataset is balanced. If the $CIF > 0.5$ in eq. (1), that dataset could be regarded as an imbalanced one. Furthermore, the larger CIF means that the PSG dataset is more imbalanced. The CIP mainly affects the training procedure of the deep model which leads to erroneous results in pattern classification tasks. For example, one of the most popular used training rules in deep learning is the back-propagation (BP) algorithm, in which the major classes are responsible for prime parts of weight update. As a consequence, the minority categories become the biased ones with relatively lower recognition rate.

The straightforward way is to increase the number of minority classes to keep equivalent with others [6]. However, the original sleep architecture is broken completely in such a way. Therefore, we only generate new epochs for the N1 stage in training set with the noise addition to maintain the intact sleep structure as far as possible. In this study, we adopt the scheme with a time-frequency image input, which is generally considered as a higher-level representation of the raw signal and can get a faster training speed [11], [15]. Furthermore, we also investigate whether the sequence order of Gaussian white noise addition (i.e., before and after the time-frequency transform) would affect the final result. To be specific, the same type of noise (Gaussian white noise)

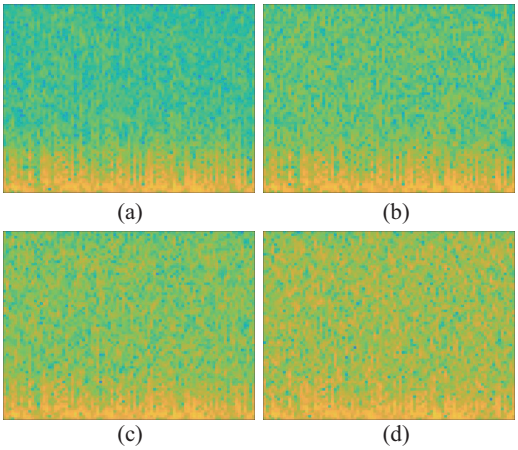


Fig. 1. (a) is the time-frequency image of raw EEG signal (N1 stage, Sleep-EDF-V1), the x-axis represents the time, the y-axis denotes the frequency. Subfigures (b), (c) and (d) illustrate the time-frequency images of raw EEG signal with 10, 5 and 1 dB Gaussian white noise addition respectively.

with three intensities is designed for comparison. The first method is to add the Gaussian white noise with 10, 5 and 1 dB (low, moderate and high intensities) to the raw EEG signal, respectively, then the noisy EEG signals are converted to time-frequency image using the short-time Fourier transform (STFT), it is a signal-driven approach to conduct the noise addition. As a comparison, the second scheme, the image-driven way, adds the similar intensities of Gaussian white noise (the mean (M) is 0, the variances (V) are 0.05, 0.1, 0.2 respectively) to the time-frequency image rather than the raw EEG signal. The scheme of two Gaussian white noise addition methods is demonstrated in Table I.

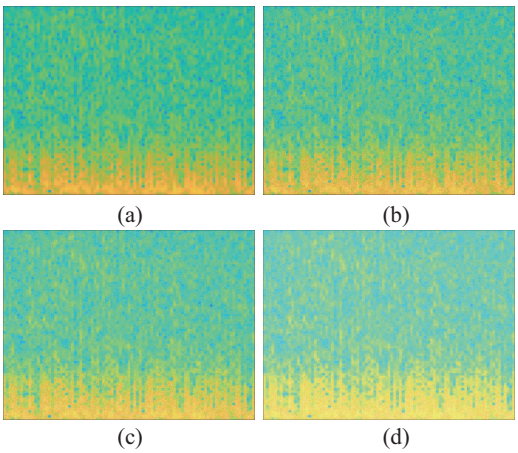


Fig. 2. (a) is the time-frequency image of raw EEG signal (N1 stage, Sleep-EDF-V1) which is the same as Fig. 1. The x-axis represents the time, the y-axis denotes the frequency. (a). Subfigures (b), (c) and (d) present the time-frequency images with three intensities of Gaussian white noise addition (variances are respectively 0.05, 0.1 and 0.2).

When attaining the optimal intensity of two balancing methods, the efficiency of the combination of two proposed

methods is further tested. We visualize the time-frequency images of two noise addition methods in Fig. 1 and Fig. 2.

III. Experiments

A. Experimental Datasets

1) Sleep-EDF-V1: The Sleep-EDF-V1 dataset has two subsets: sleep-cassette (SC) and sleep-telemetry (ST). In this study, we choose the 20 individuals with 39 overnight PSG recordings from the SC cohort, the age ranges from 25 to 34 years. As the suggestion of the American Academy of Sleep Medicine (AASM) manual, the frontal lobes Fpz-Cz channel EEG with a sampling rate of 100 Hz is adopted. More details are described in [16], [17]. The whole PSG recording was labeled with different sleep stages (i.e., W, N1-N4 and REM) based on the Rechtschaffen and Kales (R&K) [18], we merged the stages N3 and N4 into stages N3 for being consistent with the latest AASM standard.

2) Sleep-EDF: The Sleep-EDF dataset is the expanded version, including 78 subjects whose age stretches to 101 years. It has a higher proportion of N1 stages with the increase of age. In order to mitigate the negative impact of the long W stage period on overall accuracy (e.g., stage W has the highest classification accuracy), 30 minutes of W stages before and after regular sleep stages are employed for both versions of Sleep-EDF datasets.

3) CCSHS: The last PSG dataset used in this study is the Cleveland Children's Sleep and Health Study (CCSHS), which includes 515 children aged from 16-19 years. Due the absence of the FPz-Cz, we employ the the C4/A1 (sampled at 128 Hz) channel EEG instead. The main description can be found in [19], [20]. Here, we implement the many-to-one scheme which treats the combination of one 30 s epoch and its neighboring epochs as the contextual input (i.e., 90 s epoch). In Table II, we conclude the number of each sleep stage, the CIF is respectively 6.3%, 6.6% and 2.8% for the Sleep-EDF-V1, Sleep-EDF and CCSHS datasets. Although the sleep stage with the minimum number of Sleep-EDF is different from the other two datasets, we adopt the proposed balancing method only to increase the samples of stage N1 on all experimental datasets.

4) Data preprocessing: In this work, we adopt the STFT with a window size of two seconds and 50% overlap to convert the EEG signal to the image. Firstly, the EEG signal (with/without Gaussian white noise addition) is filtered by a notch filter, a high-pass filter and a low-pass filter in sequence. Hamming window and 256 points Fast Fourier Transform (FFT) [21] are further conducted to obtain the time-frequency image (efficient frequency band: 0.5-30 Hz).

B. Experimental setting

The whole dataset is divided into the training and test sets randomly based on the ratio of 4 to 1 (i.e., 80% subjects as the training set, 20% subjects as the test set). We use the Adam optimizer to train the model within 30 iterations, the model with the best performance in the test set is saved in all epochs. In addition, the learning rate would drop to half value when the test accuracy shows no enhancement within three epochs. The categorical cross-entropy is chosen as the model loss function. To find out a proper batch size, we assess four batch sizes (32, 64, 128 and 256), the batch size of 64 achieves the best performance. In our cases, a workstation with two Inter Xeon E5-2640 V4 CPUs and four Nvidia Tesla P100 GPUs with 16 GB memory is applied to conduct all experiments.

TABLE II
The data distribution of the experimental datasets

Stage	Sleep-EDF-V1	Sleep-EDF	CCSHS
W	10197 (23.1%)	69518(34.9%)	211030 (30.6%)
N1	2804 (6.3%)	21522 (10.8%)	19211 (2.8%)
N2	17799 (40.3%)	69132 (34.7%)	249681 (36.2%)
N3	5703 (13.0%)	13039 (6.6%)	110188 (16.0%)
REM	7717 (17.5%)	25835 (13.0%)	100252 (14.5%)

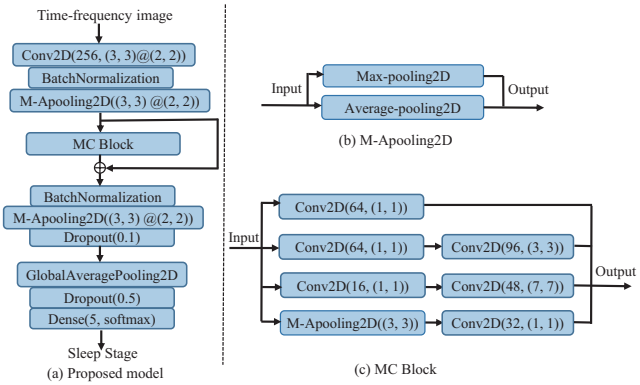


Fig. 3. The overall construct of the evaluation model.

C. The evaluation model

We construct a convolutional neural network based model to assess the efficiency of the proposed balancing method, it is treated as the baseline model (shown as Fig.3). The baseline model is mainly composed of a two-dimensional convolutional (Conv2D) layer, a multi-convolution (MC) block, two Max-Apooling2D layers and several BatchNormalization and dropout layers. The MC block, containing three filter sizes (1×1, 3×3, 7×7), is inspired by the inception module [24] to obtain the multi-scale feature representations. Similarly, we concatenate the outputs of Max-pooling2D and Average-pooling2D layers to rebuild as the Max-Apooling layer. The dropout layer aims to prevent the overfitting problem with a drop rate of 0.1 and 0.5. In addition, the Global Average Pooling (GAP) layer is used to replace the fully connected layer, which is considered more robust spatial translations of the input without parameter optimization [25]. The final dense layer employing the softmax as the activation function is implemented for predicting the sleep stage. We also apply the shortcut connection strategy to combine the input of the

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT INTENSITIES OF THE GAUSSIAN NOISE ADDITION (SIGNAL-DRIVEN AND IMAGE-DRIVEN WAYS) IN THIS WORK

	Sleep-EDF-V1			Sleep-EDF			CCSHS		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Without DA	86.3	81.1	38.9	84.5	79.0	24.6	87.0	82.0	22.9
DA (signal-driven, 10 dB)	85.4	80.0	35.2	84.5	79.0	26.1	87.3	82.4	25.7
DA (signal-driven, 5 dB)	86.8	81.1	42.7	84.3	78.6	18.7	87.2	82.3	24.1
DA (signal-driven, 1 dB)	87.1	82.3	34.8	84.7	79.3	24.0	87.5	82.5	27.3
DA (image-driven, V = 0.05)	87.0	82.1	30.8	84.6	79.0	15.6	87.1	82.1	21.9
DA (image-driven, V = 0.1)	87.0	82.2	34.5	84.6	79.1	21.1	87.3	82.3	22.2
DA (image-driven, V = 0.2)	86.1	80.7	30.3	84.6	79.1	25.9	87.3	82.3	23.0
DA (Combination, 1 dB & V = 0.1)	87.2	82.4	28.5	84.9	79.4	19.1	87.9	82.9	20.7

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHODS AND PREVIOUS METHODS ON THE CCSHS DATASET

Study	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Nakamura *et al*. [22]	HMM	C4/A1 + C3/A2	Spectrogram	515	-	73.0	-
Li *et al*. [23]	Random Forest	C4/A1	Features	116	86.0	80.5	7.3
Baseline	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.0**	**82.0**	**22.9**
DA (signal-driven, 1 dB)	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.5**	**82.5**	**27.3**
DA (image-driven, V = 0.1)	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.3**	**82.3**	**23.0**

MC block with features learned from the MC block, in which 240 filters with size of 1×1 are used to unify the dimension.

IV. EXPERIMENTAL RESULTS

A. Overall performance

We employ the overall accuracy (ACC), Cohen's kappa coefficient (K) and class-wise recall of N1 (RE_N1) to assess the performance. The RE, ACC and K are defined as follows:

$$RE = \frac{TP}{TP + FN}. \tag{2}$$

$$ACC = \frac{\sum_{i=1}^{n} x_{ii}}{N} \tag{3}$$

$$K = \frac{\frac{\sum_{i=1}^{n} x_{ii}}{N} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}. \tag{4}$$

where TP and FN denote the true positives and false negatives respectively, N is the total number of all sleep stages, x_{ii} represents the diagonal value of the confusion matrix, n refers to the number of classes.

Table III illustrates the results of the baseline model without DA methods and different intensities Gaussian white noise addition using two balancing methods. We can see that the 1 dB Gaussian white addition could obtain the most significant improvement of ACC and K for the signal-driven DA on three datasets (Sleep-EDF-V1: ACC-0.8%, K-1.2%; Sleep-EDF: ACC-0.2%, K-0.3%; CCSHS: ACC-0.5%, K-0.5%). In terms of the RE of N1 stage, the 5, 10 and 1 dB achieve 3.8%, 1.5% and 4.4% enhancement from 38.9%, 24.6% and 22.9% on the Sleep-EDF-V1, Sleep-EDF and CCSHS datasets, respectively. Regarding the image-driven DA, the low and moderate intensities (V = 0.05 and 0.1) have the same ACC improvement on Sleep-EDF-V1 and Sleep-EDF datasets. Nevertheless, only the heavy intensity (V = 0.2) gains a gentle enhancement (1.3% and 0.1%) of RE_N1 on the Sleep-EDF and CCSHS databases. It is pleasant that the combination of two intensities (1 dB and V = 0.1) realise the most considerable ACC and K improvement (ACC-0.9%, K-1.3%; ACC-0.4%, K-0.4%; ACC-0.9%, K-0.9%) on the experimental datasets, but an unfavorable decrease in the RE_N1 on three datasets. In addition, two balancing approaches fail to show remarkable distinctions concerning the accuracy improvement with the experimental datasets.

B. Performance comparison

In order to further validate the efficiency of proposed methods, we also compare the overall and N1 accuracies with other works on the same dataset in Tables IV and V. It can be observed in Table IV that the proposed methods can outperform [22], [23] on the CCSHS dataset. Similarly, the baseline model shows better overall accuracy than the performance of [14], [15], [21] on the Sleep-EDF-V1 and Sleep-EDF datasets. Moreover, the performance (i.e., accuracies of all stages and N1) obtain further enhancement with proposed balancing methods.

TABLE V
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHODS AND PREVIOUS METHODS ON THE SLEEP-EDF-V1 AND SLEEP-EDF DATASETS

Study	Database	Method	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [14]	Sleep-EDF-V1	Deep CNN	Time series	20	74.8	66.0	-
Ref. [21]	Sleep-EDF-V1	1-max CNN	Time-frequency image	20	82.6	76	29.9
Ref. [15]	Sleep-EDF-V1	CNN	Spectrogram	20	86.1	81.0	-
Baseline	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**86.3**	**81.1**	**38.9**
DA (signal-driven, 5 dB)	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**86.8**	**81.1**	**42.7**
DA (image-driven, V = 0.1)	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**87.0**	**82.2**	**34.5**
Ref. [26]	Sleep-EDF	CNN + LSTM	Time series	78	80.0	73	-
Ref. [27]	Sleep-EDF	CNN + LSTM	Time series	78	83.1	77	-
Ref. [11]	Sleep-EDF	RNN	Time series	78	84.0	77.8	-
Baseline	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.5**	**79.0**	**24.6**
DA (signal-driven, 10 dB)	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.5**	**79.0**	**26.1**
DA (image-driven, V = 0.2)	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.6**	**79.1**	**25.9**

V. CONCLUSION AND DISCUSSION

The inherent CIP existing in the PSG datasets has hindered the real-world application of automatic sleep scoring models greatly. In this paper, we try to explore the solutions for the CIP in the sleep stage classification procedures. We first define the CIF to quantify the imbalance degree in three common PSG datasets. Two balancing methods are further introduced to mitigate the undesirable effect from the types of noise addition. The first one is to add different intensities of Gaussian white noise to the raw EEG signal, the noisy EEG signals are then converted to the time-frequency images. In this way, extra frequency components could be added to the time-frequency images, it is called the signal-driven way. Another noise addition way is to add the Gaussian white noise to the time-frequency image directly, it is more similar to the implementation in the computer vision field, we name it the image-driven method. Different from previous studies balancing the PSG datasets with equal proportion [6], [13], [14], we argue that it would break the original overnight sleep structure and hide the physiological mechanism related to sleep. By contrast, we only increase the number of the minority class (N1 stage in this study) that we intend to improve to keep consistent with the test set as much as possible. The proposed methods are validated on a CNN based model with three public PSG datasets.

