Doctoral Dissertation

Representation Learning Schemes and Interpretable Scoring for Sleep Stage

Zheng Chen

Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Shigehiko Kanaya
Computational System Biology Lab. (Division of Information Science)

Submitted on December 01, 2021

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Zheng Chen

Thesis Committee:

Supervisor	Shigehiko Kanaya
	(Professor, Division of Information Science)
	Takayuki Tohge
	(Associate Professor, Division of Biological Science)
	Naoaki Ono
	(Associate Professor, Division of Information Science)
	MD Altaf-Ul-Amin
	(Associate Professor, Division of Information Science)
	Ming Huang
	(Assistant Professor, Division of Information Science)

Representation Learning Schemes and Interpretable Scoring for Sleep Stage*

Zheng Chen

Abstract

Sleep stage scoring/classification is crucial for the assessment of sleep quality and the diagnosis of sleep disorders. Recently with the progress in deep learning, the sleep community has seen successful applications of deep networks on automatic stage classification tasks fueled by large-scale public sleep datasets. Undoubtedly, the performance of a deep learning model is heavily dependent on the data representation on which they are pre-processed to reveal more latent information. Hence, how to represent the stage-dependent characteristics in human brain is crucial for the subsequent classifier. Moreover, manual scoring is a rule-based process. It needs to follow the criteria that were defined by combining the physiological evidence with the consensus of sleep experts. Hence, a reasonable and interpretable framework is precisely what the sleep scoring community requires.

Considering the frequency characteristics of the electroencephalogram (EEG) in sleep nature, this thesis first explores the different time-frequency frameworks for the representation learning of the EEG following the definition of sleep medicine. The results show the stage-dependent characteristics of EEG waveform can be embedded into a set of spectrogram of short EEG segments that is compliant with the scientific understanding of the cortical behavior during sleep. By merging with the parallel computational deep learning structure-Transformer, the pipeline attains the best performance when compared with the other works and reached a new state-of-the-art. Finally, this thesis visualizes the stage scoring

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, December 01, 2021.

process of the model decision with the layer-gradient-wise propagation method, which shows that our work is more sensitive and perceivable in the decision-making process than the existing related works.

Keywords:

Sleep stage classification, EEG, Representation Learning, Deep Learning, Transformer

Acknowledgements

I would like to thank the following people for their wisdom, guidance, and support on my work. Without these people, this thesis would never have been possible.

First and foremost, I would like to express sincere gratitude to my supervisor, Prof. Shigehiko Kanaya for giving me precious opportunities to enjoy the scenery along the way of Ph.D. programs in his laboratory. He also provides me with insightful guidance and encouragement during my student life. Without his support, I would not successfully accomplish the doctoral degree.

I would like to express my gratitude to my thesis committee, Assoc Prof. Takayuki Tohge. He gives me invaluable comments and suggestions in my presentation of colloquium to improve the quality of my research.

I would like to offer my special thanks to Assoc. Prof. Naoaki Ono and Assoc. Prof. MD Altaf-Ul-Amin. Prof. Naoaki Ono provided some insight comments in every lab-seminar and helps me to solve the issue in computational resource. Assoc. Prof. MD Altaf-Ul-Amin also took a serious attention to my research and my work-life balance, provided constructive feedback, and brought my work to a higher level. Here, I would like to show my special thanks to Prof. Amin's vegetables, especially, the white gourd, that is my favorite vegetable in winter. Once again, thank you for participating in this committee!

I am deeply grateful to Assist. Prof. Ming Huang for his assistance at every stage of the research project since the first year of my doctoral program. His wisdom helps me to resolve lots of research difficulties, improve my writing skills, and shape the critical thinking. Without his support and guidance, I would not have this beautiful journey of Ph.D.

I would like to extend sincere thanks to my lab-mates (Koki Odani and Ziwei Yang) and close friend (Dong Wang). I would not have been able to have great research experiences without their support.

In addition, I would like to thank my beloved parents for their encouragement all the time. Without them, I would not have had a chance to pursue my dream in Japan. Finally, I would like to extend my sincere thanks to Jiaxuan Zhang for her wise counsel and sympathetic ear. She is the love of my life. She is always there for me and goes through this wonderful journey together.

List of Publications

Journal paper

- **Zheng Chen**, N. Ono, W. Chen, T. Tamura, M. Altaf-Ul-Amin, S. Kanaya, M. Huang. "The Feasibility of Predicting Impending Malignant Ventricular Arrhythmias by Using Nonlinear Features of Short Heartbeat Intervals". Computer Methods and Programs in Biomedicine (CMPB), vol. 205, p.106102, 2021.

- **Zheng Chen**, Ziwei Y, M. Huang, W. Chen, T. Tamura, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, Enhancement on Model Interpretability and Sleep Scoring Performance with A Novel Pipeline Based on Deep Neural Network, IEEE Journal of Biomedical and Health Informatics. 2021. (*Under review*)

- Yongxin Zhang, **Zheng Chen**, HY. Tian, M. Huang, W. Chen, T. Tamura, N. Ono, M. Altaf-Ul-Amin, S. Kanaya. A Real-Time Portable IoT System for Telework Tracking. Frontiers in Digital Health, vol.205, p.106102, 2021

Conference paper

- **Zheng Chen**, N. Ono, M. Huang, M.D Amin, S. Kanaya, "iVAE: an Improved Deep Learning Structure for EEG Signal Characterization and Reconstruction," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'20), 12 2020, pp. 1909–1913.

- **Zheng Chen**, K. Oda, P. Gao, M. Huang, N. Ono, M.D Amin, S. Kanaya "Feasibility Analysis of Transformer Model for EEG-based Sleep Scoring," IEEE International Conference on Biomedical and Health Informatics (BHI'21), July, 2021.

- **Zheng Chen**, P. Gao, M. Huang, N. Ono, M.D Amin, S. Kanaya, "Feasibility Analysis of Symbolic Representation for Single-Channel EEG-Based Sleep Stages," The 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'43), Nov, 2021.

- **Zheng Chen** et al., "An End-to-End Sleep Staging Simulator Based on Mixed Deep Neural Networks," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 848-853.

- Pei Gao, **Zheng Chen**, M Huang, N Ono, M.D Amin, S Kanaya. Prediction of TCM Effective against Bacterial Pneumonia and Identification of Antibacterial Natural Products. The 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'43), 1-page abstract, 2021.

- Pei Gao, **Zheng Chen***, Dong Wang, Ming Huang, Naoaki Ono, Altaf Amin, Shigehiko Kanaya. Exploring Feasibility of Truth-Involved AutomaticSleep Staging Combined with Transformer. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'21), 2021pp. 2211-2216.

- Ziwe Yang, Dong Wang, **Zheng Chen***, Ming Huang*, Naoaki Ono, Altaf Amin, Shigehiko Kanaya. Exploring Feasibility of Truth-Involved AutomaticSleep Staging Combined with Transformer. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'21), 2021, pp. 2920-2923.

Contents

List of Figures

List of Tables

1 Introduction

Sleep is a human function, the characteristics of which are manifested by a sequence of physiological alterations, including neural spiking, cardiorespiratory, blood oxygen saturation, and eye activity. During sleep, the consciousness will be altered, the sensory activity is relatively inhibited while its reduced muscle activity and inhibition of nearly all voluntary muscles during rapid eye movement (REM) sleep [30]. It is distinguished from wakefulness by a decreased ability to react to stimuli, but more reactive than a coma or disorders of consciousness, with sleep displaying different, active brain patterns.

Sleep occurs in repeating periods, in which the body alternates between two distinct modes: REM sleep and non-REM sleep [65]. Although REM stands for "rapid eye movement", this mode of sleep has many other aspects, including virtual paralysis of the body. A well-known feature of sleep is the dream, an experience typically recounted in narrative form, which resembles waking life while in progress, but which usually can later be distinguished as fantasy. During sleep, most of the body's systems are in an anabolic state, helping to restore the immune, nervous, skeletal, and muscular systems; these are vital processes that maintain mood, memory, and cognitive function, and play a large role in the function of the endocrine and immune systems[50]. The internal circadian clock promotes sleep daily at night. The diverse purposes and mechanisms of sleep are the subject of substantial ongoing research. Sleep is a highly conserved behavior across animal evolution [44].

Screening sleep is not only a tool in the assessment of pathophysiology [25, 51], but an essential ingredient in the exploration of neuroscience [34, 68, 88]. Humans may suffer from various sleep disorders, including dyssomnias such as in-

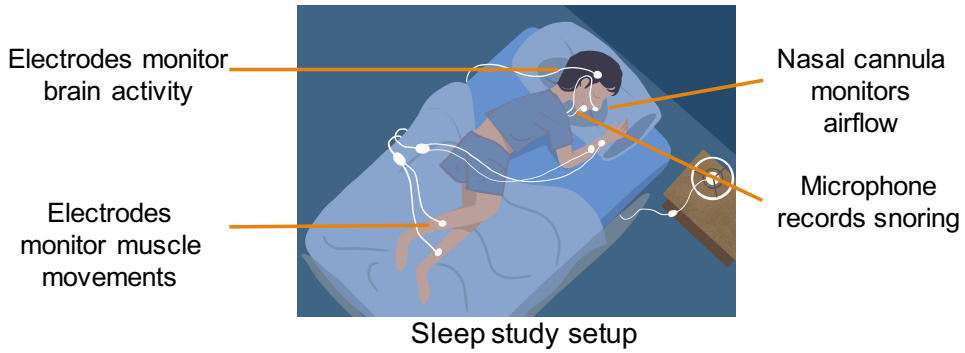


Figure 1.1: Human sleep during PSG recording

somnia, hypersomnia, narcolepsy, and sleep apnea; parasomnias such as sleep-walking and rapid eye movement sleep behavior disorder; bruxism; and circadian rhythm sleep disorders [8]. Moreover, experimental validations prove that some electroencephalogram (EEG) features shown in different sleep stages have extraordinary physiological significance [3]. For instance, slow waves contribute to memory consolidation [64], and neurophysiologists proved sleep spindle is highly correlated with tests of intellectual ability [31]. Hence, determining sleep stages and consequent macrostructures is indispensable in sleep care and sleep science.

Standardized rules for sleep stage scoring using polysomnography (PSG) (seen in Figure 1.1) were first laid out by Rechtschaffen and Kales (R&K rule) in 1968 [1]. There are seven stages: wake, stage S1, stage S2, stage S3, stage S4, REM and movement. Usually, stage S1 and S2 can be viewed as light sleep that implies human enters to beginning phase of sleep. This light sleep often contains few sleep epoch and transfers the sleep condition to deep very fast. Stage S3 and S4 are the deep sleep that human rehabilitates the physical fatigue during this phase. The American Academy of Sleep Medicine (AASM) re-defines overnight sleep into five different stages in 2007, where the stage S4 and movement is no longer existed. There is a constant cyclic pattern of sleep stages from wake to non-rapid eye movement (NREM) to REM that repeats several times one night, where NREM consists of three stages, i.e., N1, N2, and N3 [7]. Sleep staging clinically relies on overnight electrophysiological recordings using the PSG [71]. Clinically, experts are required to inspect the stage-dependent characteristics of various physiological indexes in PSG recording and then assign a sleep stage

to each 30-second epoch on the basis of distinct features. This laborious handling strongly relies on prior knowledge and inevitably limits the batch scoring of stages. Moreover, given that the recent advances in portable monitoring with fewer sensors could provide technical support for daily sleep screening [70, 84]. Therefore, a reliable automatic sleep stage scoring alternative is crucial for the sleep community.