According to the experimental results, although there is no fixed intensity Gaussian white noise suitable for the enhancement of ACC, K and the recognition of N1 stage on experimental PSG datasets, the overall and N1 stage classification rate could be improved with different intensities. In addition, two DA methods do not show significant differences regard to the improvement of model performance. It can be inferred that it should be tailored to adopt the different intensities and types of Gaussian white noise addition based on the practical results on different properties of PSG datasets. In future work, we will explore more data argumentation methods to deal with the CIP of PSG datasets. Except for balancing the samples, how to balance the deep network is another aspect that can be considered.

ACKNOWLEDGMENT

This work was supported by National Key R&D Program of China National (No.2021ZD0109803), Natural Science Foundation of China (No.91748105), National Foundation in China (No. JCKY2019110B009, 2020-JCJQ-JJ-252), the Fundamental Research Funds for the Central Universities [DUT20LAB303, DUT20LAB308, DUT21RC(3)091] in Dalian University of Technology in China, Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ, No. GML-KF-22-11), CAAI-Huawei Mindspore Open Fund (CAAIXSJLJJ-2021-003A) and the Scholarships from China Scholarship Council (No.201806060164, No.202006060226). This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä. We also thank Prof. Hämäläinen Timo from University of Jyväskylä for his great help in this study.

REFERENCES

[1] Pierre Maquet, "The role of sleep in learning and memory," *Science.*, vol. 294, no. 5544, pp. 1048–1052, 2001.

[2] Raffaele Ferri, Mauro Manconi, Plazzi, et al., "A quantitative statistical analysis of the submentalis muscle emg amplitude during sleep in normal controls and patients with rem sleep behavior disorder," *J. Sleep Res.*, vol. 17, no. 1, pp. 89–100, 2008.

[3] Vijay Kumar Chattu, MD Manzar, Soosanna Kumary, et al., "The global problem of insufficient sleep and its serious public health implications," in *Healthcare*. Multidisciplinary Digital Publishing Institute, 2019, vol. 7.

[4] Conrad Iber, Sonia Ancoli-Israel, Andrew L Chesson, et al., *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, vol. 1, Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[5] Rui Yan, Fan Li, Dongdong Zhou, et al., "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods.*, vol. 348, pp. 108971, 2021.

[6] Akara Supratak, Hao Dong, Chao Wu, et al., "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.

[7] Stanislas Chambon et al., "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.

[8] Dongdong Zhou, Jian Wang, Guoqiang Hu, et al., "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 75, pp. 103592, 2022.

[9] Huy Phan, Fernando Andreotti, Navin Cooray, et al., "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018.

[10] Wei Qu, Zhiyong Wang, Hong Hong, et al., "A residual based attention model for eeg based sleep staging," *IEEE J. Biomed. Health. Inf.*, vol. 24, no. 10, pp. 2833–2843, 2020.

[11] Huy Phan, Oliver Y Chén, Minh C Tran, et al., "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[12] Rui Yan et al., "A deep learning model for automatic sleep scoring using multimodality time series," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 1090–1094.

[13] Hao Dong, Akara Supratak, Wei Pan, et al., "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2017.

[14] Jiahao Fan, Chenglu Sun, Chen Chen, et al., "Eeg data augmentation: towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, pp. 056017, 2020.

[15] Dongdong Zhou, Qi Xu, Jian Wang, et al., "Lightsleepnet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2021, pp. 43–46.

[16] Ary L Goldberger, Luis AN Amaral, Leon Glass, 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.

[17] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, et al., "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.

[18] Allan Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, pp. 644, 1969.

[19] Guo-Qiang Zhang, Licong Cui, Remo Mueller, et al., "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[20] Carol L Rosen, Emma K. Larkin, H. Lester Kirchner, et al., "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity," *J. Pediatr.*, vol. 142, no. 4, pp. 383–389, 2003.

[21] Huy Phan, Fernando Andreotti, Navin Cooray, et al., "Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2018, pp. 453–456.

[22] Takashi Nakamura, Harry J Davies, and Danilo P Mandic, "Scalable automatic sleep staging in the era of big data," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2019, pp. 2265–2268.

[23] Xiaojin Li, Licong Cui, Shiqiang Tao, et al., "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, 2017.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, et al., "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR).* 2015, pp. 1–9, IEEE.

[25] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv Preprint. arXiv:1312.4400,*, 2013.

[26] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya, "Sleepeeg-net: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE.*, vol. 14, no. 5, pp. 1–15, 2019.

[27] Akara Supratak and Yike Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2020, pp. 641–644.

PIV

ALLEVIATING CLASS IMBALANCE PROBLEM IN AUTOMATIC SLEEP STAGE CLASSIFICATION

by

Dongdong Zhou, Qi Xu, JianWang, Hongming Xu, Lauri Kettunen, Zheng Chang, and Fengyu Cong 2022

IEEE Transactions on Instrumentation and Measurement, 75, 1-12, https://doi.org/10.1109/TIM.2022.3191710

Alleviating Class Imbalance Problem in Automatic Sleep Stage Classification

Dongdong Zhou, Qi Xu*, Jian Wang, Hongming Xu, Lauri Kettunen, Zheng Chang, *Senior Member, IEEE*, and Fengyu Cong, *Senior Member, IEEE*

***Abstract*—For real-world automatic sleep stage classification tasks, various existing deep learning based models are biased towards the majority with high proportion. Because of the unique sleep structure, most of the current polysomnography datasets suffer an inherent class imbalance problem (CIP), in which the number of each sleep stage is severely unequal. In this study, we first define the class imbalance factor (CIF) to describe the level of CIP quantitatively. Afterwards, we propose two balancing methods to alleviate this problem from the dataset quantity and the relationship between the class distribution and the applied model respectively. The first one is to employ the data augmentation (DA) with the generative adversarial network (GAN) model and different intensities Gaussian white noise to balance samples, thereinto, Gaussian white noise addition is specifically tailored to deep learning based models, which can work on raw electroencephalogram (EEG) data while preserving their properties. In addition, we try to balance the relationship between the imbalanced class and biased network model to achieve a balanced state with the help of class distribution and neuroscience principles. We further propose an effective deep convolutional neural network (CNN) model utilizing bidirectional Long Short-Term Memory (Bi-LSTM) with single-channel EEG as the Baseline. It is used for evaluating the efficiency of two balancing approaches on three imbalanced polysomnography datasets (CCSHS, Sleep-EDF and Sleep-EDF-V1). The qualitative and quantitative evaluation of experimental results demonstrates that the proposed methods could not only show the superiority of class balancing through the confusion matrix and class-wise metrics, but also get better N1 stage and whole stages classification accuracies compared to other state-of-the-art approaches.**

***Index Terms*—Sleep stage classification, Class imbalance problem, Deep neural network, Data augmentation, Generative adversarial network, Network connection.**

This work was support by National Key R&D Program of China (No.2021ZD0109803), National Natural Science Foundation of China (No.91748105), Youth Fund of National Natural Science Foundation of China (No.82102135), National Foundation in China (No. JCKY2019110B009, 2020-JCJQ-JJ-252), Fundamental Research Funds for Central Universities [DUT2019, DUT20LAB303, DUT21RC(3)091] in Dalian University of Technology in China, Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No.MMC202104), Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ, No. GML-KF-22-11), CAAI-Huawei Mindspore Open Fund (CAAIXSJLJJ-2021-003A) and the Scholarships from China Scholarship Council (No.201806060164, No.202006060226).

D. Zhou is with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland & Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (e-mail: dongdong.w.zhou@student.jyu.fi)

Q. Xu is with School of Artificial Intelligence, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (corresponding author: xuqi@dlut.edu.cn).

J. Wang and F. Cong are with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China & Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (wangjian009@mail.dlut.edu.cn, cong@dlut.edu.cn).

H. Xu is with School of Biomedical Engineering, Faculty of Electronic and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China (e-mail: mxu@dlut.edu.cn).

L. Kettunen and Z. Chang are with Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (e-mail: lauri.y.o.kettunen@jyu.fi, zheng.chang@jyu.fi).

I. Introduction

CORRECT sleep stage classification with overnight polysomnography (PSG) recordings plays an essential role in diagnosing and treating sleep-related disorders [1]–[3]. The PSG data consist of the EEG, electromyogram (EMG), electrocardiogram (ECG), electrooculogram (EOG), etc [4]. Clinically, the PSG data are divided into sequential 30-second (30s) epochs and then each epoch is labeled as one of the sleep stages by clinicians manually following the guidelines of the Rechtschaffen and Kales (R&K) [5] or the American Academy of Sleep Medicine (AASM) [6]. Regarding the AASM manual, the sleep stages can be defined as Wake (W), Rapid Eye Movement (REM), Non-REM1 (N1), Non-REM2 (N2) and Non-REM3 (N3).

However, it is cumbersome, time-consuming and prone to be subjective errors for the manual approach with visual inspection of PSG recordings [3]. Hence a large body of automatic sleep stages classification methods including the conventional machine learning [7]–[9] and the deep networks [10]–[15] have been proposed. Although these methodologies achieve promising performance in terms of overall accuracy, the inherent class imbalance problem (CIP) of PSG datasets have been barely explored. The class distribution of PSG databases is highly imbalanced on account of the specific sleep architecture. Additionally, the structure of whole-night sleep is greatly related to the subject's physiological and psychological condition and data acquisition environment. Hereinto, the stage N1 is the most challenging to be recognized and regarded as a representative of minority groups which usually accounts for 2%-5% of total sleep time, and the N1 stage plays the role of indicator in some sleep disorders. Typically, stage N1 would start within minutes of going to sleep, whereas insomnia may delay the beginning of the N1 stage. Moreover, people who have insomnia show a higher proportion of the N1 stage [16]. Besides, the sufferer with apnea may experience abnormal breathing during sleep, which would awaken the brain from deeper sleep. This could lead to an increase in stage N1 [17]. The N1 stage is also highly related to narcolepsy

[18]. Considering the importance of stage N1 recognition, the high misclassification rate of N1 has tremendously limited the practical application of automatic sleep stages classification approaches.

Only a few literature attempt to address the CIP in the sleep stage classification task. Sun *et al*. [19] introduced a DA approach employing the synthetic minority oversampling technique algorithm. A range of 8–14 dB white noise were added to enable the equal number of each sleep stage. It can be more appropriate to apply the DA to stage N1 rather than all sleep stages to maintain original structure of whole-night sleep maximally. In addition, the scope of signal noise ratio (SNR) of white noise could be extended to investigate the efficiency of different intensities noise. Tsinalis *et al*. [20] used class-balanced random sampling across sleep stages to avoid biased performance on the side of the most representative sleep stages and significantly improve the recall of the stage N1. But the overall accuracy achieved was 78%, which is not good enough compared to other state-of-the-art methodologies. One important reason is that the class-balanced random sampling diminished the importance of major classes making the primary contribution to classification performance. It should be noted that keeping a sensible equilibrium among different distribution classes. Fan *et al*. [21] investigated the efficiency of five DA approaches for sleep EEG signals. New training datasets were created with each class equals in number by means of DA algorithms. The overall classification performance was improved, nevertheless, the stage N1 showed a slight drop in terms of F1 scores. Apart from applying DA methods to balance the class distribution of PSG datasets, the correlation between categories and the trained model should not be ignored. CIP poses a big challenge for the prediction model as most machine learning or deep learning algorithms for classification were designed based on the assumption of the same number of samples in each category. The loss weight of each class is equal, which may lead to discrimination against the minority class.

We initially introduce CIP and define the class imbalance factor (CIF) in sleep PSG datasets systematically. To tackle the CIP in the field of automatic sleep stage classification, two solutions are introduced. The first one is to balance the database quantity by means of the DA approaches using the generative adversarial network (GAN) model and Gaussian white noise (GWN) addition, which increases the number of the N1 stage in the training set. The second method is to balance the relationship between the trained model and the original imbalanced dataset through setting different class weights (CW) in the loss function. To assess the efficiency of DA and CW methods, we further propose an efficient deep model that implements Bi-LSTM and CNNs to extract features across temporal and spatial scales with single-channel EEG simultaneously. In this paper, the proposed model is regarded as the Baseline, the proposed framework with the DA of GAN model, the DA of Gaussian white noise and CW are named the Baseline + GAN , the Baseline + GWN and the Baseline + CW, respectively. The main contributions of this work are summarized as follows:

i) We systematically analyze the class imbalance problem in PSG datasets. Furthermore, we propose two solutions to tackle the CIP from the database quantity and the correlation between classes and the applied model.
ii) We explore the GAN model and the method with Gaussian white noise addition to balance the PSG dataset samples. We further search for the balanced network connection from the perspectives of class distribution and neurology.
iii) We develop a novel model that utilizes one convolution block and two multi-convolution (MC) blocks with different filter sizes as the spatial feature extractor. Another temporal feature extractor consisting of one CNN and Bi-LSTM can learn the information of sleep stage transition rules.
iv) The overall performance and recognition of the N1 stage could be improved to different extents by proposed methods on three public datasets.

The rest of this paper is organized as follows. We demonstrate the experimental datasets and methodologies in Sec. II. In Sec. III, the experimental results are represented. The final discussion and conclusion are included in Sec. VI and Sec. V.

II. Materials and Methods

A. Data Description

We employ three public PSG datasets in this study: Cleveland Children's Sleep and Health Study (CCSHS) [22], [23], Sleep-EDF Database (Sleep-EDF-V1, version 2013) and Sleep-EDF Database Expanded (Sleep-EDF, version 2018) [24]. As the recommendation of the AASM manual, the central and frontal lobes are used. More specifically, C4/A1 and Fpz-Cz EEG channels are selected from the CCSHS and Sleep-EDF datasets respectively.

The CCSHS database is one of the largest pediatric cohorts, including 515 children whose ages range from 16-19 years. In our experiments, C4/A1 channel EEG signals sampled at 128 Hz are used. Each 30s epoch was labeled by trained-well sleep experts.

There are two subsets: sleep-cassette (SC) and sleep-telemetry (ST) in the Sleep-EDF dataset (Sleep-EDF-V1). We use 39 whole-night PSG recordings from 20 subjects aged 25 to 34 years in the SC cohort. Each subject has two full night PSG recordings except for subject 13. The number of individuals in SC subset is increased to 78 with 153 over-night sleep recordings in Sleep-EDF Database Expanded (Sleep-EDF). The oldest subject is 101 years. In our study, we employ Fpz-Cz EEG signals with a sampling rate (i.e., f_s) of 100Hz. It is worthy that the resampling method is not applied to restrict the sampling rate, which means our model can be adaptable to different input lengths. Besides, we only adopt 30 minutes of W epochs before and after sleep stages, as there are long W stages at the start and end of the whole-night sleep in Sleep-EDF and Sleep-EDF-V1 datasets. Considering the correlation and dependency between surrounding epochs, we use the many-to-one scheme described in our prior study that combines one 30s epoch with its neighboring epochs (i.e., three sequential 30s epochs) as the 90s epoch [25]. There is 60s overlap between the adjacent 90s epochs and the label

TABLE I
THE NUMBER OF 90S EPOCHS FOR EACH SLEEP STAGE FROM EXPERIMENTAL DATASETS

Stage	CCSHS	Sleep-EDF	Sleep-EDF-V1
W	211030 (30.6%)	69518(34.9%)	10197 (23.1%)
N1	**19211 (2.8%)**	21522 (10.8%)	**2804 (6.3%)**
N2	249681 (36.2%)	69132 (34.7%)	17799 (40.3%)
N3	110188 (16.0%)	**13039 (6.6%)**	5703 (13.0%)
REM	100252 (14.5%)	25835 (13.0%)	7717 (17.5%)
Total	690372	199046	44220
CIF	3.6	1.5	1.6

TABLE II
THE NUMBER AND PROPORTION OF N1 STAGE BEFORE AND AFTER DATA AUGMENTATION (GAN MODEL) IN THE TRAINING SET

Status	CCSHS	Sleep-EDF	Sleep-EDF-V1
Before	15721 (2.8%)	19284 (11.2%)	2024 (5.4%)
After	31442 (5.5%)	38568 (20.1%)	4048 (10.3%)

of the 90s epoch is the same as the label from the middle 30s epoch. We show in Table I the number and percentage of 90s epochs for each sleep stage from three datasets in our experiments, the class with the smallest number of samples is labeled in bold. The N1 stage occupies the smallest percentage, which equals 2.8% and 6.3% respectively in CCSHS and Sleep-EDF-V1 datasets. While the proportion of N1 in the Sleep-EDF dataset is 10.8% and the N3 stage has the smallest number of samples. Sleep architecture changes with ages [26], sleep efficiency would decline with the increase of age due to frequent arousals from sleep, these changes result in an increment of N1 stage.

B. Class Imbalance Problem

In computer vision (CV), the equal number of each category of some image datasets (e.g., CIFAR-10 database) can be guaranteed. However, the sleep pattern differs from ages, genders and physical conditions of individuals [26], [27], the sleep PSG database suffers severe CIP with imbalanced class distribution. In other words, some sleep stages occupy the dominant proportion, whereas the other stages become the minority classes. For instance, the number of the N2 stage is several times that of the N1 stage. When training a model, the majority class contributes the leading weight updating and therefore the performance of minority classes is biased with a higher misclassification rate. The severity of CIP is described using the class imbalance factor (CIF), which is calculated as follow:

$$CIF = \frac{N}{2 \cdot c \cdot \min\{N_i\}} \quad i \in \{1, 2, \ldots, c\} \tag{1}$$

Where c is the number of classes, N represents the number of all epochs, N_i refers to the number of epochs of class i. We argue that the dataset suffers CIP when CIF is greater than or equal to 1. The greater the CIF is, the more imbalanced the database is. In this study, the CIF of CCSHS, Sleep-EDF, and Sleep-EDF-V1 datasets are 3.6, 1.5 and 1.6, respectively.

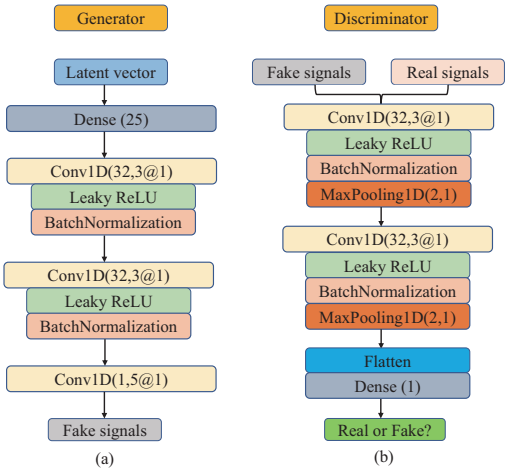


Fig. 1. The framework of the GAN model. Demonstrate the structure: (a) Generator, (b) Discriminator.

In order to alleviate the negative effect of CIP on classification performance, we propose two balancing solutions. The first one is to raise the number of N1 stage with the DA method, which could improve the severity of imbalance to some extent. Another one is to find out the inner network connection between classes and the trained model while maintaining the original dataset quantity, that is to say, setting different class weights for each category depend on the specific class distribution and the neuroscience rule.

C. Balance the Dataset Samples

The imbalanced class distribution has negative effect on the training procedure, which means the applied model could not be trained efficiently. Hence, it is natural and straightforward to increase the number of minority classes to achieve the same proportion, whereas this would break the original architecture of whole-night sleep. By contrast, we choose to produce new epochs of the N1 stage in the training set to maintain the physiologic sleep structure maximally, but the test set is kept independent without balancing sample operation.

The generative adversarial network model has attained significant achievement in the CV field, however, this technology is barely adopted to augment synthetic EEG signals. We use the GAN model as the first method to generate artificial EEG signals of the N1 stage in this study. The GAN is generally comprised of two opposing networks (i.e., generator (G) and discriminator (D)) as shown in Fig. 1. The generator mainly includes three one-dimensional convolutional (Conv1D) layers, thereinto, the first two Conv1D layers are assembled with LeakyReLU (the activation function) and the batch normalization and the last one is used to generate the demanded length signals. In addition, the padding is set as casual to keep the length unchanged. In terms of the discriminator, the Conv1D layer is followed by the LeakyReLU, batch normalization and MaxPooling1D sequentially. The final dense layer makes the prediction for the inputting signal. Given a latent vector

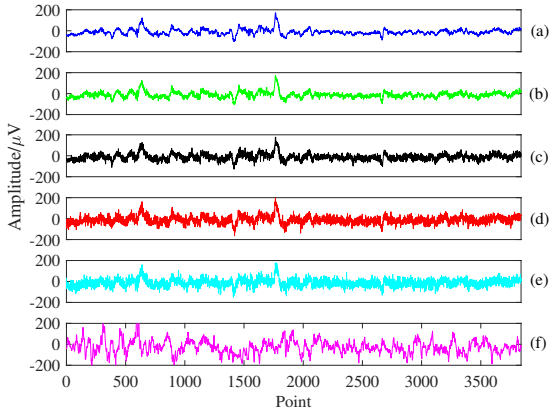


Fig. 2. Raw EEG signal (N1 stage) and Gaussian white noise addition with four SNR. (a) Raw EEG. (b) Gaussian white noise addition with 10 dB. (c) Gaussian white noise addition with 5 dB. (d) Gaussian white noise addition with 2 dB. (e) Gaussian white noise addition with 1 dB. (f) Artificial signal by the proposed GAN model.

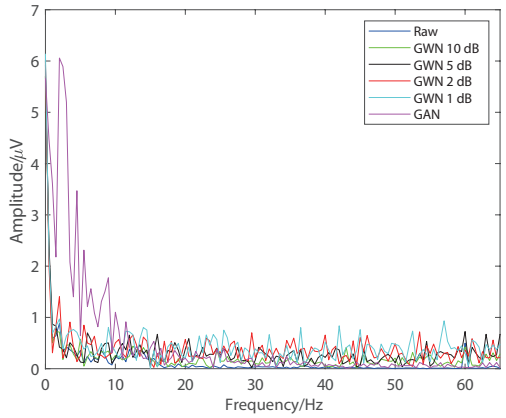


Fig. 3. Spectrogram of raw EEG signal and artificial signals generated by the Gaussian white noise addition with 10 dB, 5 dB, 2 dB, 1 dB and the GAN model.

z following the standard normal distribution ($N(0,1)$), the generator maps it to the input space and learns a distribution $\mathbb{P}_g$ to approach the distribution $\mathbb{P}_{data}$. The discriminator is designed for distinguishing the fake signals generated by the generator and real signals by estimating the correspondence between $\mathbb{P}_g$ and $\mathbb{P}_{data}$. It can be defined as the minimax objective:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\ +\mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}_{\boldsymbol{g}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]\end{aligned} \tag{2}$$

where $D(\boldsymbol{x})$ means the probability of $\boldsymbol{x}$ sampled from the real samples $\mathbb{P}_{data}$. $G(\boldsymbol{z})$ stands for the artificial signals produced by the generator. Additionally, we adopt the loss function presented by Gulrajani *et al.* [28]:

$$\begin{aligned}L\left(\mathbb{P}_{data},\mathbb{P}_g\right) = E_{x_r\sim\mathbb{P}_{data}}[D(x)] \\ -E_{x_g\sim\mathbb{P}_g}\left[D\left(x_g\right)\right] + P(\tilde{x})\end{aligned} \tag{3}$$

$$P(\tilde{x}) = \lambda \cdot E_{\hat{x}\sim\tilde{X}}\left[\max\left(0, \|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] \tag{4}$$

where $P(\tilde{x})$ is defined as the one-sided gradient penalty, λ denotes the the penalty coefficient and $\tilde{X}$ includes points sampling alone the straight line between $\mathbb{P}_{data}$ and $\mathbb{P}_g$. We employ the Adam optimizer to update the model parameters and choose five iterations to train the generator for each iteration of the discriminator. We demonstrate the number and the proportion of N1 in the training set before and after the data augmentation with the GAN model in Table II.

The second method for balancing the dataset samples is the noise addition. Unlike repeating samples of the minority stage directly [11], the data augmentation method with Gaussian white noise addition is implemented in this work for two important reasons. On the one hand, the acquisition of EEG signals always accompanies with noise, a Gaussian noise that imitates the line-related noise that is commonly found in electrophysiology recordings, hence the data generated by noise addition can be more real-like sleep EEG signals. On the other hand, generated data with noise addition can provide the trained model with new features and enhance the generalization. To be specific, we investigate the efficiency of the DA algorithm with four different intensities Gaussian white noise ranging from 1-10 dB. Fig. 2 and Fig. 3 show an example of this DA procedure with different intensities and the DA with GAN model in terms of the amplitude and spectrogram, we can find that these implementations with Gaussian white noise addition retain wave properties of the raw EEG signal. We further explore the effectiveness of various times noise addition. Specifically, once obtaining the optimal intensity (x dB), the intensities of three and five times noise addition are defined as (x - 0.2, x, x + 0.2) dB and (x - 0.2, x - 0.1, x, x + 0.1, x + 0.2) dB with a type of arithmetic progression, respectively. Compared with the way of repeating corresponding times noise addition with x dB, this could provide with the trained model with additional information.

D. Balance Relationship Between the Imbalanced Dataset and Trained Model

The CIP is not only the imbalance of class distribution but also the imbalanced network connection. Although the DA method could mitigate the imbalance of PSG datasets, whether DA with the GAN model or DA with noise addition, the generated data are still fake. More importantly, we could not ignore the corresponding physiological information behind the PSG dataset for real-world application. In other words, it would be more meaningful to achieve the performance improvement without changing the distribution of class. Therefore, another alternative is to balance the network connection between the sample distribution and the trained model with the original imbalance PSG dataset. By default, the weight of each class is the same. As a consequence, the majority class occupies the dominant weight updating with a more considerable length of the gradient component. Furthermore, the performance of the minority classes is prejudiced by the trained model. To eliminate the discrimination, we reassign

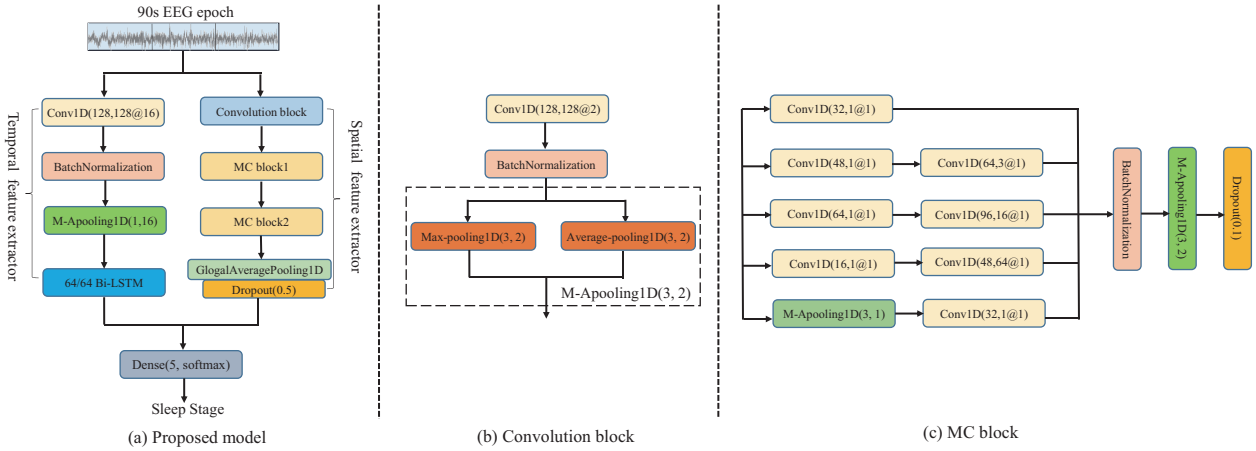


Fig. 4. The schematic diagram of the proposed model.

TABLE III
PARAMETERS OF THE PROPOSED MODEL

Layer	Layer Type	Filters	Size	Stride	Activation	Output dimension
SFE1	Input	-	-	-	-	$(90 \times f_s, 1)$
SFE2	Convolution block	128	128	2	relu	$(\lceil 45 \times f_s/2 \rceil, 256)$
SFE3	MC block1	-	-	-	-	$(\lceil 45 \times f_s/4 \rceil, 544)$
SFE4	MC block2	-	-	-	relu	$(\lceil 45 \times f_s/8 \rceil, 544)$
SFE5	GAP	-	-	-	-	544
SFE6	Dropout (0.5)	-	-	-	-	544
TFE1	Input	-	-	-		$(90 \times f_s, 1)$
TFE2	Conv1D	128	128	16	relu	$(\lceil (90 \times f_s - 128)/16 \rceil, 128)$
TFE3	BatchNormalization	-	-	-	-	$(\lceil (90 \times f_s - 128)/16 \rceil, 128)$
TFE4	M-Apooling1D	-	1	16	-	$(\lceil (90 \times f_s - 128)/256 \rceil, 256)$
TFE5	Bi-LSTM	64	-	-	tanh	128
Decision	Dense	-	-	-	softmax	5

the weight (W) of each class based on the class distribution and the brain-inspired rule, namely CW (Ratio), CW (Log_R) and CW (E_I), respectively. The W_i, W_j using CW (Ratio), CW (Log_R) methods are shown in equations (5), (6):

$$W_i = \frac{N}{N_i} \quad i \in \{1, 2, \ldots, 5\} \tag{5}$$

$$W_j = \ln \frac{N}{N_j} \quad j \in \{1, 2, \ldots, 5\} \tag{6}$$

Where N, N_i and N_j are respectively the numbers of whole classes, class i and j samples. The CW (Ratio) is a direct way to get the W_i by calculating the ratio of the numbers of all samples and each class. Additionally, we attempt a more moderate and sensible approach, the CW (Log_R), to attain the natural logarithm of W_i of the CW (Ratio) method. The CW (E_I) algorithm considers the allocation of neurons during information processing procedures in the human brain [29], [30], namely the ratio of excitatory neurons to inhibitory neurons. Zeng *et al.* [29] investigated the effect of the proportion of inhibitory neurons on the spiking neural networks. As a result, the 15% of inhibitory neurons are the optimal for good performance. Inspired by this brain-inspired rule, we regard the samples of N1 stage as the excitatory neurons, other stages as the inhibitory neurons. To be specific, we set the weight of N1 stage with the value of 8.5, other stages with the weight of 1.5. Three CW methods adopted in this study aim to strengthen the contribution of the minority class and ultimately mitigate the bias towards the majority class.

E. Proposed Model

To evaluate the efficiency of two balancing methods used in this study, we propose a CNN based model for automatic sleep stage classification. The proposed framework is composed of two key parts as illustrated in Fig. 4. The first part is the temporal feature extractor (TFE), which could learn the temporal information (e.g., transition rules between stages). Another part is the spatial feature extractor (SFE) for extracting spatial features. The concatenation of feature maps extracted from the

temporal and spatial feature extractors is fed into the dense layer with the activation function of softmax to make the final decision.

The temporal feature extractor consists of a one-dimensional convolutional (Conv1D) layer, batch normalization, M-Apooling layer and Bi-LSTM layer. The main function of the Conv1D is to attain the feature map from the raw EEG signal. Then the Bi-LSTM is responsible for leaning the temporal information, such as the transition rule between successive stages. Practically, the clinicians decide the next probable stage based on the prior stage on some occasions.