1 Problem Statement

At the level of human electrophysiology, quantitative analysis of EEG is the priority to illuminate the structure of sleep and its control mechanisms [34]. Automatic sleep scoring stimulated by the field of deep learning and recently shows the promises [67, 76, 102]. With benefiting from the increasingly available sleep databases, numerous advanced frameworks with large-scale parameters proposed to explore the feasibility in several aspects [4, 9, 72, 75, 80, 92]. However, ***the current research findings do not break a blockade by the manual process yet.***

The performance of the current deep learning models is overly dependent on data and its representation [5]. Back to the sleep nature, the functional cooperative interaction of brain dynamics always has heterogeneous characteristics of inter-subject, even the same subject records at the different physical or emotional conditions [3]. Without considering and introducing this nature, the methodology of sleep stage scoring alternative is limited for revealing the inter-subject variability inherent in real clinical settings. Therefore, the representation learning for the stage-dependent characteristics in sleep nature is a corn stone but essential study for sleep staging alternative.

Some studies pay attention to finding a entire alternative. The network architectures are designed on top of reckless pursuing the scoring performance, with overly relying on the ability of deep learning. With the architectures getting more complex, the derived black-box issues lead the decision-making of the networks to lack the explanations. A reasonable deep learning framework is precisely what prerequisites the automatic sleep scoring community requires.

More recently, some studies aim to construct the end-to-end framework from

mining the latent features in data to stage scoring. However, manual scoring is a rule-based process. It needs to follow the criteria that defined by combining the physiological evidence with the consensus of sleep experts [72]. Without considering constraints of clinical sleep truth, it is hard to evaluate whether the results can assist the sleep scoring. Therefore, the effectiveness of the proposed frameworks is somewhat unconvincing. In conclusion, an interpretable sleep scoring needs to involve an adaptive framework designed from the perspective of data attributes, while explaining the decision-making of the networks. However, manual scoring is a rule-based process that has to follow the criteria which defined by combining the physiological evidence with the consensus of sleep experts [72]. Building the framework cannot completely escape the constraints of clinical sleep truth to ensure effectiveness. In another word, interpretability sleep scoring needs to involve an adaptive framework designed from the perspective of data attributes.

Existing works have proved that combined with the feature mapping layers of deep learning, the performance has a significant improvement [4, 80, 92]. Meanwhile, the recent advances in portable monitoring with fewer sensors provide technical support for daily sleep screening [70, 84]. Despite the above attempts, it still remains unclear about (i) the representation learning method for sleep stage-dependent characteristics, (ii) how to construct a context-sensitive flexible pipeline that automatically adapts attributes of sleep data itself, (iii) the Interpretability of the model decision-making. Therefore, I state this thesis as follows:

Thesis Statement: Considering the time-consuming issue in sleep stage scoring, a reasonable and accurate automatic sleep stage framework is required. Presuming the natural frequency characteristics in sleep medicine, this thesis aims to proposes a time-frequency framework for the representation learning of the electroencephalogram following the definition of sleep. To meet the temporal-random and transient nature of the defining characteristics of sleep stages, this thesis presents a context-sensitive flexible pipeline that automatically adapts to the attributes of data itself. Through the thinking and mimicking the sleep nature and scoring process, the proposed framework provides the potential values and feasibility for future sleep health-care.

2 Contributions

The main contributions of this thesis can be classified into two categories: empirical observations and future research directions.

Empirical Observations

1. The multi-resolution latent features of EEG have a more accurate reconstruction and it is benefit for downstream sleep stage classification. (Chapter 3)

2. Latent Dirichlet Allocation topic model proved the feasibility of the multi-topics representation for sleep signals. (Chapter 3)

3. Sleep scoring framework requires to extract the relevant features of sleep stage that considers the intrinsic characteristics of stage defining characteristics expressed in EEG signal, e.g., Delta (δ), Theta (θ), Alpha (α), Sigma (σ), Beta (β). (Chapter 3)

4. The transient characteristics of the frequency domain can be parallel captured from the different 1-second time duration. (Chapter 3)

5. By constructing a delicate network structure, this paper reaches state-of-the-art performance in scoring accuracy of intra-epoch classification under the EEG-signal-only restriction. (Chapter 4)

6. By using the layer-wise relevance propagation method in the proposed model, the resultant attention-derived matrix can be explained clinically and therefore provides interpretability for the staging decision. (Chapter 4)

7. The sleep-mechanism based framework is able to elevate the scoring accuracy on one hand, and provides better interpretability on the other hand; (Chapter 5)

Future research directions

1. Sleep stage scoring alternative requires more available dataset to tackle the variability of inter-subject. (Chapter 3)

2. However, the training issue, especially, sequence-level training needs to adopt quantified strategy. (Chapter 4)

3. A further redundant information refinement and reduction will speed up the real use in automatic sleep stage scoring. (Chapter 4)

3 Thesis Outline

In this section, I provide structure of the thesis and the potential research outcomes. In the remainder of the thesis, first of all I introduce the studies that are related to this thesis.

- **Chapter 2**— Three main related topics are discussed as follows: (i) representation leaning method, (ii) automatic sleep stage scoring model, and (iii) model explanation.

I then present the main studies that are split into three chapters. These chapters are organized into **Part I Representation Learning of Sleep Nature** and **Part II Accurate Staging Framework and Decision Interpretability**. Each chapter has corresponding potential research outcomes.

- **Chapter 3**— This chapter introduces three case study to investigate the representation methods of sleep stage-dependent characteristics in EEG. The results show that extracting sleep frequency characteristics requires to meet the criterion of sleep medicine and to consider feature representations in different frequency bands. Moreover, a time-domain refinement can further improve the sleep stage scoring performance.

- **Chapter 4**— This chapter proposes a novel local-context-sensitive pipeline to extract the relevant features of sleep stage that considers the intrinsic characteristics of stage defining characteristics expressed in EEG signal. Also, the proposed framework can capture the key stage-depdent features in parallel and it is able to elevate the scoring accuracy on one hand.

- **Chapter 5**— This chapter investigates the feasibility of parameter-tracking (model weight and output layers) visualization. By using the layer-wise

relevance propagation method in the proposed framework, the resultant attention-derived matrix can be explained clinically and therefore provides interpretability for the staging decision.

Finally, **Chapter 6** provides the conclusion that includes the main research results and contributions of this thesis. I also provide the potential opportunities for future sleep stage scoring alternative research.

2 | Related Studies

Representation leaning method - Current deep learning methods strongly rely on the data and its representation. During sleep, the noise involved issue and the entangled multiply frequency waves bring the burden to the discriminative model. A reliable representation of the signal is crucial for the sleep stage classification task.

To capture/represent these sleep-specific features, early works have used hand-craft feature design methods based on prior knowledge [46, 74, 97]. Based on these clinical features, effort has been put into the searching for more salient features in sleep scoring. The proposed features include spectrogram [9, 67, 72, 75, 76, 102], power spectral density [40, 89, 97], entropy [46, 89], wavelet transformation [39, 59, 74]. Considering the interpretability of the feature representation, some studies introduced word distribution [49], graph representation [22, 41], statistics [22, 55, 89] to feature extraction phase. These studies transfer the signal dimensions to a explainable dimension and then fed the transformations into the classifiers. However, the reconstructed feature space sometimes cannot retrieval the key features back from the signal sequence. Therefore, empirical mode decomposition [42, 104], wavelet[43], feature mapping neural network [4, 12, 80, 92] are widely used in sleep stage classification study in recent. However, extracting hand-crafted features is time-consuming and requires heavy prior knowledge. Practitioners often select features that are expected to work well with the model from a statistical learning perspective, and seldom consider the feature extraction problem from a neuroscientific viewpoint that explicitly considers the dynamic nature of human brain.

Existing works have proved that combined with the feature mapping layers of

deep learning, the performance has a significant improvement [53, 86, 92, 102]. Even, some researchers try to construct a fully staging model from the feature extraction to end classification decision.

Due to the convolutional neural networks (CNNs) becoming the defacto standard for visual classification tasks, CNNs have been used in feature extraction from the frequency or time-frequency domain [4, 80, 92]. By interleaving a collection of multi-size convolutional filters with non-linear activation functions and downsampling operators, proposed scoring frameworks expect to capture sophisticated local features and attain an inductive representation of the EEG epoch [35]. To named a few, Tsinalis *et al.* [97] employed a set of CNNs to learn task-specific filters for sleep classification. Enrique *et al.* [28] and Arnaud *et al.* [86] explored the extended CNNs-based architectures to capture more meaningful frequency features within the deeper latent space.

Since CNNs does not consider the local features against the global context, it will results in a local inductive bias [98]. Consequently, it may draw more attention from the classifier onto a pattern of the neighboring area while distracting the classifier from really important but transient information and its relevance. Meanwhile, feature extraction in each EEG epoch is handled by the sharing weights of filters. This mechanism expects key rhythms to occur in a relatively inherent position. Therefore, there is a mismatching between the translation-invariant constraint of CNNs and the temporally random and transient nature of the relevant defining characteristics.

Automatic sleep stage scoring model - The sleep community has witnessed successes in developing automatic staging systems combined with deep learning methods. Recurrent neural networks (RNNs), such as long short-term memory (LSTM) and gated recurrent neural networks (GRUs), have been established as the state-of-the-art (SOTA) in automatic sleep scoring [75, 80, 85, 89]. Different from CNNs, the RNNs-based models pay attention to the information of the global context. By generating a set of hidden states, the decision-making considers the influences of a sequence of previous time steps or future steps (by the Bidirectional LSTM) [52].

Since RNNs-based models allow the sequential modeling of dependencies and transfer the temporal influence into the data, some work utilized RNNs models

into the construction of the staging system. Dong *et al.* [92] proposed a mixed neural network that combines the RNNs with the LSTM. The 2-layer RNNs were used to extract hierarchical handcraft features, while the LSTM was applied to classify the sequence data. Supratak *et al.* [92] constructed a fully deep learning model by combining the bidirectional-LSTM with the feature map CNNs. In this work, CNNs were directly applied to the raw EEG signal to output the feature vector which is further used for staging in Bi-LSTM. More recently, Xu *et al.* [102] transformed the EEG+EOG signals to the multi-channel spectrogram and then fed it into the LSTM model. They aim to find the time-frequency characteristics from the second-to-second time duration and claim the performance has an competitive overall accuracy. Considering the parallel computation, recent works attempt to introduce the attention mechanism to the sequential model. In the CNNs-LSTM model proposed by Chen *et al.* [15], an attention network was merged to adjust the significance of two groups of features. Sun *et al.* [89] proposed a multi-flow RNN which is utilized to learn temporal information by fusing the hand-crafted features and network-trained features.

Although the periodic frequency waves, for example, *theta* oscillations meet the assumption of a recurrent-based model, such transiently burst rhythms, such as non-periodic sharp-wave ripples and spindle activities are unpredictable [2, 10]. Here, the previous time duration of the EEG segment cannot provide the indicator for key rhythms. Further, the inherently sequential nature (time-invariant) of the recurrent-based model precludes the possibility of parallelization within feature capture [99]. This constraint but is a non-trivial consideration for screening the occurrence of transient sleep rhythms.

Model Explanation - Sleep experts point to the scepticism of deep learning models being a black box, which is a common criticism when it comes to the application of artificial intelligence in healthcare and medicine [78]. Model explanation, especially, to explicitly interpret the model decision-making process is crucial for a machine scoring system to work alongside practitioners in an interactive and collaborative manner. Some previous studies pay attention to providing the interpretability for their designed scoring framework. Phan *et al.* visualizes the learned weights from a sequence of the sleep epochs by using spectrogram of short-time Fourier transformation [75]. Sun *et al.* visualizes the different forms

of signal characteristics from the network trained features. Meanwhile, this work validates the resulted features with the sleep experts annotation. They found the high similarities in the morphology between network trained features and the sleep events defined by AASM [89]. However, the aforementioned methods still focus on providing the interpretability of result, but the decision-making itself. More recently, Phan *et al.* [78] adopts an attention-based model as the sleep stage classifier. The epoch-level attention scores in this work are used as a heat map applied to the EEG signal input to highlight the features the model attends to. Also, the attention scores at the sequence level is interpreted as the influence of different neighboring epochs to the recognition of a target epoch in an input sequence. It provides a certain quantilization analysis for model-based scoring decision.

Part I

Representation Learning of Sleep Nature

3 | Representing Stage-Dependent Characteristics

1 Introduction

The sleep stage annotation is crucial to sleep disorder/insomnia diagnosis, and the medical evidence is based majorly on the analysis of the EEG. Quantitative electrophysiology, which allows for the sophisticated processing of EEG signals, has revealed how widespread human neuronal systems generate the characteristic electrical rhythms of different sleep stages [34].

The EEG-based technical specifications of AASM define the staging rule with a 0.5 to 30-35 Hz range in frequency oscillation [7]. By inferring stage-dependent occurrences (or increases) in EEG features, the clinical setting assigns a stage to each 30-second epoch.

Specifically, the transition from quiet wakefulness to eyes closed in humans is characterized by the onset of 8-12 Hz posterior dominant rhythm (alpha rhythm), while scoring epochs as wakefulness when more than 50% of the epoch contains alpha rhythm [2]. Beta waves (16-32 Hz) are the states associated with wakefulness. The slow oscillations provide the nested characteristic waveforms (spindles, K-complexes, theta and delta waves) for NREM sleep. Theta waves (4-8 Hz) that consist of the low-amplitude and mixed frequency activity are the criteria for stage N1. When one or more K-complex and sleep splines under the low beta waves (12-16 Hz) appear, the corresponding epoch will be scored as stage N2. Stage N3 is characterized by large-amplitude, slow oscillations (<1 Hz) and delta waves (0-4 Hz) [2]. Although the significant characteristic of REM is the move-

Table 3.1: EEG frequency definitions of AASM

Rhythm	Frequency band (Hz)	Target stages
Delta (δ)	1-4	N3
Theta (θ)	4-8	N1, R
Alpha (α)	8-12	N1, W
Sigma (β_1)	12-16	N2
Beta (β_2)	16-32	W, R

ment of the eye and muscle, the saw-tooth waves are alternative scoring criteria. The summary of stage-dependent frequency waves has shown in TABLE 3.1.

Moreover, these frequency rhythms are reliant upon different modulations from distinctive subcortical regions of the human brain. For example, the high energy distribution of the occipital region commonly reflects activation of the alpha rhythm [7]. However, the distributed networks of brain structures make the information sharing from different activated regions and even synchronized activation [26]. This mechanism results in a large number of stage-irrelevant disturbances (or noise) in EEG records. Sometimes, the continuum of frequency waves across stages even dominates the current stage [2]. As a consequence, the stage-dependent characteristics are not always unequivocal among the same staged EEG epochs.

However, some physiological nature of sleep still hinders the mapping from the inherent feature space to inductive subspace. Reflecting the mechanism of burst firing at the cell level to the frequency rhythm in EEG, the spontaneous rhythms of the met criteria of staging sometimes are transient events [18]. For instance, the spindling activity is even burst within a duration of 0.5–1.5 second [2]. The duration of vertex sharp waves which represent stage N1 is less than 0.5 seconds.

Meanwhile, the transient rhythms exhibit dynamic or temporal-random characteristics while it bursts with other concomitant frequency waves. Although this truth leads to the stage-dependent waves cannot dominate one epoch, the intense burst firing is important in stage scoring processes. Sometimes, an assignment of the stage relies on the key waves which solely appear at one time

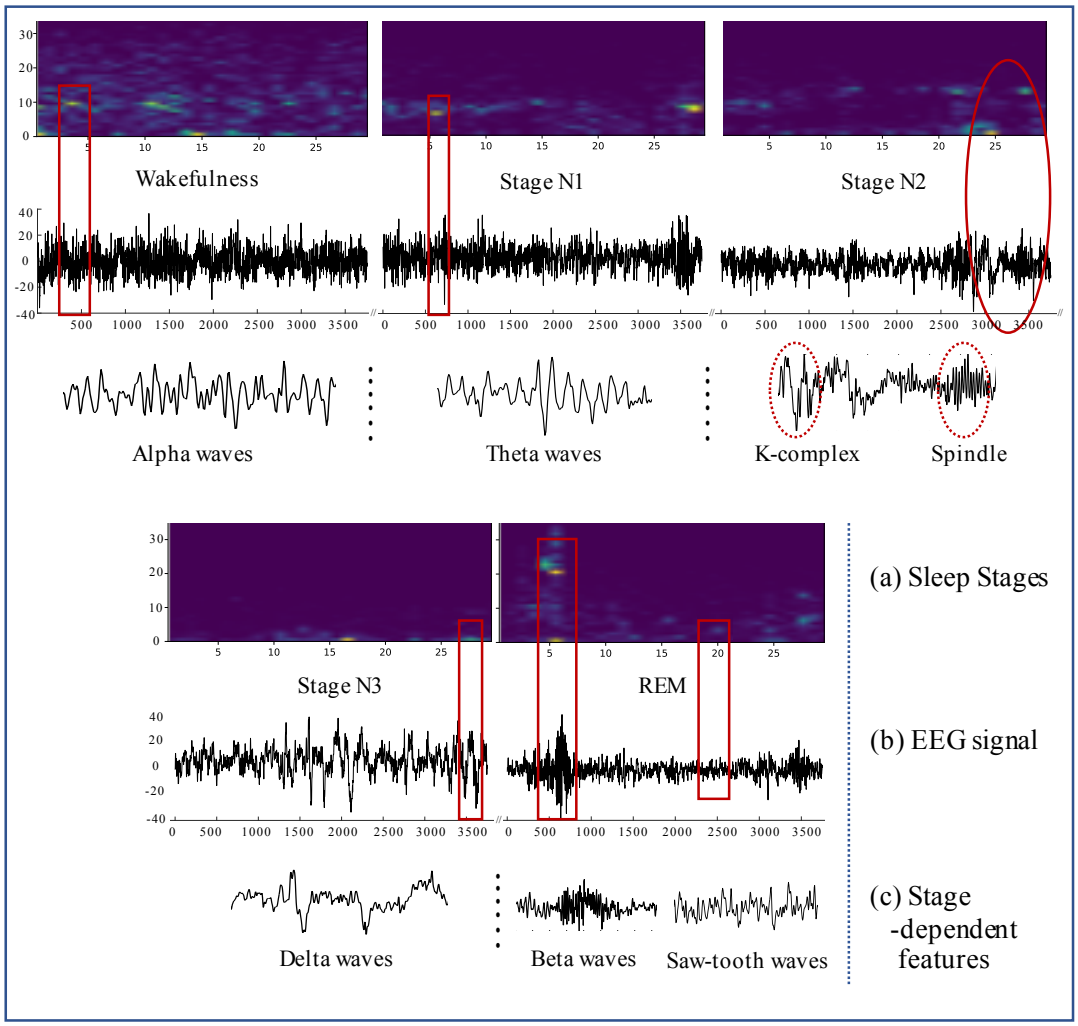


Figure 3.1: Stage-dependent features in sleep medicine: (a) exhibits a sample spectrogram for its sleep stage, while (b) is the corresponding EEG epoch signal. The stage-dependent feature following the TABLE I has highlighted in red. (c) shows the morphological details for different sleep rhythms.

(e.g., K-complex for stage N2 [7]). In addition, the issue of stage transition remains. That is, multi-segment of an epoch often meets the criteria for different stages. The temporal nature inevitably limits to capture the stage-dependent characteristic from the time-invariant space. Importantly, how to represent the key transient non-stationary rhythms, e.g., stage-dependent feature is a runoff effort for the staging system.

The end-to-end pipeline implemented by deep learning models in recent years has made feature generation more flexible and problem-specific [75, 80]. The subject-wise generalization gap, especially in models with complicated representation learning, can be problematic [81]. Therefore, where to draw a line between predefined and adjustable representation generation should be discussed.

We can think of predefined-feature-only input and raw EEG input to a learning model as two extremes and assume that a better solution lies between them. To finding a property representation method that is to server for downstream staging/classification deep learning framework, this work aims to explore a feature reasonable feature method in both learning method and hand-crafted method. There are three case studies (CR) of representation learning are as below:

- **CR1: A deep generative model: variational-autoencoder-based representation**
- **CR2: A multifold symbolic representation with Latent Dirichlet Allocation topic model**
- **CR3: A context-sensitive flexible pipeline based on time-frequency representation**

The key results of each CR are as follow:

For CR1, we constructed a shallow iVAE (Inception-based variational autoencoder) model, which will capture the multi-scale features of the spectrogram of EEG by replacing the main structure in encoder and decoder with the inception-like structure. By comparing with the vanilla VAE and convolutional autoencoder (CAE), a more accurate reconstruction and a better classification using the latent features of the iVAE can be confirmed.

For CR2, we investigated feasibility of the EEG-based symbolic representation for sleep stages. By combining the Latent Dirichlet Allocation topic model

and comparing with different feature extraction methods, the work proved the feasibility of multi-topics representation for sleep stages and physiological signals.