The spatial feature extractor includes four components: a convolution block, two multi-convolution (MC) blocks (inspired by the inception module [31]), a GlobalAveragePooling (GAP) layer and a dropout layer. The convolution block is followed by a Conv1D layer with 128 filters of size 128 and a stride of 2, batch normalization and M-Apooling layer in sequence. Analogously, the MC block comprises different sizes of filters, batch normalization, M-Apooling layer and dropout layer. The purpose of different filter sizes is to capture feature representations in multi-scales. We optimize the filter sizes with small (3, 5 and 7), medium (16 and 32) and large (64, 128 and 256) sizes to adapt to the long input length. In addition, the filter size of 1 is applied to enhance the nonlinearity of the network. The filter sizes are selected with 1, 3, 16 and 64 as they provide the optimal results in our testing. We use the M-Apooling layer, the concatenation of the average-pooling and max-pooling layer, to replace the conventional max-pooling layer in our model. The GAP layer plays the role of the traditional fully connected layer to flat the previous output without introducing extra trainable parameters, which can prevent the overfitting problem efficiently [32]. Table III shows the detailed information of the proposed model, the length of input is $90 \times f_s$, which is related to the sampling rate.

F. Experimental Setup

We divide the whole dataset into the training and test sets randomly based on the subject-wise scheme (i.e., 80% subjects for training, 20% subjects for test). Only recordings from the CCSHS dataset are employed to tune the hyper-parameters of the proposed model. Besides, we choose the Adam as the model optimizer with the algorithm of learning rate (LR) reducing, and the LR would decrease to half of it when the accuracy of test set shows no improvement within three epochs. The value of LR ranges from 10^{-7} to 10^{-3}. In addition, the size of mini-batch is set to 64 chosen from four batch sizes (32, 64, 128, and 256). We select the categorical cross entropy as the loss function, which is always employed for the multi-class model. The number of iteration is 40 as the proposed model could achieve the convergence state within 40 epochs. Furthermore, we save the model with the best test accuracy in all iterations.

To prevent the overfitting problem, we adopt two regularization strategies in this study. The first strategy is the L2 regularization, which adds a squared magnitude of coefficient as penalty term to the loss function. Then we test four regularization rates (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}), and 10^{-3} is adopted finally. The second technology is the dropout that drops units from the model with a probability from 0-1. In the MC block and dropout layer, the probabilities are set to 0.1 and 0.5 respectively.

In our cases, we conduct the experiments on a workstation with two Inter Xeon E5-2640 V4 CPUs and four Nvidia Tesla P100 GPUs with 16 Gbytes memory.

III. Experimental Results

A. Performance Metrics

We use class-wise recall (RE), overall accuracy (ACC) and Cohen's kappa coefficient (K) to evaluate the performance. Similar to the binary classification, we regard each class as a positive class, other classes as a negative class to compute the class-wise metrics. The calculation of RE, ACC and K are shown as follows:

$$RE = \frac{TP}{TP + FN}. \tag{7}$$

$$ACC = \frac{\sum_{i=1}^{n} x_{ii}}{N} \tag{8}$$

$$K = \frac{\frac{\sum_{i=1}^{n} x_{ii}}{N} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}. \tag{9}$$

where TP and FN, respectively, stand for the true positive and false negative, N is the total number of test epochs, c represents the number of classes. In this study, c equals 5, x_{ii} ($1 \leq i \leq 5$) refers to the diagonal value of the confusion matrix.

B. Efficiency of Balancing the Dataset Samples

Table IV illustrates the performance of DA methods with proposed GAN model and different intensities and times Gaussian white noise addition, the bold format stands for the best performance of each index. Compared to the Baseline model, the proposed GAN model can improve the overall accuracy, however, show an slight decrease in terms of the RE of N1 stage (RE_N1) on the experimental datasets. By contrast, the ACC, K and RE_N1 have been enhanced to a different extent on three datasets with the GWN method. Specifically, the RE_N1 has an increase of 9.7%, 16.2%, 12.0% with systems of Baseline + GWN (1 dB), Baseline + GWN (1 dB) and Baseline + GWN (10 dB) on the CCSHS, Sleep-EDF, Sleep-EDF-V1 databases, respectively. In addition, ACC and K are also improved with a range of 0.1% to 2.2%. The improvement of N1 performance (RE_N1) is the priority thing to be considered in the situation of comparable ACC and K. Besides, the enhancement of N1 recognition should not sacrifice the overall performance. Considering the overall and N1 performance, the optimal intensity of Gaussian white noise addition is set as 1 dB. Hence, the intensities of GWN methods with three and five times are respectively set to (0.8, 1.0, 1.2) dB and (0.8, 0.9, 1.0, 1.1, 1.2) dB. Generating more samples of N1 stage could not achieve better overall (ACC and K) and N1 (RE_N1) performance simultaneously compared to the

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED GAN MODEL AND DIFFERENT INTENSITIES AND TIMES GAUSSIAN WHITE NOISE ADDITION IN THIS WORK

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Baseline	88.2	83.8	23.0	86.4	81.1	24.7	85.4	79.9	33.6
Baseline + GAN	88.5	84.3	21.4	86.9	82.0	20.4	86.5	81.5	32.1
Baseline + GWN (1 dB)	88.3	84.0	**32.7**	86.5	81.5	**40.9**	86.3	81.4	44.6
Baseline + GWN (2 dB)	88.3	84.0	30.0	86.7	81.7	28.3	**86.8**	**82.1**	35.9
Baseline + GWN (5 dB)	88.4	84.1	28.4	86.6	81.5	26.7	86.1	80.9	34.6
Baseline + GWN (10 dB)	88.4	84.0	29.0	**87.0**	**82.2**	32.4	86.0	80.9	45.6
Baseline + GWN (three times)	**88.6**	**84.3**	28.2	86.2	80.9	38.2	85.8	80.8	**49.0**
Baseline + GWN (five times)	88.4	83.9	31.0	86.5	81.6	30.2	85.9	80.9	47.4

TABLE V
THE WEIGHT OF EACH CLASS WITH DIFFERENT CW METHODS

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	CW (Ratio)	CW (Log_R)	CW (E_I)	CW (Ratio)	CW (Log_R)	CW (E_I)	CW (Ratio)	CW (Log_R)	CW (E_I)
W	3.3	1.2	1.5	2.5	0.9	1.5	3.7	1.3	1.5
N1	34.9	3.6	8.5	9.0	2.2	8.5	18.4	2.9	8.5
N2	2.8	1.0	1.5	3.1	1.1	1.5	2.6	1.0	1.5
N3	6.3	1.8	1.5	19.4	3.0	1.5	8.0	2.1	1.5
REM	6.9	1.9	1.5	8.6	2.1	1.5	5.9	1.8	1.5

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT CW METHODS IN THIS WORK

	CCSHS			Sleep-EDF			Sleep-EDF-V1		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Baseline	**88.2**	**83.8**	23.0	86.4	81.1	24.7	85.4	79.9	33.6
Baseline + CW (Ratio)	85.3	80.3	**75.0**	**86.5**	**81.5**	30.4	**87.3**	**82.7**	**42.6**
Baseline + CW (Log_R)	87.8	83.4	51.3	86.3	81.1	33.9	85.9	80.8	36.4
Baseline + CW (E_I)	86.6	81.8	67.7	85.9	80.4	**35.7**	85.8	80.8	34.5

Baseline + GWN (1 dB). It is noteworthy that we do not apply the DA operation to the test set, which means the sleep structure of the test set is not destroyed. Employing more times noise addition stands for the worse consistency of training and test sets, which may hinder the classifier from achieving better performance.

C. Efficiency of Balancing the Network Connection

Table V shows the class weight of the training set using three CW methods from the experimental datasets. To further demonstrate how different CW methods may affect the performance, we make a performance comparison in Table VI. The performance obtained by CW methods differs significantly on three datasets. It can be seen that the RE_N1 shows a dramatic increase by all CW approaches, corresponding to 52.0%, 28.3%, and 44.7% (by CW (Ration), CW (Log_R) and CW (E_I) respectively) on the CCSHS dataset. Nevertheless, ACC and K decrease slightly instead. By contrast, on the Sleep-EDF and Sleep-EDF-V1 databases, ACC and K attain slight improvements except by the CW (Log_R) and CW (E_I) methods on the Sleep-EDF dataset. Additionally, the improvements of RE_N1 is relatively lower than those on the CCSHS dataset.

We show in Fig. 5 the confusion metrics of three datasets utilizing four systems (the Baseline, the Baseline + GAN, the Baseline + GWN, and the Baseline + CW). For both CCSHS and Sleep-EDF datasets, the Baseline + GWN (1 dB) and the Baseline + CW (Log_R) are selected as the optimal decision considering the overall performance and the accuracy rate of N1 stage. Whereas, we choose the Baseline + GWN (1 dB) and the Baseline + CW (Ratio) based on experimental results of the Sleep-EDF-V1 database. We further in Fig. 6 reveal the hypnogram comparison labeled by experts and the predictions of four systems for one subject (ccshs-trec-1800905) of the CCSHS dataset. Fig. 7 demonstrates the distribution of weights in the layer with the largest number of parameters (without and with the CW method). We also calculate the kurtosis and skewness of two weight distributions, the kurtosis and skewness of the weight distribution without and with the

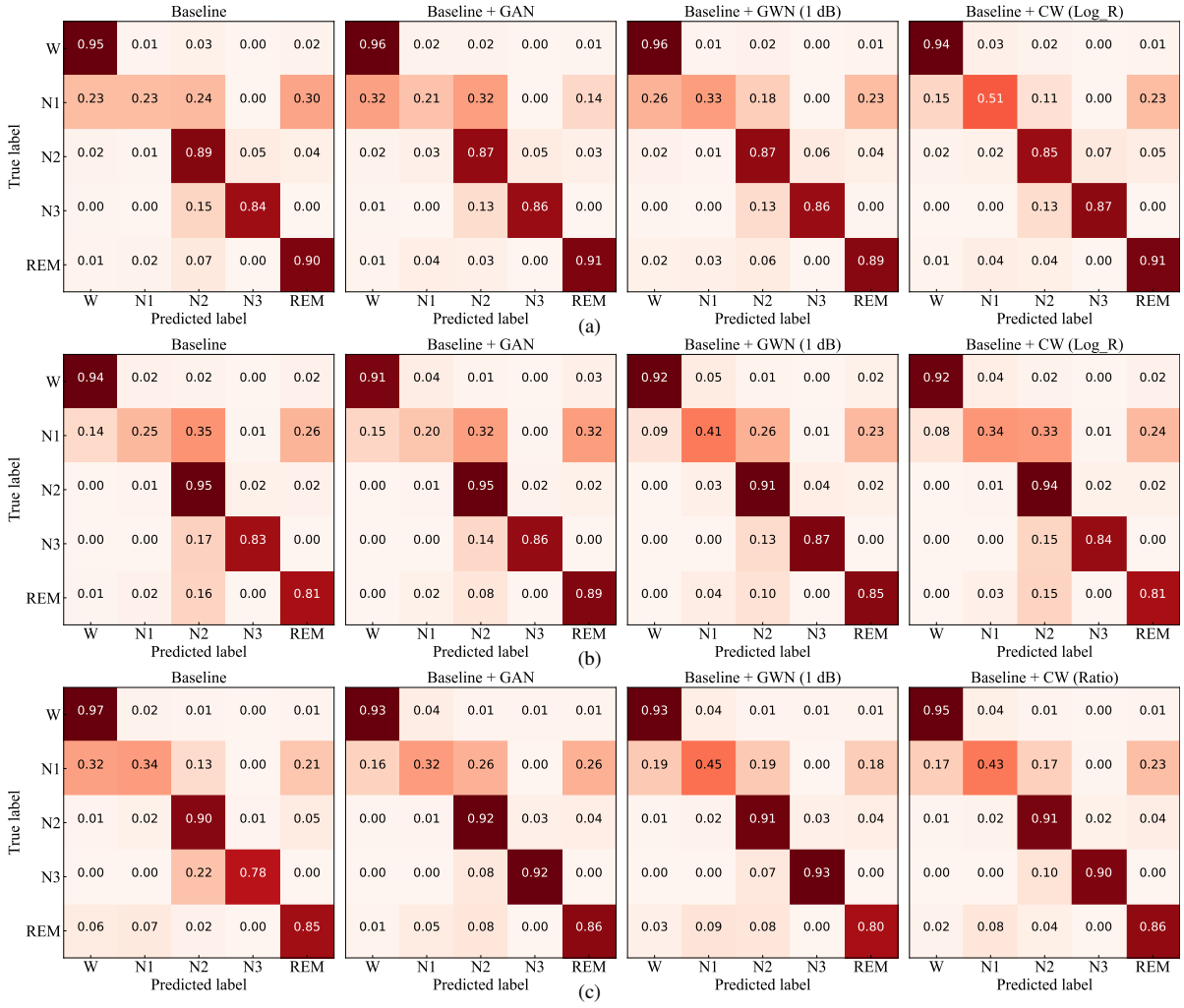


Fig. 5. The confusion matrices of three datasets with four systems. (a) the CCSHS dataset. (b) the Sleep-EDF dataset and (c) the Sleep-EDF-V1 dataset.

CW (Log_R) method are (0.034, -0.414) and (0.094, -0.090), respectively. The CW method more closely resembles a normal distribution (i.e., (0,0)). In such way the network convergence velocity becomes faster [33] and achieving more efficient training procedure for the minority class.

D. Performance Comparison

To see an overall picture, we demonstrate the performance comparison with previous works on the three datasets in Tables VII and VIII. Only a few studies employ the CCSHS dataset, the proposed systems (the Baseline, the Baseline + GWN (1 dB)) could achieve better performance compared to [34], [35]. Similarly, we compare the performance of the Baseline, the Baseline + GWN (1 dB) with [25], [36]–[39] on the Sleep-EDF database, the best ACC, K and RE_N1 are obtained by the Baseline + GWN (1 dB). Those literature [21], [38], [40], [41] utilize the Sleep-EDF-V1 dataset to develop automatic sleep stage classification model, the Baseline + CW (Ratio) framework shows a better ACC, K and a more favorable RE_N1 compared with them.

IV. Discussion

Class imbalance problem is one of the critical factors in real-world automatic sleep stage classification tasks especially using deep learning based models. Here in this paper, we introduce the CIP and define the CIF in the currently common PSG datasets. Correspondingly, this paper introduces two balancing methods to alleviate its negative effect from the dataset quantity and the relationship between the class distribution and the applied model respectively. One is to balance the dataset quantity through increasing the number of samples in the N1 stage, the other aims to balance the relationship between the original imbalanced datasets and deep neural networks while keeping the original dataset quantity. Embedding with two introduced methods, this paper propose a deep convolution neural network based model with Bi-LSTM units for automatic sleep stage classification tasks with single-channel EEG.