For CR3, the proposed pipeline is validated against a large database, i.e., the Sleep Heart Health Study (SHHS), and the results demonstrate that the competitive performance for the wake, N2, and N3 stages outperforms the state-of-art works.

1.1 Chapter Organization

The remainder of this chapter is organized as follows: Section 2 introduces the details of *CR1*. Section 3 describes the symbolic representation of EEG (*CR2*). Section 4 introduces the details of *CR3*. Finally, I conclude this chapter and discusses the implications from the findings in Section 5.

2 CR1: iVAE

2.1 Variational Autoencoder

By borrowing the data-driven approaches based on deep learning, researchers have started to use the generative model in data augmentation [48], add latent features extraction [79]. However, one thing we need to be very careful is that many physiological signals have crucial local features, for example, the existence of P-wave in electrocardiogram (ECG), and it puts a very strong restriction on applying the unsupervised methods in biomedical signals.

Recent studies emphasize the importance of representation learning rather than feature engineering [6]. Hence, some researches turn their attention to automatically extract the informative features by using deep-learning-based methods, especially in, DGMs. DGMs is a unsupervised representation learning technology, which can work on feature representation learning of images and reconstruct the feature space into a lower latent [36]. Variational autoencoder (VAE) is one of the type of unsupervised DGM which became widely utilized in analysis of pathological images or biomedical problems.

VAE is now widely used in computer vision, natural language processing and communication, etc [47]. By bringing in the assumption that the latent features

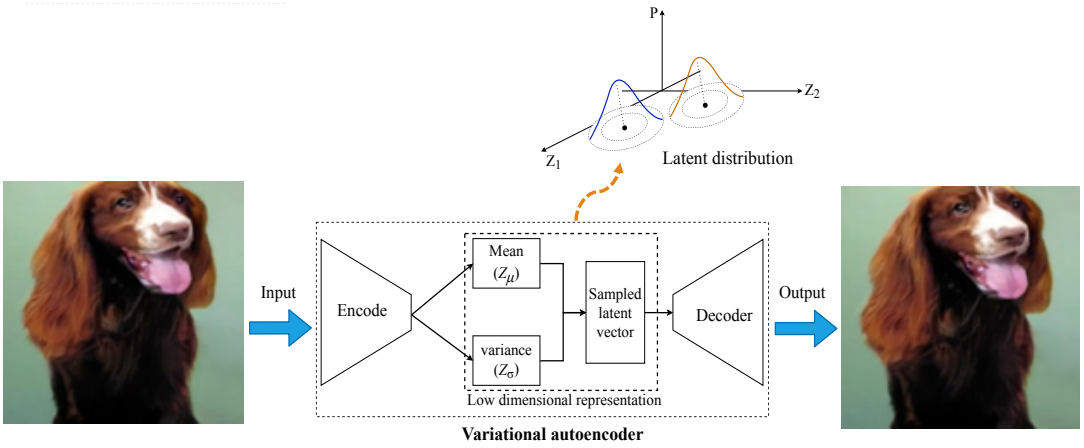


Figure 3.2: VAE mechanism

are of a simple prior distribution and are independent, VAE can capture the complicated distribution of the latent features, which may be further used in downstream pipeline e.g., classification, of the data explicitly. On the other hand, the disadvantages of the VAE are obvious, one of which is the roughness of the rebuilt output. The other one is the prior distribution may not always go well with the real data. The mechanism of VAE can be seen in Figure 3.2.

2.2 Motivating Example

Numerous researchers have tried to use the VAE in biomedical signals. Yildirim *et al.* have used the convolutional autoencoder (CAE) to compress the normal ECG signal [103], which has shown the potential of the VAE in learning the local and general features. Targeting at the practical application, Dai *et al.* have used the CAE model in motor imagery classification with EEG signal and have attained the highest performance in experimental data [19]. Li *et al.* used the RNN-VAE in multi-channel EEG emotion recognition and have shown a better performance over the other recurrent deep learning models [54]. Phan *et al.* used the convolutional neural network in EEG, EMG, and EOG signal to classify the sleep stages into W, N1–N3, and REM [76]. Li *et al.* used the predefined features of EEG extracted from sleep stages and stage transition periods to classify the sleep stages. Combined with assembly random forest classifier, they got satisfactory

accuracy in open datasets [62].

Even though the number of open biomedical databases is increasing in recent years and especially sleep-related datasets have spurted. The gap between trained classification models based on open datasets and their uses in local system remains. One of the propeller in the sleep-staging filed has been the design and validation of new features, which were taken in by the classifier naturally, to manifest the inherent characteristics of sleep stages [38], [21]. Zhang and Wu had tried to choose the predefined features with the unsupervised k-means method and combined them with the complexed-valued CNN to classify the sleep stages [74], which was a trial in searching for new features or features space for sleep stages.

2.3 Dataset and Preprocessing

This case study used two public databases which are available online on the Physionet website: St. Vincent's University Hospital/ University College Dublin Sleep Apnea Database (UCDDB) [1] and MIT-BIH Polysomnographic Database (SLPDB) [2] in this study [32].

The UCDDB contains 25 full overnight PSG recordings by utilizing the Jaeger-Toennies system (sex: 21 males and 4 females; age: 50 $\pm$ 10, range 26-28 years; AHI: 24.1 $\pm$ 20.3, range 1.7-90.9) [37] . Two EEG channels were recorded in this database (C3-A2 and C4-A1). Meanwhile, sleep onset time and sleep stages were scored by an experienced sleep technologist according to standard R&K guidelines. Considering the capacity of two databases, only the signal of C4-A1 channel (128 Hz sample rate) is used in this work.

The SLPDB includes 18 recordings of multiple physiological signals during sleep from 16 subjects. Each record contains the EEG signal of one of the three channels (C4-A1, O2-A1, and C3-O1) with a 250 Hz sampling rate. In considering the activation difference from different brain parts under the same stage, we finally chose five records with EEG signal of C4-A1 channel.

In pre-processing phase, five records in SLPDB are firstly selected and resampled to 128 Hz same as UCDDB. Since sleep EEG recordings have been

[1] https://physionet.org/content/ucddb/1.0.0/

[2] https://physionet.org/content/slpdb/1.0.0/

typically contaminated by various types of artifacts, a 20th order Butterworth bandpass filter with cutoff frequencies between 0.5 and 30 Hz is implemented into two datasets. These cutoff frequencies for bandpass filtering were selected because the brain activities have significant information in 0.5 to 30-35 Hz range during sleep [69]. After the preprocessing of filtering, each record is labeled by its corresponding sleep stage (Wake, REM, Lightsleep, and Deepsleep). Noteworthily, we merge s1 and s2 stages into the lightsleep, thus, s3 and s4 stages are regarded as deepsleep class.

The spectrograms (size: 129×29) are generated for each EEG epoch to transform the signal into the log-power spectra via a short-time Fourier transform (STFT) with a 256 points Hamming window, 50 % overlap, and 256 sampling points Fast Fourier Transform (FFT). Whereafter, min-max normalization is implemented for each spectrogram across different subjects to take the inter-subject variation into the model.

2.4 Network Architecture

Preprocessed spetrogram is input to the proposed VAE-based model (iVAE) to generate a latent representation. VAE is an unsupervised generative model for estimating distributions in lower dimensional latent feature spaces (z) to represent the data itself. The random variable z is learned from the given input X. The posterior $p_\theta(z|X)$ parameterized by weights θ is modeled by an encoder network. Since $p_\theta(z|X)$ is an intractable posterior distribution, VAE framework proposes a substitutive distribution $q_\varphi(z|X)$ to approximate $p_\theta(z|X)$ by variational inference, while $q_\varphi(z|X)$ is usually assumed to be a Gaussian distribution $N(0, 1)$. Meanwhile, a distribution $p_\theta(X|z)$ is constructed by the input data and then evaluate the latent feature spaces z and to represent the distribution of X itself via a decoder network.

$$p_\theta(X) = p_\theta(X|z)p(z) \tag{3.1}$$

The loss function of VAE is minimized following the evidence lower bound, is given by:

$$\mathcal{L}(X; \theta, \varphi) = E_{q_\varphi(z|X)}[\log p_\theta(X|z)] - KL(q_\varphi(z|X)||p(z)) \tag{3.2}$$

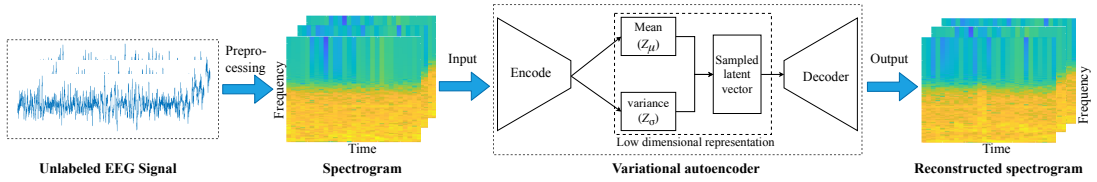
where the KL-Divergence between approximate distribution and the true posterior can be interpreted as the loss of the inference network (encoder), and the generative network (decoder) need to minimize the reconstruction error.

Inception module: The inception-like multi-scale module is used in the encoder following the initial SequenceNet and this module is mirrored in the decoder. The Inception architecture has been proposed in [93] [95], and the key concept of this architecture is based upon finding the optimal local sparse structure in a convolutional vision network that can be approximated and applied [57]. The proposed Inception-like module in our work has shown in Figure 3.3 (b).

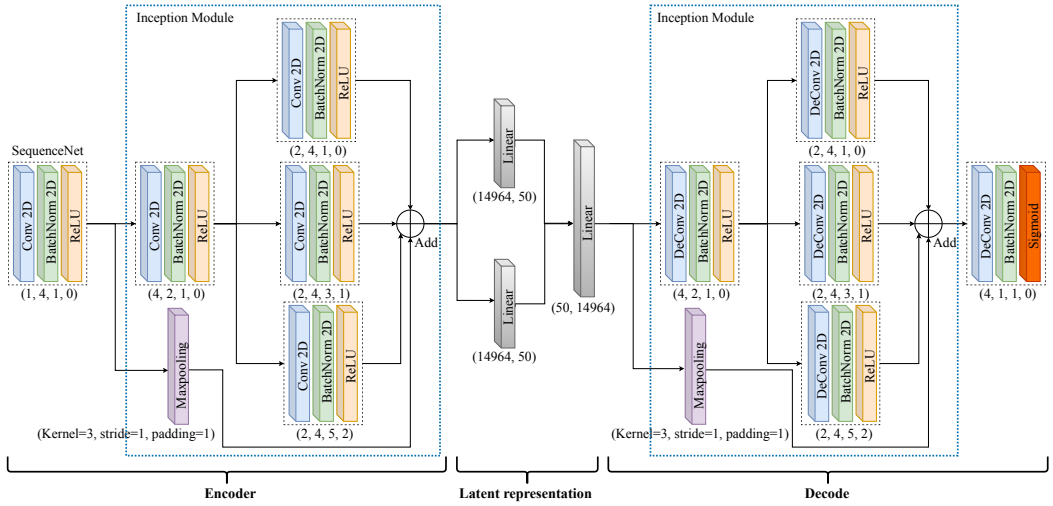
A SequenceNet is first applied to input data, which consists of a 2-D CNN (kernel size = 1) layer, a 2-D batch normalization (BatchNorm) manipulation and rectified linear (ReLU) activation function. The Inception module in this work is modified, which includes multiple SequenceNets of 1×1 CNN, 3×3 CNN, 5×5 CNN and 3×3 Maxpooling layer. The first layer of Inception consists of a Maxpooling layer and a SequenceNet (CNN kernel size = 1), whose aim is to map and store the feature of input.