In order to enhance the ability of feature extraction, we use the MC block with four sizes of filters to capture spatial features from different scales. The small and large filters are responsible for capturing local features and big context, respectively [10]. In addition, the Bi-LSTM is designed as the

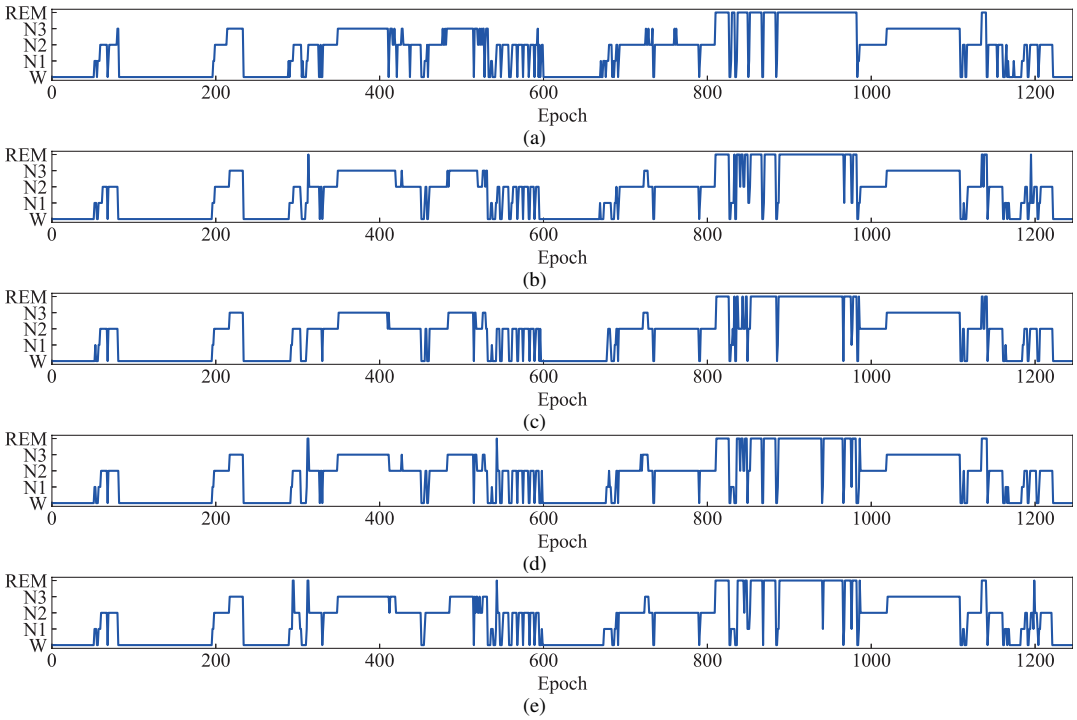
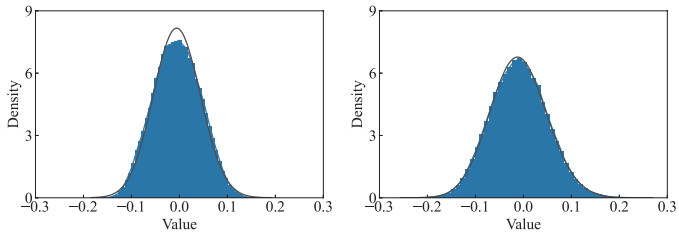


Fig. 6. Hypnogram of one subject from the test set (ccshs-trec-1800905). (a) the ground truth. (b) the prediction of the Baseline. (c) the prediction of the Baseline + GAN. (d) the prediction of the Baseline + GWN (1 dB). (e) the prediction of the Baseline + CW (Log_R).



(a) imbalanced network connection (b) balanced network connection

Fig. 7. The distribution of weights, black lines represent the curve of normal distribution, Y-axis refers to the probability density. (a) without the CW method, (b) with the CW (Log_R) method.

temporal feature extractor to learn the information of sleep stage transition rules. It further enriches features learned from the proposed model. The principle of the DA method here is quite different from previous studies [19], [21], in which the number of each category is designed as the same proportion. However, doing so, the original sleep structure is seriously destroyed. We argue that the physiological correlation between successive sleep stages should not be ignored. That is to say, the initial architecture of whole-night sleep needs to be intact maximally for clinical significance. By contrast, we only increase the number of N1 stage as it is typically considered as the archetype of minority classes with the highest misclassification rate. Different from duplicating selected samples from minority classes in [11], this paper adopts two DA methods with the proposed GAN model and Gaussian white noise addition to generate EEG signals. Although the under-sampling method can also improve the proportion of the minority class and does not need to generate new samples, the evaluation model may suffer from the underfitting problem with the decrease in the training samples. Employing the proposed DA methods, we could not only achieve the goal of increasing the samples of the minority class, but also introduce additional features to enhance the generalization of the applied model. As can be seen from Table IV, the applied GWN method could obtain different degrees of improvement of overall accuracy and recall of N1 stage simultaneously on three datasets compared to those of the baseline model. Nevertheless, the performance of N1 stage showed a slight decrease in [21]. Unlike the image database with independent classes, it is not necessary to keep the equal percentage of each class for mitigating the CIP in PSG datasets. More importantly, we should take into consideration in maintaining inherent characteristics of PSG datasets when employing DA methods. On the other hand, we should develop tailor made DA methods (e.g., different intensities and times noise addition) to deal with the diversity of subjects in different PSG datasets. For instance, both the macro-level (including the sleep stages and duration) and micro-level (such as the quality and quantity of sleep oscillations) structure of sleep would change with the older age [42] and sleep disorders.

Nevertheless, the generated EEG signals by the DA approaches are still artificial. Apart from balancing the class distribution of datasets, another method is to discover the balanced network connection with the original imbalanced dataset. Compared to the DA method, this method could enable the original architecture of sleep and handle general imbalanced PSG datasets. More specifically, we try to balance the relationship between the class and the trained model from the data distribution and the brain-inspired rule. According

TABLE VII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED SYSTEMS AND PREVIOUS METHODS ON THE CCSHS DATASET

Study	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [34]	HMM	C4/A1 + C3/A2	Spectrogram	515	-	73.0	-
Ref. [35]	Random Forest	C4/A1	Features	116	86.0	80.5	7.3
Baseline	**CNN + LSTM**	**C4/A1**	**Time series**	**515**	**88.2**	**83.8**	**23.0**
Baseline + GWN (1 dB)	**CNN + LSTM**	**C4/A1**	**Time series**	**515**	**88.3**	**84.0**	**32.7**

TABLE VIII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED SYSTEMS AND PREVIOUS METHODS ON THE SLEEP-EDF AND SLEEP-EDF-V1 DATASETS

Study	Database	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [25]	Sleep-EDF	CNN	Fpz-Cz	Time series	78	83.9	77.8	-
Ref. [36]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	80.0	73	-
Ref. [37]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	83.1	77	-
Ref. [38]	Sleep-EDF	RNN	Fpz-Cz	Time series	78	84.0	77.8	-
Ref. [39]	Sleep-EDF	CNN	Fpz-Cz	Spectrogram	78	83.4	76.7	-
Baseline	**Sleep-EDF**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**78**	**86.4**	**81.1**	**24.7**
Baseline + GWN (1 dB)	**Sleep-EDF**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**78**	**86.5**	**81.5**	**40.9**
Ref. [21]	Sleep-EDF-V1	Deep CNN	Fpz-Cz	Time series	20	74.8	66.0	-
Ref. [38]	Sleep-EDF-V1	RNN	Fpz-Cz	Time series	20	83.9	77.1	-
Ref. [40]	Sleep-EDF-V1	CNN + LSTM	Fpz-Cz	Time series	20	83.9	78.0	40.0
Ref. [41]	Sleep-EDF-V1	1-max CNN	Fpz-Cz	Time-frequency image	20	82.6	76	29.9
Baseline	**Sleep-EDF-V1**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**20**	**85.4**	**79.9**	**33.6**
Baseline + CW (Ratio)	**Sleep-EDF-V1**	**CNN + LSTM**	**Fpz-Cz**	**Time series**	**20**	**87.3**	**82.7**	**42.6**

to the experimental results demonstrated in Table VI, we conclude some important findings. Firstly, it is essential to keep a sensible equilibrium between minority and majority classes, there is a trade-off between the overall accuracy and the recognition of the N1 stage on the CCSHS dataset, the RE improvement of the N1 stage is accompanied by the sacrifice of ACC and K. Secondly, even the same rule of relationship may result in different results on experimental datasets. The overall and N1 performance could be improved simultaneously on the Sleep-EDF and Sleep-EDF-V1 databases, but much lower enhancement of N1 stage than that on the CCSHS dataset. As mentioned in Sec. II. A, three experimental datasets comprise of subjects from different age groups (CCSHS: 16-19 years, Sleep-EDF: 25-101 years, Sleep-EDF-V1: 25-34 years).

In summary, the CW method is suitable for avoiding generating new EEG samples and keeping the dataset intact for retaining overnight sleep structure. In addition, when recognizing the N1 stage for diagnosing some related sleep disorders, the CW method is prone to show better performance (CCSHS dataset). If we prefer to enhance the performance of all stages and N1 simultaneously, the GWN method can improve the accuracy of the N1 stage without the sacrifice of overall accuracy. In this study, although the GAN model can enhance the overall accuracy, the stage N1 shows a slight drop in recall on three datasets.

V. CONCLUSION

In this study, we aim to deal with the widely existing class imbalance problem in the field of automatic sleep stage classification through balancing the dataset quantity and network connection. The attained results suggest that the proposed methods could make positive contribution to the improvement of biased performance. In most cases, the accuracies of N1 and whole stages are enhanced simultaneously on three public PSG datasets. In addition, our frameworks could outperform the state-of-the-art studies on the same dataset. This study paves new avenues for enhancing the sleep stage classification performance with class imbalance and monitoring the sleep equality and disorders. However, there are some aspects worthy of further exploration in future works. Firstly, more DA methods for balancing the dataset quantity could be investigated, such as the Variational Auto-Encoding network (VAE), which has obtained significant achievements in CV field. In terms of the imbalanced network connection, we will take into consideration of the activation function simulating the operation of neural's synapse for the duration of information processing procedures.

ACKNOWLEDGMENT

This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä for his great help to the authors and Prof. Tapani Ristaniemi has supervised this study very much.

REFERENCES

[1] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[2] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal," *IEEE J. Biomed. Health. Inf.*, vol. 18, no. 6, pp. 1813–1821, 2014.

[3] H. Phan *et al.*, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018.

[4] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2017.

[5] A. Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, p. 644, 1969.

[6] C. Iber *et al.*, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, vol. 1. Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[7] 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.

[8] D. Silveira *et al.*, "Single-channel eeg sleep stage classification based on a streamlined set of statistical features in wavelet domain," *Med. Biol. Eng. Comput.*, vol. 55, no. 2, pp. 343–352, 2017.

[9] P. Memar and F. Faradji, "A novel multi-class eeg-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, 2017.

[10] R. Yan *et al.*, "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods.*, vol. 348, p. 108971, 2021.

[11] A. Supratak *et al.*, "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.

[12] S. Chambon *et al.*, "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.

[13] Q. Wei *et al.*, "A residual based attention model for eeg based sleep staging," *IEEE J. Biomed. Health. Inf.*, vol. 24, no. 10, pp. 2833–2843, 2020.

[14] R. Yan *et al.*, "A deep learning model for automatic sleep scoring using multimodality time series," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1090–1094, 2021.

[15] B. Yang *et al.*, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Trans. Instrum. Meas.*, 2022.

[16] Y. Wei *et al.*, "Sleep stage transition dynamics reveal specific stage 2 vulnerability in insomnia," *Sleep.*, vol. 40, no. 9, 2017.

[17] D. Shrivastava *et al.*, "How to interpret the results of a sleep study," *J. Community Hosp. Intern. Med. Perspect.*, vol. 4, no. 5, p. 24983, 2014.

[18] T. Nakamura *et al.*, "Automatic detection of drowsiness using in-ear eeg," in *Proc Int Jt Conf Neural Netw (IJCNN).*, pp. 1–6, IEEE, 2018.

[19] C. Sun *et al.*, "A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation," *IEEE Access.*, vol. 7, pp. 109386–109397, 2019.

[20] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann Biomed Eng.*, vol. 44, no. 5, pp. 1587–1597, 2016.

[21] J. Fan *et al.*, "Eeg data augmentation: towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, p. 056017, 2020.

[22] G. Zhang *et al.*, "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[23] C. L. Rosen *et al.*, "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity," *J. Pediatr.*, vol. 142, no. 4, pp. 383–389, 2003.

[24] B. Kemp *et al.*, "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.

[25] D. Zhou *et al.*, "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 75, p. 103592, 2022.

[26] B. A. Edwards *et al.*, "Aging and sleep: physiology and pathophysiology," in *Semin Respir Crit Care Med.*, vol. 31, pp. 618–633, 2010.

[27] V. Krishnan and N. A. Collop, "Gender differences in sleep disorders," *Curr Opin Pulm Med.*, vol. 12, no. 6, pp. 383–389, 2006.

[28] I. Gulrajani *et al.*, "Improved training of wasserstein gans," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[29] Y. Zeng, T. Zhang, and B. Xu, "Improving multi-layer spiking neural networks by incorporating brain-inspired rules," *Sci. China Inf. Sci.*, vol. 60, no. 5, pp. 1–11, 2017.

[30] D. J. Heeger and D. Ress, "What does fmri tell us about neuronal activity?," *Nat. Rev. Neurosci.*, vol. 3, no. 2, pp. 142–151, 2002.

[31] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR).*, pp. 1–9, IEEE, 2015.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv Prepr. arXiv:1312.4400.*, 2013.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Conf. Artif. Intell. Stat (AISTATS)*, pp. 249–256, 2010.

[34] T. Nakamura, H. J. Davies, and D. P. Mandic, "Scalable automatic sleep staging in the era of big data," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 2265–2268, 2019.

[35] X. Li *et al.*, "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, 2017.

[36] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE.*, vol. 14, no. 5, pp. 1–15, 2019.

[37] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 641–644, 2020.

[38] H. Phan *et al.*, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[39] D. Zhou *et al.*, "Lightsleepnet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 43–46, 2021.

[40] H. Seo *et al.*, "Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 61, p. 102037, 2020.

[41] H. Phan *et al.*, "Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, pp. 453–456, 2018.

[42] B. A. Mander, J. R. Winer, and M. P. Walker, "Sleep and human aging," *Neuron.*, vol. 94, no. 1, pp. 19–36, 2017.

Dongdong Zhou received the B.S. and M.S. degrees in biomedical engineering from Dalian University of Technology, Dalian, China, in 2015, and 2018, respectively. Supported by China Scholarship Council through the Dalian University of Technology, he is currently pursuing the Ph.D. degree in software and communications engineering with the University of Jyväskylä, Jyväskylä, Finland.

His research interests include biomedical signal processing, sleep analysis, deep learning.

Qi Xu received the B.S. degree from the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China, in 2015, and the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2021. He was ever granted as the honorary Visiting Fellow of the Centre for Systems Neuroscience, University of Leicester, Leicester, U.K., in 2019.

He is curently a tenure-track Associate Professor with the School of Artificial Intelligence, Dalian University of Technology, Dalian, China. His research interests include brain-inspired computing, neuromorphic computing, neural computation, computational neuroscience, biomedical signal processing, sleep analysis and cyborg intelligence.

Jian Wang received the B.S. degree in Department of Japanese from Dalian Neusoft University of Information, Dalian, China in 2011 and the M.S. degree in School of Kinesiology and Health Promotion from Dalian University of Technology, Dalian, China, in 2017. She is currently pursuing the Ph.D. degree in biomedical engineering from Dalian University of Technology, Dalian, China and the visiting doctor of Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland.

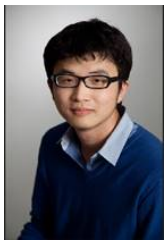
Her research interests include biomedical signal processing, rehabilitation.

Hongming Xu received B.S. and M.S. degrees from College of Inormation Engineering at Northwest A&F University, Yangling, China, in 2009 and 2012, respectively. He received his Ph.D. degree in Department of Electrical and Computer Engineering at University of Alberta in Edmonton, Canada, in 2017.

He is currently an Associate Professor in the School of Biomedical Engineering at Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China. His research focuses on artificial intelligence in biomedical imaging fields (particularly Pathology AI), medical image computing, imaging informatics, and machine learning.

Lauri Kettunen received the Ph.D. degree in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1992.

He is currently a Professor of Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His research interests include mathematical physics, category theory, boundary value problems.



Zheng Chang (Senior Member, IEEE) received Ph.D. degree from the University of Jyväskylä, Jyväskylä, Finland in 2013. He has published over 130 papers in journals and conferences, and received best paper awards from IEEE TCGCC and APCC in 2017 and has been awarded as the 2018 IEEE Best Young Research Professional for EMEA and 2021 IEEE MMTC Outstanding Young Researcher. He has been served as symposium chair, publicity chair and workshop chair and also participated in organizing workshops and special sessions for many IEEE flagship conferences, such as Infocom, ICC and Globecom. He is an editor of Springer Wireless Networks, International Journal of Distributed Sensor Networks, and IEEE Wireless Communications Letters. He was the exemplary reviewer of IEEE Wireless Communication Letters in 2018. He also acts as a guest editor of IEEE Communications Magazine, IEEE Wireless Communications, IEEE Networks, IEEE Internet of Things Journal and IEEE Transactions on Industrial Informatics. His research interests include IoT, cloud/edge computing, security and privacy, vehicular networks, and green communications.