The output of the first SequenceNet will be separately passed onto three parallel SequenceNets with different size of CNNs to extract the feature of different scales. Here, unlike the conventional Inception module that concatenates the output from previous layers, we replace the concatenation function to the summation of the outputs to decrease the computational complexity.

Two linear layers are applied to Inception output and they are interpreted as a vector of means and a vector of variances of the latent features following normal distribution. In the decoder, Then a new latent vector is sampled from its distribution learned by the encoder and subsequently be mapped and reshaped back to the size of the original spectrogram and used as the input to the decoder. With the decoder structure, it is expected to be reconstructed similarly to the original spectrogram. Moreover, the decoder has the same Inception architecture of encoder which is reconstructing data by using deconvolution layer. The activation function in the last SequenceNet uses the sigmoid function to output the reconstructed spectrogram.



(a)



(b)

Figure 3.3: Framework of proposed iVAE model: (a) illustrates the system overview that starts from the raw EEG data go through the system to the output. The design of iVAE architecture is illustrated by (b), the encoder and decoder are separately constructed by a Inception module along with a SequenceNet. Moreover, the Inception also consists of a Maxpooling and four SequenceNet units where each SequenceNet includes three functional layers. The index of SequenceNet is the parameter setting of CNN that is to represent (input channel, output channel, 2-D kernel size, padding value), while all the stride in this work we set as 1. The latent space consists of three linear layers as the common VAE.

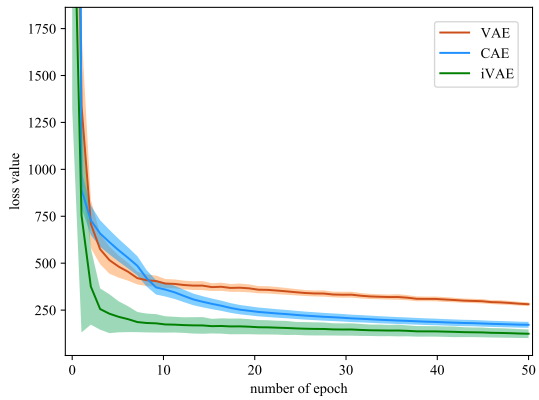


Figure 3.4: Training loss of three VAE architecture. The shadow shows the standard deviation of loss function for different models in three times training.

2.5 Evaluation and Result

To evaluate the feasibility of the Inception module in VAE architecture, we store the reconstructed sprectrograms for each input data and assess the reconstruction of different sleep stages. Given the performance metric of iVAE and a necessity of interpretability, we construct the vanilla VAE and the CAE and compare the performance of them with iVAE architecture. The encoder and decoder of the vanilla VAE and the CAE models are constructed as two layers, with a latent vector of 50 dimensions. Thus, for each model, experiment is implemented three times to analyze the mean of training loss and the generalizability of the models.

Since the inference network can be regarded as the feature generation network. We unitize the latent feature encoder to reconstruct the feature space of original data under fifty dimensions, and then, random forest machine learning model is used to evaluate the generated latent feature. That is, four classes classification is conducted for these VAE models and 80% of the sample as training data is implemented with 5-fold cross-validation.

Figure 3.5 and 3.6 provides the comparison between original input data and reconstructed spectrogram. It can be seen that the model has good performance in the lower frequency band (0-10 Hz) for the wake stage. Moreover, even though the reconstruction become blurry in the high frequency (30–60 Hz), a bound-

ary of power intensity can be seen around 15 Hz from both the original and reconstructed spectrogram. For the REM stage, a distinguishable feature of a relatively high power intensity in the frequency band from 0–30 Hz is preserved in the reconstructed spectrogram.

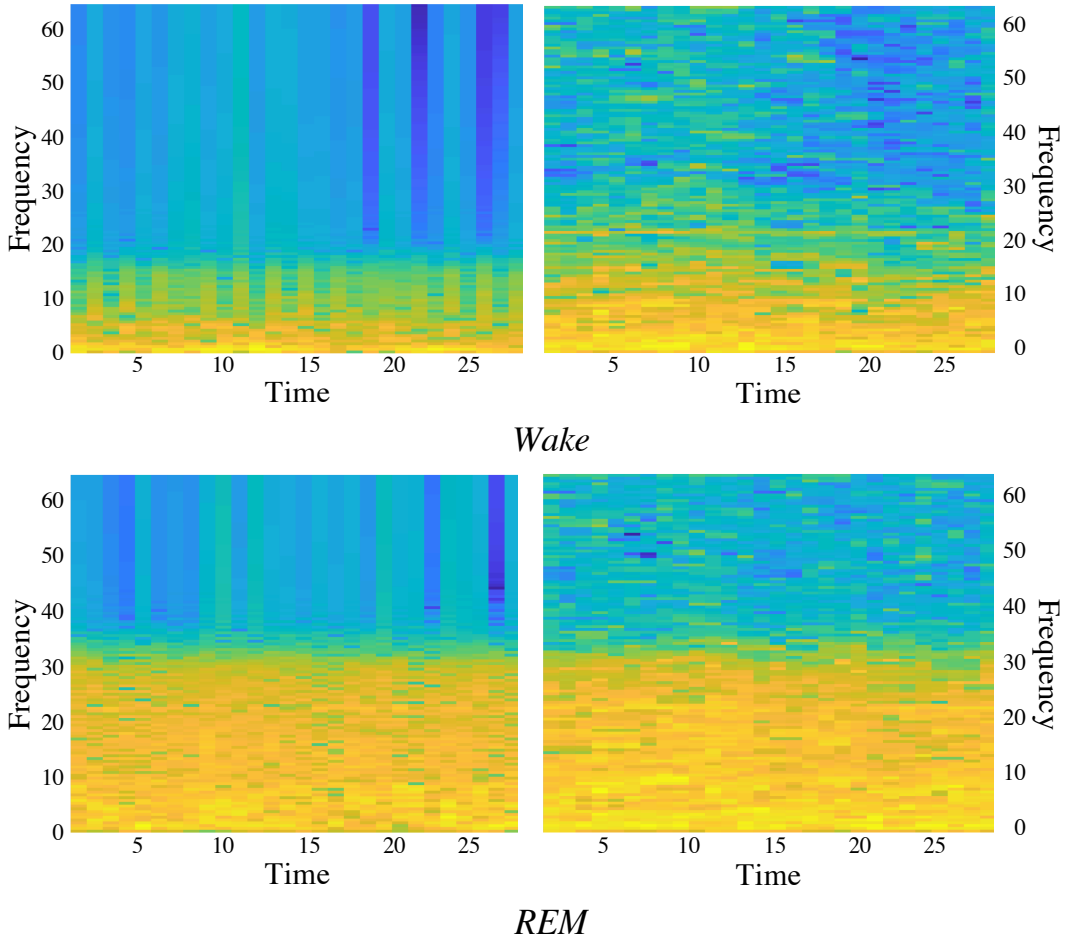


Figure 3.5: Sample of reconstructed spectrograms. The left column shows the spectrograms transformed from raw EEG data, since the right side is the generated image for each sleep stage.

In the lightsleep stage, the brain wave activity begins to shift from the waking state to sleep state, and the power intensity of the mid frequency band (15–30 Hz) increases. This change can be seen by comparing the reconstructed spectrograms of the Wake and the lightsleep stages in Figure 3.6. Moreover, the output

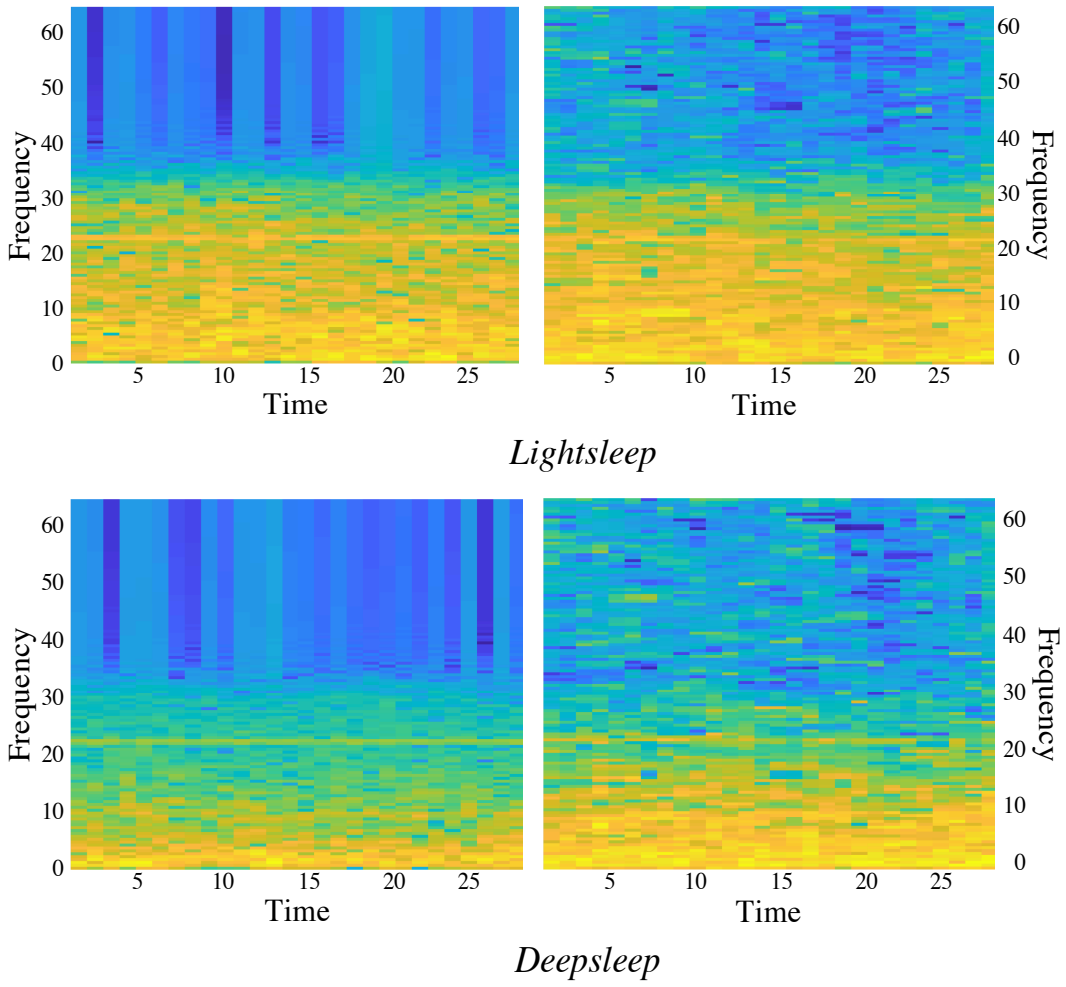


Figure 3.6: Sample of reconstructed spectrograms. The left column shows the spectrograms transformed from raw EEG data, since the right side is the generated image for each sleep stage.

in the deepsleep stage can generally be reconstructed. As it can be seen from the original spectrogram, only the very low frequency (0–4 Hz) has a relatively high power density, and the reconstructed spectrogram can faithfully conserve this characteristic, albeit a little overshoot can be seen in the reconstructed spectrogram.