Fengyu Cong (Senior Member, IEEE) received the B.S. degree in power and thermal dynamic engineering and the Ph.D. degree in mechanical design and theory from Shanghai Jiao Tong University, Shanghai, China, in 2002 and 2007, respectively, and the Ph.D. degree in mathematical information technology from the University of Jyväskylä, Jyväskylä, Finland, in 2010.

He is currently a Professor with the School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China and the visiting Professor with Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His current research interests include brain signal processing, acoustic signal processing, independent component analysis, tensor decomposition, and pattern recognition/machine learning/data mining.

PV

INTERPRETABLE SLEEP STAGE CLASSIFICATION BASED ON LAYER-WISE RELEVANCE PROPAGATION

Dongdong Zhou, Qi Xu, Jiacheng Zhang, Lei Wu, Lauri Kettunen, Zheng Chang, Hongming Xu, and Fengyu Cong 2023

Submitted to IEEE Transactions on Cognitive and Developmental Systems, under revision

Interpretable Sleep Stage Classification Based on Layer-wise Relevance Propagation

Dongdong Zhou, Qi Xu*, Jiacheng Zhang, Lei Wu, Lauri Kettunen, Zheng Chang, *Senior Member, IEEE*, Hongming Xu, and Fengyu Cong, *Senior Member, IEEE*

***Abstract*—Many deep learning-based approaches have been proposed for conducting the automatic sleep stage classification tasks. Nevertheless, the black-box nature of these approaches is one of the skeptical factors hindering clinical application. Towards model interpretability, this study presents a novel interpretable sleep stage classification scheme based on layer-wise relevance propagation (LRP). We first adopt the short-time Fourier Transform (STFT) to convert the raw electroencephalogram (EEG) signals to the time-frequency images, which could visually demonstrate EEG patterns of each sleep stage. Moreover, we introduce an efficient convolutional neural network (CNN) based model, namely MSSENet, that assembles with the Multi-Scale CNN module and residual Squeeze-and-Excitation block for the image input. The LRP method is eventually applied to evaluate the contribution of each frequency pixel in the input time-frequency image to the model prediction. Experimental findings show that the MSSENet could exceed other state-of-the-art approaches on three polysomnography (PSG) datasets. Furthermore, through utilizing the heat mapping, the LRP-based explainability results validate the high relevance of specific EEG patterns to the prediction of the corresponding sleep stage, which is consistent with the sleep scoring guidelines.**

***Index Terms*—Sleep stage classification, Model interpretability, Layer-wise relevance propagation, Neural networks.**

I. Introduction

SLEEP is a vitally important physiological activity that makes up about one-third of human life. There is proof that getting good sleep helps human cognitive function [1]. In addition, people have various sleep problems (e.g., apnea, insomnia), which immensely interrupt their daily activities. Therefore, there is an urgent requirement to quantify sleep quality and detect sleep-related disorders precisely.

Sleep stage classification is the initial and fundamental step in evaluating sleep quality and identifying sleep disorders. Clinically, whole-night polysomnography (PSG) data, containing EEG, electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG), etc [2], are first collected in the hospital or laboratory. Afterwards, sleep specialists classify each 30-second epoch of PSG recordings into distinct sleep stages manually, in which the intensive time-cost can be expected. Two gold rules, the Rechtschaffen and Kales (R&K) manual [3] and American Academy of Sleep Medicine (AASM) rule [4], can be referred by the sleep clinician to label the hypnogram. Sleep stages can be classified as six stages, Wake (W), Rapid eye movement (REM), and None rapid eye movement (N1, N2 N3 and N4), with the R&K manual. However, stages N3 and N4 are merged into N3 in the AASM rule.

In recent years, machine and deep learning methods have been applied successfully in various fields [5]–[11]. For the automated sleep stage classification tasks, the deep learning techniques have gained popularity and obtained significant achievements due to their unbelievable power of feature extraction. These works can be summarized as convolutional neural networks (CNNs) [12]–[15], recurrent neural networks (RNNs) [16]–[18], CNN + RNN [19]–[21], etc. In addition, a many-to-one scheme [22] has been further explored to handle the limitation of the short input context. Despite all this progress, the black-box property of deep learning-based models still lacks convincing interpretability and they receive scepticism from sleep experts for the real-world automatic sleep stage classification application. It is a necessary step to clearly explain how the deep model makes decisions and establish trust among the practitioners.

Some studies attempt to give evidence to support the proposed models' sleep staging decision from different aspects. Thereinto, the t-distributed stochastic neighbor embedding (t-SNE) method has been employed to visualize the output of the model layer in numerous automatic sleep stage classification studies [23]–[27]. The t-SNE approach can transform high-dimensional data to two-dimensional data to offer feature visualization of each layer, which can demonstrate the classification results of each sleep stage in each model layer [28]. Yan et al. [25] presented a new deep model integrating the Long Short-Term Memory (LSTM) and CNN for sleep scoring with multimodal time series. The proposed model was then

This work was partially supported by National Natural Science Foundation of China (No.62206037, No.91748105), National Key R&D Program of China (No.2021ZD0109803), Youth Fund of National Natural Science Foundation of China (NO. 82102135), National Foundation in China (No.JCKY2019110B009, No.2020-JCJQ-JJ-252), Fundamental Research Funds for Central Universities [No.DUT2019, No.DUT20LAB303, No.DUT21RC(3)091] in Dalian University of Technology in China, and the Scholarship from China Scholarship Council (No.201806060164).

D. Zhou and F. Cong with School of Biomedical Engineering, Dalian University of Technology, 116024, Dalian, China & Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (e-mail: dongdong.w.zhou@student.jyu.fi, cong@dlut.edu.cn)

Q. Xu is with School of Artificial Intelligence, Dalian University of Technology, 116024, Dalian, China (corresponding author: xuqi@dlut.edu.cn).

J. Zhang is with School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China (e-mail: zhangjc@nuist.edu.cn).

L. Wu is with Department of Neurology, Second Affiliated Hospital of the Zhejiang University School of Medicine, Hangzhou, China (e-mail: wulei301@zju.edu.cn).

L. Kettunen and Z. Chang are with Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland (e-mail: lauri.y.o.kettunen@jyu.fi, zheng.chang@jyu.fi).

H. Xu is with School of Biomedical Engineering, Dalian University of Technology, 116024, Dalian, China (e-mail: mxu@dlut.edu.cn).

visualized using the t-SNE through analysis of the compressed layer outputs. Nevertheless, the features learned by each layer of the applied model were not illustrated. An ablation method is proposed in [29] to validate the importance of each modality data to the applied CNN-based model for the classification of each sleep stage.Each modality's signals were substituted by a mixture of a 40 Hz sinusoid with an amplitude of 0.1 and Gaussian noise with a mean of 0 and a standard deviation of 0.1. To make the performance comparison before and after ablation of each modality, the weighted and individual F1 scores were calculated, respectively. However, the ablation approach can not be efficiently adaptive to the automatic sleep classification models using the single-channel EEG. Moreover, it is still unknown what elements the model captures from the input and whether these feature presentations are related to specific sleep stages. Ellis et al. [30] proposed a new local spectral explainability method to evaluate the importance of different frequency bands over time by perturbing EEG signals. However, using perturbation techniques might result in unrealistic, atypical samples that do not represent a classifier's learning fairly. Moreover, the temporal information of different EEG patterns was discarded using the spectral power.

In this study, we present an explainable scheme to explore the inner connection between the input and prediction of the applied model, which is depicted in Fig.1. First, we acquire time-frequency images containing the EEG patterns information using the short-time Fourier Transform (STFT), which are fed into the proposed model. We further propose a novel CNN-based model assembling with Multi-Scale and Residual Squeeze-and-Excitation block for automatic sleep stage classification. A conservative relevance redistribution method, layer-wise relevance propagation (LRP), is finally applied to detect significant pixels (corresponding to frequency features) in the time-frequency image input that contribute the most to the final layer and receive the most relevance from it. We aim to verify whether specific EEG patterns existing in each sleep stage can be identified properly by the proposed model for making the final decision. The following are the main contributions:

i) We present a novel CNN-based model containing a Multi-scale CNN module for extracting feature presentation from different scales and a Residual Squeeze-and-Excitation block to recalibrate learned features and enhance the performance.
ii) We design an interpretable system to demonstrate the contribution of different frequency bands in the time-frequency image input to the model prediction with the layer-wise relevance propagation method.
iii) Our proposed MSSENet model could achieve remarkable performance on three public datasets and certificate the high relevance of specific EEG patterns to the prediction of the corresponding sleep stage visually.

The remaining of this work is structured as follows: The experimental datasets and methodologies are primarily described in Sec. II. In Section III, we provide the experimental findings. Additionally, Section IV has the discussion and conclusion.

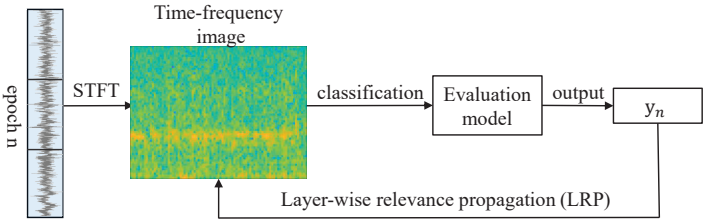


Fig. 1. The overall schematic diagram of the interpretable sleep stage classification with layer-wise relevance propagation.

II. Materials and Methods

A. Data Description

In this paper, we conduct experiments using the following three open PSG datasets:

Cleveland Children's Sleep and Health Study (CCSHS): This database is one of the largest pediatric cohorts for studying objective sleep. The age of 515 subjects starts from 16 to 19 years, and the EEG channel C4/A1 (sampled at 128 Hz) is employed in this study. More details are available in literature [31], [32]. (Data link: https://sleepdata.org/datasets/ccshs).

Sleep-EDF-V1: It is a subset of the Sleep-EDF Expanded dataset called Sleep Cassette (SC), published in 2013, containing 20 subjects between the ages of 25 and 34. Each subject has two consecutive overnight PSG recordings, excluding subject 13 due to the device error. We adopt the Fpz-Cz EEG channel instead as there is no C4/A1 EEG channel. (Data link: https://www.physionet.org/content/sleep-edfx/1.0.0/).

Sleep-EDF: This is the expanded edition of Sleep-EDF-V1 (version 2018), the number of subjects (aged 25-101) increases to 78 with 153 full-night PSG recordings. Each individual receives two consecutive nighttime PSG recordings, with the exception of subjects 13, 36, and 52 on account of the device failure. More detailed information can be found in [33]. We use the Fpz-Cz EEG channel in this study. For Sleep-EDF series datasets, the 30-second EEG epoch sampled at 100 Hz, and the sleep stages are unified as five stages W, N1, N2, N3, and REM for all experimental datasets based on the AASM standard. As we concentrate more on the sleep stages than the wake stage, the samples of 30-minute wake periods are kept prior to and following the sleep periods. [19].

Besides, we implement the many-to-one scheme, in which three consecutive 30-second epochs are reconstructed as the contextual 90-second epoch [22]. Table I provides the details of employed PSG datasets.

B. Data preprocessing

As shown in Table II, we provide the EEG patterns of the five phases of sleep, including Delta, Theta, Alpha, Beta, and others. We aim to validate whether these EEG patterns are also crucial for the proposed model to make the final decision. To display the EEG patterns associated with various sleep stages more visually, we first use the short-time Fourier Transform with a window size of two seconds and 50% overlap to convert the raw EEG signal to the time-frequency image. The time-frequency image is also considered the representation

TABLE I
SPECIFICATIONS OF THE EXPERIMENTAL DATASETS EMPLOYED IN THIS STUDY (EACH EPOCH REFERS TO THE 90-SECOND EPOCH)

Dataset	Subject	EEG channel	Sampling Rate	W	N1	N2	N3	REM	Total
CCSHS	515	C4/A1	128 Hz	211030	19221	249681	110188	100252	690372
Sleep-EDF-V1	20	FPz-Cz	100 Hz	10197	2804	17799	5703	7717	44220
Sleep-EDF	78	FPz-Cz	100 Hz	69518	21522	69132	13039	25835	199046

TABLE II
AN OVERVIEW OF EEG PATTERNS FOR DISTINCT SLEEP STAGES

Stage	Delta (<4 Hz)	Theta (4 - 8 Hz)	Alpha (8 - 13 Hz)	Beta (13 - 30 Hz)	Other EEG Patterns
W			✓	✓	
N1		✓	✓		Vertex waves
N2		✓			K-complexes; Sleep spindles
N3	✓				Sleep spindles may persist
REM		✓	✓		Saw tooth waves

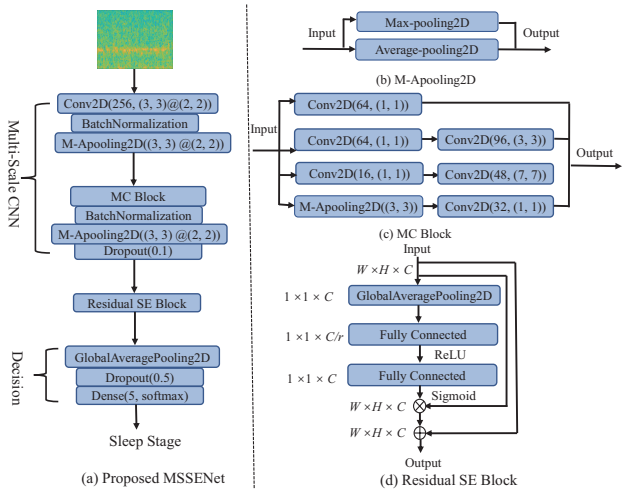


Fig. 2. The structure of proposed MSSENet. a) is an overall of the proposed framework. b) is the M-Apooling, which combines the Max-pooling and Average-pooling. c) illustrates the structure of MC Block, which includes different filter sizes. d) is the SE Block with the shortcut connection strategy.

of the original signal's higher-level features [34]. In addition, Hamming window and 256 point Fast Fourier Transform (FFT) are performed. We retain information from the efficient frequency band of 0.5-30 Hz for the subsequent step analysis.

C. Overall of MSSENet

With the use of time-frequency image input, we design a CNN-based deep model called MSSENet for automatical sleep scoring, which is shown in Fig.2. It includes three main key parts: the Multi-Scale CNN, the Residual Squeeze-and-Excitation (SE) block and the classification part. The number of model parameters of the MSSENet model is around 0.34 million.

1) Multi-Scale CNN (MSCNN): This part mainly comprises one 2-dimensional convolutional (Conv2D) layer and the multi-convolution (MC) block. The original size of the time-frequency image is (656, 857) and then we resize it to (64, 64) as the model input. Feature maps are acquired by the 256 filters with the size of (3, 3), the MC block is then tailored to acquire diversiform feature representations with distinct sizes of filters. As each sleep stage corresponds to different frequency ranges, three different CNN kernel sizes (i.e., 1×1, 3×3, 7×7) are designed to obtain various frequency characteristics (i.e., low and high frequencies). More specifically, the smaller kernel size of (3, 3) is expected to learn the local feature, while the bigger filter size of (7, 7) is prone to capture the big context. Similarly, the combination of Max-pooling and Average-pooling layers, termed the M-Apooing, is desired to further enhance the feature extraction ability [22]. The Conv2D and MC layers are followed by one BatchNormalization layer and a M-Apooling layer, the dropout layer with the drop rate of 0.1 is responsible for preventing the overfitting issue.