Architecture	Random forest classification accuracy (in %)
Vanilla VAE	63.57
CAE	66.35
Proposed iVAE	**68.12**

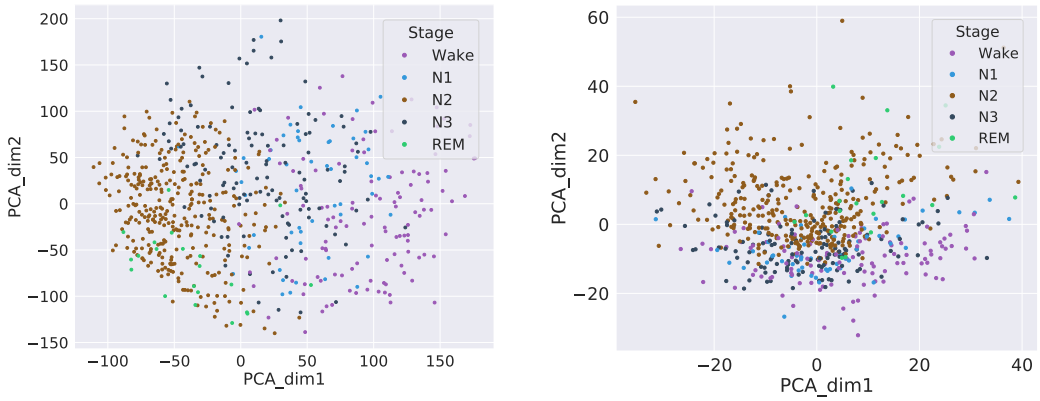
Table 3.2: Classification result of proposed iVAE architecture versus two basic VAE architectures.

The results of training loss of the three VAE architectures can be seen in Figure 3.4. Among the three architectures, the convergence of iVAE is the fastest and the loss of iVAE becomes stable after 10 epochs and the iVAE has the best performance after 50 epochs training. It can be interpreted that the iVAE can learn the stage specific features of input data with less parameters (the number of parameters: 3912449 in vanilla VAE; 2259999 in CAE; 2260345 in iVAE). The classification results using the latent features of three VAE architectures are presented in Table 3.2, where we can see an obvious improvement as we change from the vanilla VAE to the iVAE architecture.

2.6 Conclusion and Discussion

The prior distribution of the latent features in the iVAE is a Gaussian distribution. It matches the current understanding of the EEG that its characteristics follow Gaussian distribution [101]. By imposing the prior the latent features can be further explored and integrated into the pipeline of classification.

The inherent problem of VAE that the output of the decoder can only preserve the major characteristics of the input, is untouched in this paper because according to the medical guideline, only a few major features are used in the



(a) SLPDB + UCD

(b) SHHS

Figure 3.7: (a) shows the clustering results (by PCA) of SLPDB+UCDD dataset, while (b) exhibits the result in SHHS dataset.

stage annotation. From the input-output examples shown in Figure 3.5, we can see that the iVAE can preserve the major changes in different frequency bands.

However, VAE fits the data representation in latent spaces to multiple Gaussian distributions, where the Gaussian assumption is for tractability. Although we make an ablation study that experiments the proposed iVAE with large dataset (SHHS), the clustering results are not satisfied (seen in Figure 3.7). Such strong assumption poses a challenge for EEG modeling, since the EEG data is usually massive, redundant, and noisy. A refinement processing of the feature is necessary.

3 CR2: Symbolic Representation

3.1 Latent Dirichlet Allocation

One representation that the data mining community has been considered transforming real valued data into symbolic representations, noting such representations would potentially allow researchers to avail of the wealth of data structures and algorithms from the text processing and the machine learning [58]. Moreover,

such studies have more recent attention in the sleep stage analysis. Herrera *et al.*, proposed the application of a novel method for symbolic representation of the EEG and evaluated its potential as information source for a sleep stage classifier [33].

The Latent Dirichlet Allocation (LDA) is a Natural Language Processing (NLP) model which is based on the hypothesis that a document has certain topic or a mixture of different topics. When the documents with a similar topic, the topic can be reflected in the particular vocabulary and the probability distribution of words from the dictionary. The LDA facilitates the explanation of dataset by clustering the features of the data into latent unobserved sets. In this work, the symbolic represented epochs can be regards as a series of documents.

3.2 Motivating Example

To meet the criticism and reveal the latent sleep states, Koch *et al.*, utilized symbolic aggregate approximation (SAX) to transform the sleep epoch of EEG to a mixture of probabilities of latent sleep states and developed an automatic sleep classifier using the LDA topic model [49]. Christensen *et al.* inspired the idea of Koch *et al,.* and used the same method to analyze the sleep EEG of people with insomnia disorder with a frequency-based sleep analysis procedure, which is describing each epoch as a mixture vigilance states [16]. However, the proposed SAX has a major limitation, that is, symbols are mapped from the average values of segments. Different segments with similar average values may be mapped to the same symbols, and the represented distance between them is 0 [90].

The topic model is competent in discovering latent, multi-faceted summaries of documents or symbolic data in the NLP community. Therefore, this case study further investigated that the multifold symbolic representation with LDA model from frequency domain is appropriate for representing and manifesting the latent characterization of sleep stages. The study attempted to explore the representation capability of the different statistical methods to evaluate relevant transformation and capture the spontaneous characteristics of different stages. Evaluating by the data-driven method, the EEG-based symbolic representation of sleep stages can be further incorporated into the pipeline of sleep studies.

3.3 Dataset and Preprocessing

The dataset used in this work is from the SHHS. The purpose of SHHS is to test whether sleep-related breathing is associated with an increased risk of coronary heart disease, stroke, all-cause mortality, and hypertension. Access to the SHHS was permitted via the National Sleep Research Resource[3]. The database consists of two rounds of at-home PSG recordings (SHHS-1 and SHHS-2), and only SHHS-1 is used in our work. Nine institutions cooperatively created SHHS-1, for which the full PSG recordings of 5793 individuals are collected between 1995 and 1998. The participants were restricted to those who met the recruited criteria, including, age (older than 40 years), no sleep-related diseases, etc.

Sleep stages were scored by consensus between two sleep technicians blind to the condition of the participants for six classes (wake, REM, S1, S2, S3, S4) according to the R&K guidelines [1]. Noteworthily, this work merges S3 and S4 into stage N3 in reference to the AASM criteria Due to unscored epochs, invalid labels, and misaligned records, 5736 subjects were selected to construct the experimental dataset in this study. Each recording provided two channels (C4-A1 and C3-A2) of EEG, sampled at 125 Hz. The EEG records have been typically contaminated by various types of artifacts, a 8th order Butterworth bandpass filter with cutoff frequencies between 0.5 and 30 Hz was implemented over all records. These cutoff frequencies for band pass filtering were selected since the brain activities have significant information in 0.5 to 30-35 Hz range during sleep [7].

3.4 Symbolization of sleep epoch

As shown in Figure 3.8, to reveal the inherent characteristics of sleep stages, five filtered signals for each EEG record were generated by clinical frequency bands (δ, θ, α, σ, β). The single-sided spectral analysis was then implemented into each band-pass filtered signal by fast Fourier transform within 1 second/bin window and no-overlapping. Three feature extraction methods (trapezoidal integration, spectral entropy and mean) performed the numerical representation of cumulative spectrum on each bin, in order to create subject-wise distribution of spectral

[3] www.sleepdata.org

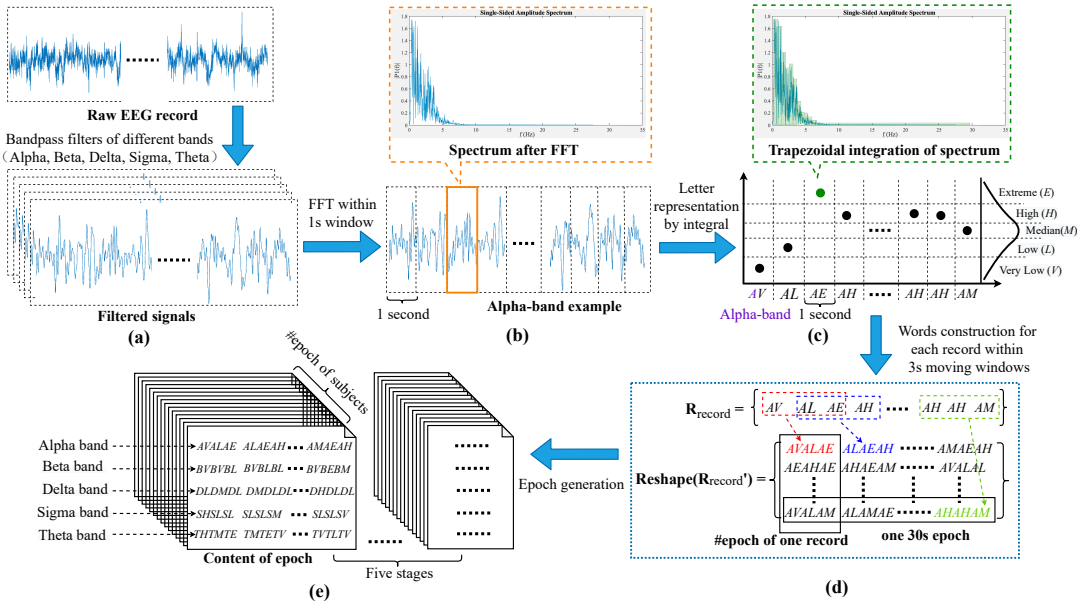


Figure 3.8: The framework of symbolization: (a) illustrates the generation of the filtered signals within five classical frequency bands. The FFT is implemented to generate the spectrum for each 1 second showed by (b). Then, each 1s spectrum is to calculate different features (Mean, Trapezoidal integration, and spectral entropy), subsequently, the numerical values of each record are converted to a word sequence by categorizing the relative distribution. (c) illustrates an example for the Trapezoidal integration process within α-band. A 3 s moving window with one stride is utilized to generate the epochs from each record as shown in (d). Finally, the content of one sleep epoch containing 150 words and five sentences (five frequency bands) is illustrated by (e).

feature within each classical frequency band. Since the generated distributions were normalized to $N(0,1)$, each subject a histogram was produced for each frequency band and the four boundaries producing five equal proportion bins derived. Meanwhile, five quantification categories were labeled as different symbols (Extreme, High, Median, Low, or Very low). Each 1-second numerical value of trapezoidal integration was converted to a word according to the cutoff boundaries and corresponding frequency band. That is, each record was represented into five symbolic sequences.