2) Residual SE Block (R-SE): To further recalibrate features attained from the Multi-Scale CNN, the Squeeze-and-Excitation block [35] with the residual strategy is used to enhance the performance. As demonstrated in Fig.2.(d), we define feature maps generated from the Multi-Scale CNN as $\mathbf{F} = \{F_1, \ldots, F_C\} \in \mathbb{R}^{W \times H}$, where C is the number of filters. To mine more contextual information outside of a local receptive field in each filter, the global average pooling (GAP) is first used to squeeze global spatial information by shrinking the $\mathbf{F} = \{F_1, \ldots, F_C\} \in \mathbb{R}^{W \times H}$ to $\mathbf{z} = \{z_1, \ldots, z_C\} \in \mathbb{R}^{C}$,

which is shown as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j), \quad c \in \{1, 2, \ldots, C\} \tag{1}$$

The next step is to take advantage of the aggregated information by implementing two fully connected (FC) layers. The first FC layer with the ReLU activation function is responsible for reducing the channel numbers with the reduction ratio r. Another FC layer is used as the dimensionality-increasing layer with the sigmoid activation function. This processing is described in Eq.(2).

$$\mathbf{s} = \sigma\left(\mathbf{W}_2 \delta\left(\mathbf{W}_1 \mathbf{z}\right)\right) \tag{2}$$

where σ and δ stand for sigmoid and ReLU functions respectively, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. The output of the SE block is generated by rescaling $\mathbf{F}$ with $\mathbf{s}$:

$$\mathbf{O} = \mathbf{F} \otimes \mathbf{s} \tag{3}$$

where $\otimes$ denotes the channel-wise multiplication. Additionally, we utilize the shortcut connection technique to merge the $\mathbf{F}$ with the output of the SE block $\mathbf{O}$ as the final output.

3) Decision Part: This part consists of a GAP layer, a dropout layer, and a dense layer. Here, we u substitute the traditional fully connected layer with the GAP layer for flatting the output of Residual SE Block without introducing trainable parameters. Another dropout layer with a drop rate of 0.5 is employed before the final decision. The dense layer calculates the likelihood of each sleep stage using the softmax activation function, and the final forecast is chosen based on the sleep stage with the best probability.

D. Layer-wise relevance propagation

The layer-wise relevance propagation (LRP) is proposed to explore the contribution of each pixel of the input image x to the prediction $f(x)$ when conducting the image classification task [36]. We relate the diagram of LRP method in Fig.3. LRP assumes that the prediction $f(x)$ can be explained by a sum of terms of the separate input dimensions x_d:

$$f(x) = R_f = \sum_d R_d(x) \tag{4}$$

where $R_d(x)$ is the resulting relevance for the pixel x_d of input image x and R_f refers to the relevance of prediction $f(x)$. Note that the sum of relevance of all nodes of each layer should be equal:

$$\sum_d R_d^{(1)} = \ldots = \sum_i R_i^{(l-1)} = \sum_j R_j^{(l)} = \ldots = R_f \tag{5}$$

The relevance can be regarded as information flowing along the network connection, the flow direction is from the output node to the input node. As demonstrated in Fig.3, we can decompose the relevance layer by layer along the sub-paths between nodes referring to the idea of back propagation. Here we always assume that i represents the sequence number of the lower layer neuron, and j denotes the sequence number of the higher layer neuron:

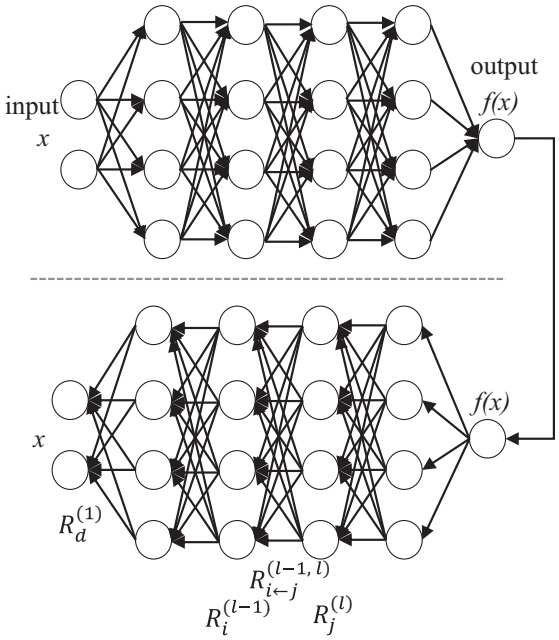


Fig. 3. The diagram of the layer-wise relevance propagation method.

$$R_{i\leftarrow j}^{(l-1,l)} = factor_{ij}^{(l-1,l)} \cdot R_j^{(l)} \tag{6}$$

where $factor_{ij}$ stands for the distribution factor, which belongs to 0-1 and satisfies:

$$\sum_i factor_{ij}^{(l-1,l)} = 1 \tag{7}$$

For any higher layer neuron, its input $z_j^{(l)} = W_j^{(l)} a^{(l-1)}$ and $a^{(l-1)}$ is the activation output vector of the lower layer neuron. Each component $z_{ij}^{(l)}$ of $z_j^{(l)}$ can be considered as the relevance distribution factor between the lower layer neuron i and higher layer neuron j. To satisfy the constraint of Eq.(7), it is divided by a normalization parameter $z_j^{(l)}$ as follows:

$$factor_{ij} = \frac{z_{ij}^{(l)}}{z_j^{(l)}} = \frac{w_{ij}^{(l)} a_i^{(l-1)}}{\sum_i w_{ij}^{(l)} a_i^{(l-1)}} \tag{8}$$

Hence we can rewrite Eq.(6) as follows:

$$R_{i\leftarrow j}^{(l-1,l)} = \frac{w_{ij}^{(l)} a_i^{(l-1)}}{\sum_i w_{ij}^{(l)} a_i^{(l-1)}} \cdot R_j^{(l)} \tag{9}$$

We can map the resulting relevance $R_d(x)$ for each input pixel x_d to color space and visualize it with a conventional heat mapping. In this study, each pixel of input time-frequency images represents the frequency at each time point. We can easily decide whether the EEG patterns (described in Table II) corresponding to a particular sleep stage can be captured and are essential for the model to recognize this specific sleep stage with the help of the heat mapping.

E. Training setup

To acquire the training and test sets, we divided the experimental datasets at random into a 4:1 ratio (i.e., 80% of the subjects served as the training set, while 20% served as the test set). In addition, The categorical cross entropy (CE) and Adam are chosen as the loss function and model optimizer, respectively. The definition of categorical cross entropy is given in Eq. (10), where x stands for the input sample, C presents the class numbers to be categorized, y_i refers to the true label for the class i, and $f_i(x)$ denotes the model's output value.

$$CE(x) = -\sum_{i=1}^{C} y_i \log f_i(x) \tag{10}$$

Adam's learning rate (LR), beta1 and beta2 begin with 10^{-3}, 0.9 and 0.999. To train the model more efficiently, we reduce the LR by half when the test accuracy fails to improve with three epochs. The training iteration is set to 40 since the proposed MSSENet reaches the optimal solution in 40 epochs. Besides, the batch size is set as 64.

III. Experimental Results

A. Performance metrics

We employ per-class and overall metrics to assess the efficiency of the proposed MSSENet model, namely, precision (PR), recall (RE), F1 score ($F1$), overall accuracy (ACC) and Cohen's kappa coefficient (K) [37]. Following are the definition of these metrics:

$$PR = \frac{TP}{TP + FP}. \tag{11}$$

$$RE = \frac{TP}{TP + FN}. \tag{12}$$

$$F1 = 2 \cdot \frac{RE \cdot PR}{RE + PR}. \tag{13}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{14}$$

$$K = \frac{\frac{\sum_{i=1}^{5} x_{ii}}{N} - \frac{\sum_{i=1}^{5}\left(\sum_{j=1}^{5} x_{ij} \sum_{j=1}^{5} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{5}\left(\sum_{j=1}^{5} x_{ij} \sum_{j=1}^{5} x_{ji}\right)}{N^2}}. \tag{15}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative and false negative respectively. The N is the total number of the testing samples. x_{ij} ($1 \leq i \leq 5$, $1 \leq j \leq 5$) is the element of the confusion matrix.

TABLE III
Confusion Matrix of Our MSSENet on C4/A1 EEG Channel from the CCSHS Dataset

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	$PR(\%)$	$RE(\%)$	$F1(\%)$	$ACC(\%)$	$K(\%)$
W	40405	465	785	50	1225	96.1	94.1	95.1		
N1	555	975	565	0	1440	33.0	27.6	30.0		
N2	555	590	40965	2270	2440	87.7	87.5	87.6	87.7	83.3
N3	70	0	2775	19185	25	89.2	87.0	88.1		
REM	460	925	1640	5	18705	78.5	86.1	82.1		

TABLE IV
Confusion Matrix of Our MSSENet on FPz-Cz EEG Channel from the Sleep-EDF-V1 Dataset

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	$PR(\%)$	$RE(\%)$	$F1(\%)$	$ACC(\%)$	$K(\%)$
W	2032	63	7	23	38	93.0	93.9	93.4		
N1	88	295	144	14	161	62.4	42.0	50.2		
N2	38	51	3248	206	134	91.9	88.3	90.1	86.9	82.2
N3	5	0	83	1031	1	80.9	92.1	86.1		
REM	23	64	52	0	1297	79.5	90.3	84.6		

TABLE V
Confusion Matrix of Our MSSENet on FPz-Cz EEG Channel from the Sleep-EDF Dataset

	Predicted					Per-class Metrics			Overall Metrics	
	W	N1	N2	N3	REM	$PR(\%)$	$RE(\%)$	$F1(\%)$	$ACC(\%)$	$K(\%)$
W	7734	227	153	31	355	92.4	91.0	91.7		
N1	430	428	596	18	712	44.1	19.6	27.1		
N2	120	177	11889	598	356	85.6	90.5	88.0	84.3	78.3
N3	20	0	360	3595	0	84.7	90.4	87.5		
REM	62	139	884	3	4395	75.5	80.2	77.8		

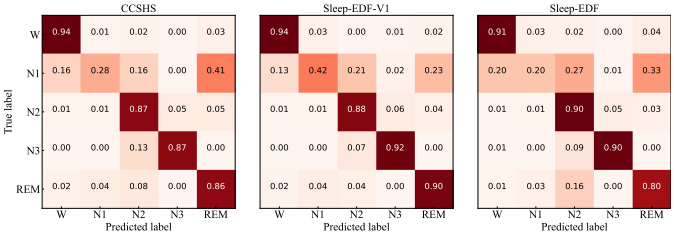


Fig. 4. The normalized confusion matrices of three PSG datasets. x-axis and y-axis represent the predicted and true labels, respectively.

We demonstrate the confusion and performance matrices of the proposed MSSENet implemented on the C4/A1 channel in CCSHS dataset and on the FPz-Cz channel in Sleep-EDF-V1 and Sleep-EDF datasets in Tables III, IV and V. The normalized confusion matrices are also related in Fig.4. The values of each row and collum in aforementioned tables represent the number of each sleep stage labeled by the sleep expert and the proposed model. The corresponding performance metrics are also calculated based on the confusion matrices. We can observe that stage N1 attains the most unfavorable performance, with $F1$ less or close to 50%, and it is prone to be misclassified as REM, N2 and W. By contrast, stage W achieves the most promising performance, with $F1$ more than 90% on three datasets. In terms of the overall performance, ACC of three datasets are more than 84% and CCSHS obtains the highest ACC of 87.7%. In Fig.5, we show the comparison between expert-labeled hypnogram and the model's prediction. The solid blue and dotted red lines represent the ground truth and the prediction of the MSSENet, respectively.

B. Performance comparison

We also compare the overall performance of our MSSENet based on the time-frequency image and the many-to-one

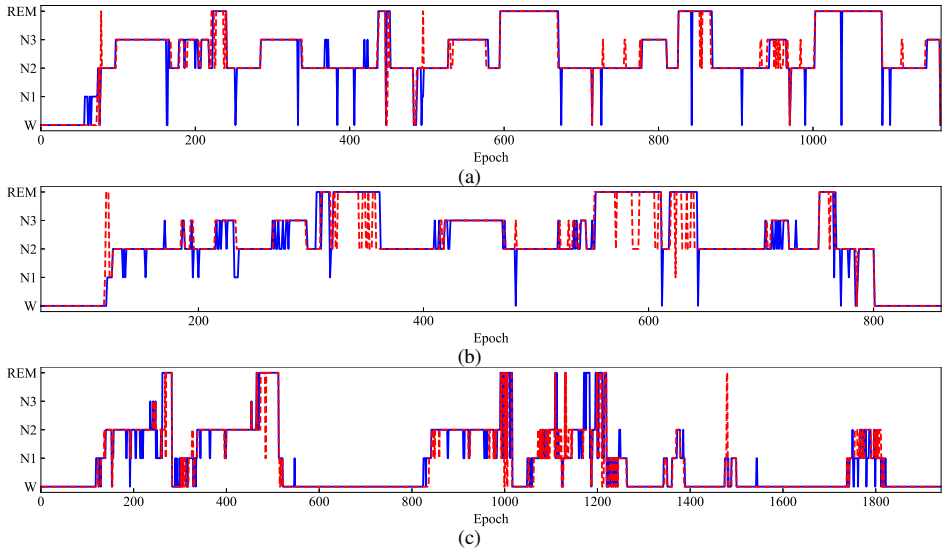


Fig. 5. The comparison between expert-labeled hypnogram and the model's prediction. The solid blue and dotted red lines represent the ground truth and the prediction of the MSSENet, respectively. (a) CCSHS dataset (ccshs-trec-1800399), (b) Sleep-EDF-V1 (SC4112), and Sleep-EDF (SC4622).

TABLE VI
PERFORMANCE EVALUATION OF THE PROPOSED METHOD USING THE CCSHS DATASET COMPARED TO PREVIOUS METHODS

Study	Database	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$
Ref. [38]	CCSHS	HMM	C4/A1 + C3/A2	Spectrogram	515	-	73
Ref. [39]	CCSHS	Random Forest	C4/A1	Features	116	86.0	80.5
MSSENet	**CCSHS**	**Deep CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.7**	**83.3**

TABLE VII
PERFORMANCE EVALUATION OF THE PROPOSED METHOD USING THE SLEEP-EDF-V1 AND SLEEP-EDF DATASETS COMPARED TO PREVIOUS METHODS

Study	Database	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$
Ref. [19]	Sleep-EDF-V1	CNN + LSTM	Fpz-Cz	Time series	20	82.0	76
Ref. [40]	Sleep-EDF-V1	CNN + LSTM	Fpz-Cz	Time series	20	83.9	78
Ref. [41]	Sleep-EDF-V1	Deep CNN	Fpz-Cz	Time series	20	84.3	78
Ref. [42]	Sleep-EDF-V1	1-max CNN	Fpz-Cz	Time-frequency image	20	82.6	76
Ref. [22]	Sleep-EDF-V1	Deep CNN	Fpz-Cz	Time series	20	86.1	80.5
MSSENet	**Sleep-EDF-V1**	**Deep CNN**	**Fpz-Cz**	**Time-frequency image**	**20**	**86.9**	**82.2**
Ref. [18]	Sleep-EDF	RNN	Fpz-Cz	Time-frequency image	78	82.6	76
Ref. [43]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	80.0	73
Ref. [44]	Sleep-EDF	CNN + LSTM	Fpz-Cz	Time series	78	83.1	77
Ref. [13]	Sleep-EDF	CNN	Fpz-Cz	Spectrogram	78	83.4	76.7
Ref. [34]	Sleep-EDF	RNN	Fpz-Cz	Time series	78	84.0	77.8
Ref. [22]	Sleep-EDF	Deep CNN	Fpz-Cz	Time series	78	83.9	77.8
MSSENet	**Sleep-EDF**	**Deep CNN**	**Fpz-Cz**	**Time-frequency image**	**78**	**84.3**	**78.3**

scheme with other state-of-the-art approaches employing the same PSG dataset in Tables VI and VII. We can conclude from Table VI that our MSSENet could achieve better performance (ACC and K) compared to studies [38] employing the multi-modal signal and [39] using the same single-channel EGG (C4/A1). In Table VII, note that all methods use the single-channel EEG (FPz-Cz) with different representation types, our MSSENet could also outperform the CNN-based approaches [22], [41], [42] and the CNN + LSTM frameworks [19], [40] on Sleep-EDF-V1 dataset. Considering Sleep-EDF dataset, the proposed model could also show superiority of ACC and K compared with the method based on the RNN [18], [34], CNN [13], [22], and the combination of CNN and RNN [43], [44].