During the sleep cycles, idiopathic neural oscillatory activities are generated into different sleep stages. For instance, sleep spindles are derived by interplay of the thalamic reticular nucleus during stage N2 of non-REM sleep with a duration of 0.5–2 seconds. To facilitate the mining and representation of the exclusive patterns, a 3-second moving window with one stride (2 bins overlapping) was used to generate the epoch sample which contains 30 6-letter words for each frequency band. Each 6-letter word represents a 3 second spectral pattern of corresponding frequency band (e.g., $AHALAE : \alpha(Alpha) - high - low - median$). Consequently, one sleep epoch was described by 150 words and five sentences.

3.5 Latent Dirichlet Allocation Topic Model

The Latent Dirichlet Allocation is an NLP model which is based on the hypothesis that a document has certain topic or a mixture of different topics. When the documents with a similar topic, the topic can be reflected in the particular vocabulary and the probability distribution of words from the dictionary. The LDA facilitates the explanation of dataset by clustering the features of the data into latent unobserved sets. In this work, the symbolic represented epochs can be regards as a series of documents (seen in Figure 3.9). Assuming the document (epoch) j has 150 words and w_{ij} is the observed value of word i in document j. LDA will traversal and cluster all the words from each document into T topics to label this document by derived the probability of each topic. This process of LDA is described as following:

- For a topic t, a multinomial parameter σ_t is sampled from Dirichlet prior $\sigma_t \sim D(\omega_1)$.

- For a document j, a multinomial parameter φ_j over the T topics is sampled from Dirichlet prior $\varphi_j \sim D(\omega_2)$.
- For a word i in document j, a topic label τ_{ji} is sampled from discrete distribution $\tau_{ji} \sim Discrete(\varphi_j)$.
- The value w_{ji} of word i in document j is sampled from the discrete distribution of topic $\tau_{ji}, w_{ji} \sim Discrete(\sigma_{\tau_{ji}})$.

Where ω_1 and ω_2 are Dirichlet prior hyperparameters. σ_t and φ_t are hidden variables to be inferred while τ_{ji} can be sampled through a Gibbs sampling procedure.

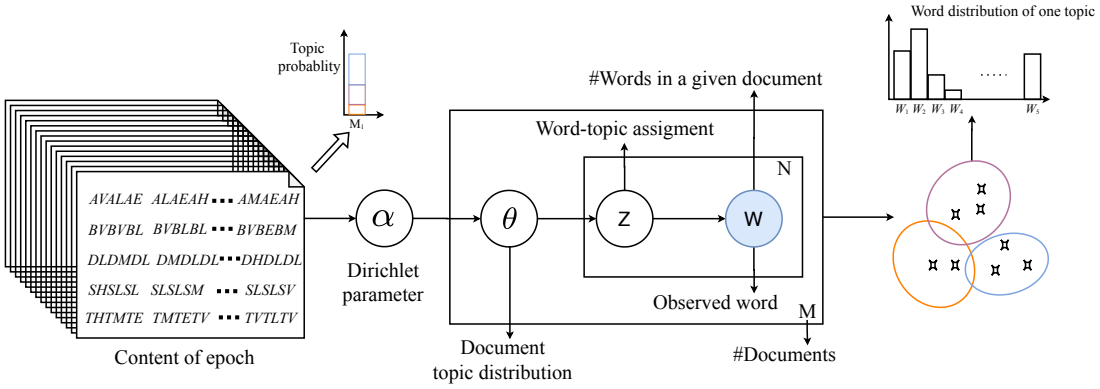


Figure 3.9: LDA is an probabilistic model which is based on the hypothesis that a document has certain topic or a mixture of different topics.

To write about a topic then means to pick a word with a certain probability from the pool of words of that topic. A whole document can then be represented as a mixture of different topics. When the author of a document is one person, these topics reflect the person's view of a document and her particular vocabulary. In the context of tagging systems where multiple users are annotating resources, the resulting topics reflect a collaborative shared view of the document and the tags of the topics reflect a common vocabulary to describe the document.

3.6 Evaluation and Result

A grid search [5:1:13] for the number of topics was first to find the optimal distributions of words based on the topic coherence metric within different feature en-

gineering. The topic coherence measures score a single topic which is the artifact of statistical inference by measuring the degree of semantic similarity between high scoring words in the topic. This word utilized C_v measurement which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that by using mutual information and the cosine similarity. Moreover, the LDA results (topics) had been decomposed via principal component analysis (PCA) to visualize the topic construction and distribution by using the pyLDAvis python packages [4]. To map the topics to the sleep epochs, the probability of topics for epochs were resulted by trained LDA model, and the statistical distribution was generated to evaluate the co-occurrence and importance for each sleep stage.

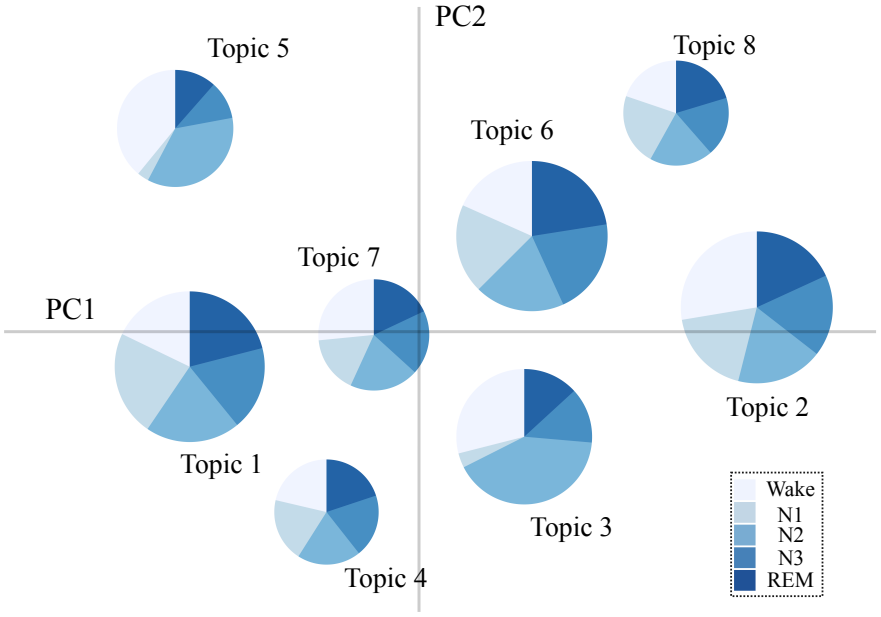


Figure 3.10: The visualization of PCA of 8 topics. The histograms show the word distributions and estimated word frequency for each topic. The distribution of the pie chart illuminates the statistical contributions for five stages and the area of each pie chart is proportional to how many documents feature each topic.

Table 3.3 illustrates the coherence metrics for different methods of feature extraction. Over all the grid search of topics, the trapezoidal integration of spec-