C. Ablation study

The proposed MSSENet comprises two essential modules: Multi-Scale CNN (MSCNN) and Residual Squeeze-and-Excitation (R-SE) block. To validate the efficiency of each module, we conduct the ablation study on three experimental datasets. Fig.6 illustrates the results of the ablation study. In specific, the MSCNN means the MSCNN module only and the MSCNN + R-SE refers to the MSCNN module and Residual Squeeze-and-Excitation block together (i.e., MSSENet). Although MSCNN + R-SE fails to improve the performance of

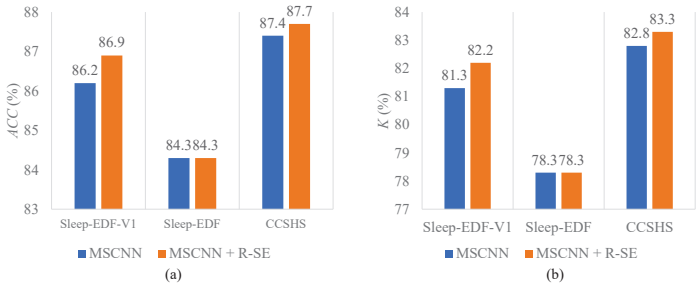


Fig. 6. Ablation study conducted on three PSG datasets. a) is the comparison of ACC, and b) presents the comparison of K.

Sleep-EDF dataset compared to MSCNN only, it can be seen that ACC shows an improvement of 0.7% and 0.3% with the addition of R-SE block on Sleep-EDF-V1 and CCSHS datasets, respectively. Similarly, 0.9% and 0.5% are enhanced on Sleep-EDF-V1 and CCSHS datasets concerning K with the proposed MSSENet.

D. LRP-based explainability results

We reveal the explainability results for each sleep stage using the LRP method in Fig.7. Panel (a) shows the raw EEG signals of five sleep stages and panel (b) illustrates the corresponding time-frequency images through the short-time Fourier Transform, which contains the EEG patterns information. The heat mapping of each sleep stage with the LRP is depicted in panel (c). We can find that the heat mapping of stage W shows high relevance in a high-frequency band (i.e., Beta waves), which is consistent with the main EEG patterns of stage W. In addition, stage N3, as the deep sleep, is characterized by the slowest EEG rhythms (i.e., Delta waves). The LRP-based results show that Delta waves contribute the most to the prediction of stage N3. As for stage N1, the significant contribution of Alpha and Theta waves (4-13 Hz) can also be detected, as illustrated in the heat mapping. From the heat mapping of stage N2, we can observe that the contribution of Theta waves (4-7 Hz) is vital for the model's prediction. During REM sleep, the brain activity closes to the pattern of stage N1, and Theta waves similarly show high relevance to the prediction of stage REM.

IV. Discussion and Conclusion

Model interpretability is one of the critical factors in promoting the practical application of deep learning-based models. This study presents a novel, interpretable scheme for the automatic sleep stage classification task. The main principle is to find the direct correlation between the input and prediction, miming the processing of sleep experts' visual inspection of PSG recordings. To depict the characteristic EEG patterns of each sleep stage more intuitively, we first convert the raw EEG signal to the time-frequency image with the short-time Fourier Transform. Moreover, we propose an efficient CNN-based model, MSSENet, to realize the automatic sleep stage classification with the time-frequency image input. Once the applied model predicts the input time-frequency image, the contribution of each pixel for the input image (i.e., the frequency at each time point) could be encoded to a color space and visualized with a heat mapping using the layer-wise relevance propagation approach. By checking each EEG pattern's contribution to the prediction, we can validate whether the LRP-based explainability results fit with the sleep scoring manuals.

Performance comparison results reveal that our MSSENet could outperform than other state-of-the-art methods employing the same PSG dataset. The Multi-Scale CNN is designed for learning feature presentations from different scales through different kernel sizes of filters. Moreover, we also implement a Squeeze-and-Excitation block with the residual strategy (R-SE) to recalibrate the multi-scale features captured from the Multi-Scale CNN. The efficiency of the recalibration ability of the R-SE block to the multi-scale features is validated on three datasets using the ablation study. Although we do not see the performance enhancement on the Sleep-EDF dataset with the R-SE block, the ablation study on CCSHS and Sleep-EDF-V1 verifies the positive contribution of the R-SE block to performance improvement. The major cause for the disparities in results might be due to the varied data distributions in the three datasets. It should be noted that three experimental datasets include participants of various ages (CCSHS: 16-19 years, Sleep-EDF-V1: 25-34 years, and Sleep-EDF: 25-101 years). The performance improvement of the Sleep-EDF dataset is relatively more challenging than the other two datasets due to a more complex age distribution. Besides, we can observe that the Sleep-EDF dataset has the lowest ACC and K. In addition, the LRP-based heat mapping in Fig.7 can provide visual interpretability to the model prediction. The EEG patterns (i.e., Delta, Theta, Alpha, and Beta waves) of specific sleep stages show high relevance to the correct prediction, which is consistent with the sleep scoring guidelines. Different from the study [45], in which the LRP was also implemented for interpretable sleep staging. We adopt the time-frequency image as the model input rather than the power spectral density, which contains the time-frequency domain information simultaneously. Besides, it is more visually accessible by demonstrating the relevance of different EEG patterns with the heat mapping.

There are also some limitations of our study. First, while our MSSENet could achieve favorable overall performance, it is still challenging to classify stage N1 accurately. The $F1$ of stage N1 in three datasets is less or close to 50%, which is much lower than other stages. Considering Tables III, IV, and V, we can see that stage N1 is easily misclassified as stages REM, N2 and W. As shown in Table II, EEG patterns of stage N1 are similar to stage W (i.e., Theta and Alpha waves). Besides, stage N1 has the same brain activity as stage N2 (Theta wave). Although the single-channel EGG-based method can significantly reduce computational costs, the contribution of other modality signals to the correct identification of stage N1 is disregarded. For instance, EOG signals can be shown to differentiate between stages N1 and REM [46]. Similarly, EMG signals also benefit the accurate recognition between stages N1 and W [47]. The multi-modality scheme with channel selection strategy could be investigated in future

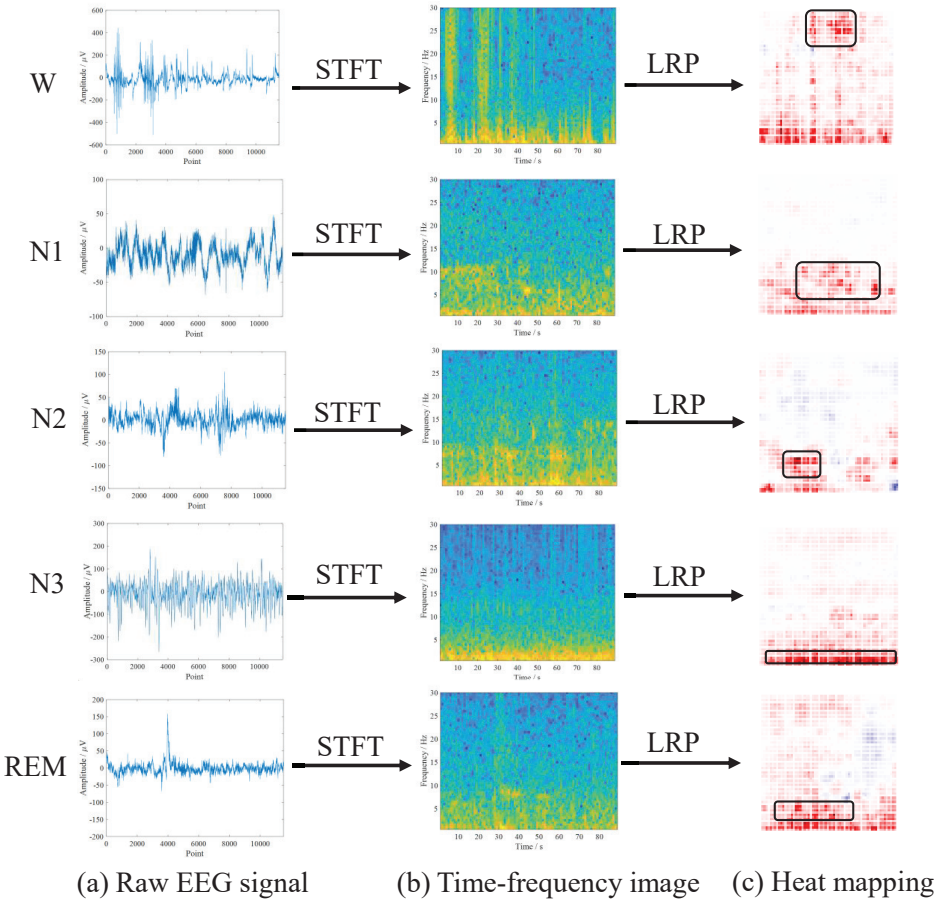


Fig. 7. The explainability results for each sleep stage implementing the LRP method.

work. Secondly, the proposed interpretable scheme could not detect other EEG patterns of particular sleep stages (e.g., K-complexes in stage N2), which also benefits sleep scoring. The study [47] provided insight into the possible detection of other EEG patterns with the raw EEG signals using the LRP method, a more morphological representation of different EEG patterns that mimic the manual visual inspection. In addition, we will explore other modalities to understand the learning course of the deep learning-based method for automatic sleep scoring.

ACKNOWLEDGMENT

This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä for his great help to the authors and Prof. Tapani Ristaniemi has supervised this study very much.

REFERENCES

[1] F. S. Luyster *et al.*, "Sleep: a health imperative," *Sleep*, vol. 35, no. 6, pp. 727–734, 2012.

[2] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng*, vol. 26, no. 2, pp. 324–333, 2017.

[3] A. Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol*, vol. 26, no. 6, p. 644, 1969.

[4] 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: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine," *J. Clin. Sleep Med*, vol. 8, no. 5, pp. 597–619, 2012.

[5] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR)*, pp. 12873–12883, 2021.

[6] H. Ullah *et al.*, "Internal emotion classification using eeg signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.

[7] Q. Xu *et al.*, "Hierarchical spiking-based model for efficient image classification with enhanced feature extraction and encoding," *IEEE Trans. Neural Netw. Learn. Syst*, vol. Early Access, 2022.

[8] T. Zhang *et al.*, "Frequency-aware contrastive learning for neural machine translation," in *Proc. Conf AAAI. Artif. Intell (AAAI)*, vol. 36, pp. 11712–11720, 2022.

[9] S. D. Khan, "Congestion detection in pedestrian crowds using oscillation in motion trajectories," *Eng. Appl. Artif. Intell*, vol. 85, pp. 429–443, 2019.

[10] Y. Ding, L. Hua, and S. Li, "Research on computer vision enhancement in intelligent robot based on machine learning and deep learning," *Neural. Comput. Appl*, vol. 34, no. 4, pp. 2623–2635, 2022.

[11] E. Felemban *et al.*, "Deep trajectory classification model for congestion detection in human crowds," *Comput. Mater. Contin*, vol. 68, no. 1, pp. 705–725, 2021.

[12] S. Chambon *et al.*, "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.

[13] D. Zhou *et al.*, "Lightsleepnet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. Eng. Med. Biol. Soc (EMBC)*, pp. 43–46, IEEE, 2021.

[14] R. Yan *et al.*, "A deep learning model for automatic sleep scoring

using multimodality time series," in *Proc. Eur. Signal Process. Conf (EUSIPCO)*, pp. 1090–1094, IEEE, 2021.

[15] Q. Xu *et al.*, "Convolutional neural network based sleep stage classification with class imbalance," in *Proc. Int. Jt. Conf. Neural Netw, (IJCNN)*, pp. 1–6, IEEE, 2022.

[16] H. Phan *et al.*, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *Proc. Eng. Med. Biol. Soc (EMBC)*, pp. 1452–1455, IEEE, 2018.

[17] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded lstm recurrent neural network for automated sleep stage classification using single-channel eeg signals," *Comput. Biol. Med*, vol. 106, pp. 71–81, 2019.

[18] H. Phan *et al.*, "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.

[19] A. Supratak *et al.*, "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.

[20] Z. Chen *et al.*, "An attention based cnn-lstm approach for sleep-wake detection with heterogeneous sensors," *IEEE J. Biomed. Health Inform*, vol. 25, no. 9, pp. 3270–3277, 2020.

[21] D. Zhou *et al.*, "Alleviating class imbalance problem in automatic sleep stage classification," *IEEE Trans. Instrum. Meas*, vol. 71, pp. 1–12, 2022.

[22] D. Zhou *et al.*, "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *Biomed. Signal Process. Control*, vol. 75, p. 103592, 2022.

[23] D. Jiang, M. Yu, and W. Yuanyuan, "Sleep stage classification using covariance features of multi-channel physiological signals on riemannian manifolds," *Comput. Methods. Programs. Biomed*, vol. 178, pp. 19–30, 2019.

[24] B. Yang *et al.*, "A single-channel eeg based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden markov model," *Biomed. Signal Process. Control*, vol. 68, p. 102581, 2021.

[25] R. Yan *et al.*, "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods*, vol. 348, p. 108971, 2021.

[26] R. Zhao, Y. Xia, and Y. Zhang, "Unsupervised sleep staging system based on domain adaptation," *Biomed. Signal Process. Control*, vol. 69, p. 102937, 2021.

[27] Z. He, L. Du, P. Wang, P. Xia, Z. Liu, Y. Song, X. Chen, and Z. Fang, "Single-channel eeg sleep staging based on data augmentation and cross-subject discrepancy alleviation," *Comput. Biol. Med*, vol. 149, p. 106044, 2022.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *J. Mach. Learn. Res*, vol. 9, no. 11, 2008.

[29] C. A. Ellis *et al.*, "Explainable sleep stage classification with multimodal electrophysiology time-series," in *Proc. Eng. Med. Biol. Soc. (EMBC)*, pp. 2363–2366, IEEE, 2021.

[30] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A novel local explainability approach for spectral insight into raw eeg-based deep learning classifiers," in *Proc. IEEE Int. Conf. Bioinformatics. Biomed. (BIBE)*, pp. 1–6, IEEE, 2021.

[31] G.-Q. hang *et al.*, "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc*, vol. 25, no. 10, pp. 1351–1358, 2018.

[32] C. L. Rosen *et al.*, "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity," *J. Pediatr*, vol. 142, no. 4, pp. 383–389, 2003.

[33] B. Kemp *et al.*, "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.

[34] H. Phan *et al.*, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell*, 2021.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR)*, pp. 7132–7141, 2018.

[36] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas*, vol. 20, no. 1, pp. 37–46, 1960.

[38] T. Nakamura, H. J. Davies, and D. P. Mandic, "Scalable automatic sleep staging in the era of big data," in *Proc. Eng. Med. Biol. Soc (EMBC)*, pp. 2265–2268, IEEE, 2019.

[39] X. Li *et al.*, "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform*, vol. 22, no. 2, pp. 375–385, 2017.

[40] H. Seo *et al.*, "Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg," *Biomed. Signal Process. Control*, vol. 61, p. 102037, 2020.

[41] W. Qu *et al.*, "A residual based attention model for eeg based sleep staging," *IEEE J. Biomed. Health Inform*, vol. 24, no. 10, pp. 2833–2843, 2020.

[42] H. Phan *et al.*, "Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification," in *Proc. Eng. Med. Biol. Soc (EMBC)*, pp. 453–456, IEEE, 2018.

[43] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE*, vol. 14, no. 5, pp. 1–15, 2019.

[44] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. Eng. Med. Biol. Soc (EMBC)*, pp. 641–644, IEEE, 2020.

[45] C. A. Ellis *et al.*, "Hierarchical neural network with layer-wise relevance propagation for interpretable multiclass neural state classification," in *Proc. Int. IEEE/EMBS Conf. Neural Eng (NER)*, pp. 351–354, IEEE, 2021.

[46] M. Ronzhina *et al.*, "Sleep scoring using artificial neural networks," *Sleep Med. Rev*, vol. 16, no. 3, pp. 251–263, 2012.

[47] Banluesombatkul *et al.*, "Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning," *IEEE J. Biomed. Health Inform*, vol. 25, no. 6, pp. 1949–1963, 2020.