[4]https://pypi.org/project/pyLDAvis/

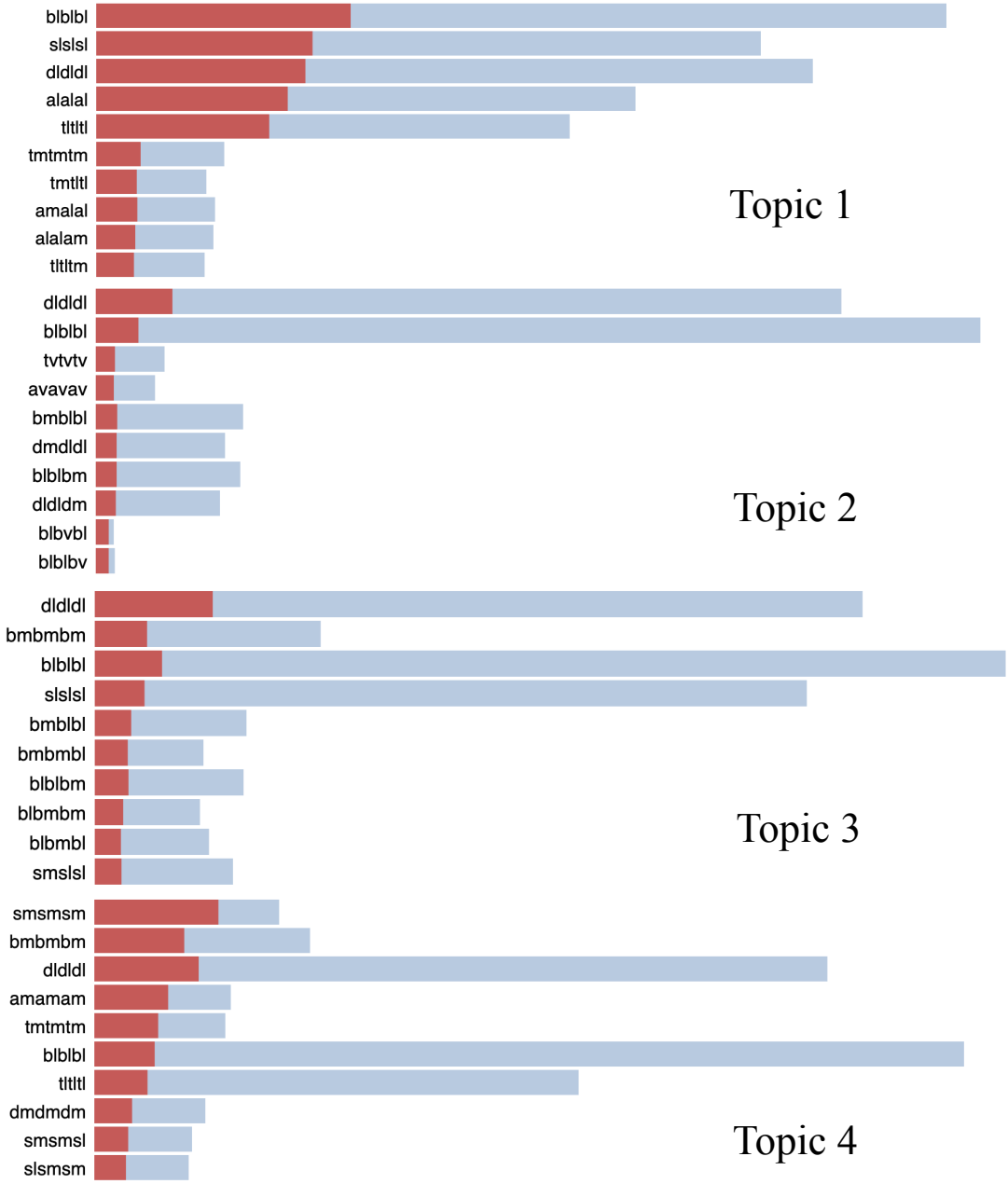


Figure 3.11: Topic 1 to 4

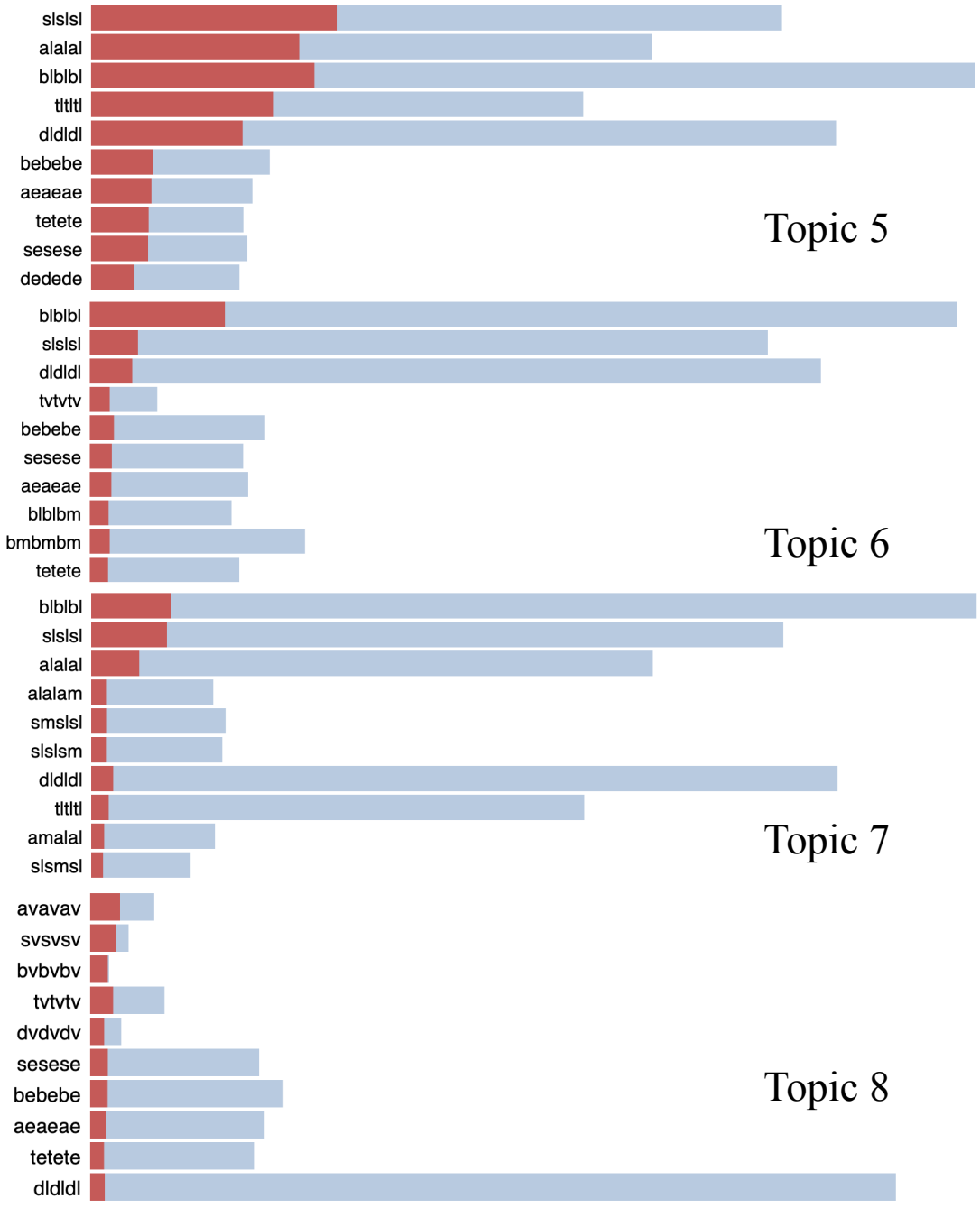


Figure 3.12: Topic 5 to 8

trum has a better performance than the common SAX (mean value) and the spectral entropy. Meanwhile, the fine-tuning of the number of topics is 8 in this word shows the highest coherence for trapezoidal integration.

Table 3.3: A grid search of number of topics for different feature process methods

#Topics	5	6	7	8	9	10	11	12	13
Mean Value	0.281	0.275	0.305	0.316	0.319	0.298	0.291	0.287	0.253
Trapezoidal Integration	0.313	0.306	0.327	**0.351**	0.316	0.295	0.315	0.314	0.314
Spectral Entropy	0.229	0.226	0.234	0.228	0.224	0.222	0.224	0.221	0.213

Table 3.4: Top 3 importance of topics for five sleep stages

Stages	Wake	N1	N2	N3	REM
Top 3 topics	T2; T5; T7	T1; T8; T4;	T1; T5; T8	T5; T6; T3	T1; T8; T5

The PCA result of the optimal number of topics for trapezoidal integration has shown in Figure 3.10, 3.11, and3.12. The topics (8-dimensional in our case) were flattened to be only 2-dimensional. There is no overlapping area among topics, that is, the topics and their word distributions are mutually independent. The center of pie charts represent the position of 8 topics in the latent feature space, while the distances between topics illustrate similarity or dissimilarity. Moreover, according to the word distributions, it can group topics 6 and 8 into one since the pie distribution are more similar, and have half estimated very low frequency of θ over all term frequency.

The topics 2, and 7 are mainly contributed to wakefulness while these topics have a smaller metric distance. Topics 1 and 2 have high term frequency for σ and β bands according to Figure 3.10 and 3.11, and the high term frequency of θ band can be reflected to N2 sleep. Topics 5 is related to N2 and wakefulness stages as shown in Table 3.4, while the amount of estimated words of θ band (e.g., $tltltl : \theta - low - low - low$) is represented to REM stages. Table 3.4 lists the top 3 importance of topics for different stages. The N1 and N2 have similar construction of topics since these two stages belong to light sleep. The most relevant term for topic 3 is the δ band word while the delta waves (more than 20%) are more related to stage N3.

3.7 Conclusion

In this study, we investigated the symbolic representation of EEG records for sleep stages. Comparing with the different feature extraction methods, the work proved the feasibility of an improved representation method of sleep stages and physiological signals based on the data-driven. The further study will extend to systematically explore the sleep construction and to evaluate the downstream study in the sleep community.

4 Patch Embedded Representation

4.1 Embeddings

In NLP, word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

Word embeddings can be obtained using a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves the mathematical embedding from space with many dimensions per word to a continuous vector space with a much lower dimension. Methods to generate this mapping include neural networks, dimensionality reduction on the word co-occurrence matrix, probabilistic models, explainable knowledge base method, and explicit representation in terms of the context in which words appear. Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing and sentiment analysis.

Due to the appealing progress of self-attention architecture, some attention-based attempts for automatic sleep scoring have recently been proposed [15, 27, 53]. Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regarding to their distance in the input or output sequences [99]. In particular, Transformer, a model architecture relying entirely on an attention mechanism, has reached a new state of the art in the visual classification tasks

[24, 60], and has been extended to sleep scoring [15]. Similar to the word embedding, the visual task divides one input image into a set of sub-images (patches). Each patch containing a self local information is fed to the attention-based neural networks. This patch embedded process allows the model to query the local significant feature from the global context.

4.2 Motivating Example

Considering the parallel computation, recent works attempt to introduce the attention mechanism to the sequential model. In the CNNs-LSTM model proposed by Chen *et al.* [14], an attention network was merged to adjust the significance of two groups of features. Sun *et al.* [89] proposed a multi-flow RNN which is utilized to learn temporal information by fusing the hand-crafted features and network-trained features. The proposed model was evaluated by 147 full night recordings with an overall accuracy of 0.88, and an F1-score of 0.82. Among these preceding studies, Qu *et al.* applied residual blocks to the EEG signal after Hilbert-Transform-like preprocessing and the Transformer on the epoch level for accurate sleep scoring[80]. The results are plausible since the macro sleep structure is relatively constant with cyclic patterns and can be modeled by using sequence-to-sequence strategy. According to the AASM guideline, sleep stages can be generally determined on the basis of intra-epoch signals. Moreover, as has been validated both computationally and experimentally, the structure of shorter signals embeds information related to sleep stages [16, 96].

Inspired by the embedding scaling successes in visual tasks, this case study introduces the patch embedding to time-frequency spectrogram and explores the feasibility representation for sleep transient stage-dependent frequency features. Assuming the mentioned transient characteristics can be revealed into 1-second EEG signals. Here, I segment the spectrogram of each epoch into the temporal patches, in order to capture the transient characteristics from 1-second frequency band. Afterward, a sequence of linear embeddings of these patches is fed into a attention model expecting to parallel find the global dependencies among the 1-second duration.

4.3 Dataset and Preprocessing

The dataset used in this study is same to the *CR2* from the SHHS. Due to the unscored epochs and misaligned records, here, we use only the SHHS Visit 1 containing two channel EEG records (C4-A1 and C3-A2) from 5736 subjects sampled at 125 Hz. The EEG records have been typically contaminated by various types of artifacts, a 8th order Butterworth bandpass filter with cutoff frequencies between 0.5 and 30 Hz is implemented over all records. According to our previous work, the spectrograms (size: 30×30) are finally generated for each filtered 30s epoch. This transformation utilizes the log-power spectra via a consecutive Fourier transform with 1-second segmentation, no overlap, and 125 sampling points Fast Fourier Transform. Table 3.5 contains the description of dataset, and the preprocessing has been shown in Figure 3.13.

Table 3.5: Dataset Description of SHHS

	Wakefulness	N1	N2	N3	REM
#Sample	2371496	166619	809155	732389	214985

4.4 Embedding and Classification Model

Inspired by this visual attention work [24], we segment each spectrogram $S \in \mathbb{R}^{F \times T \times C}$ into a sequence of patches $\widehat{S} \in \mathbb{R}^{T \times (F \cdot C)}$. T is the number of patches (30 patches) for generated sequence $\widehat{S}$ so that each patch represents the 1-second brain waves in frequency domain from two channels. Then all patches are mapped through a linear projection layer to a high dimension D (64 in this work) for further training step. This study we adopt a attention model, e.g., Transformer, as the classification model (The detail of Transformer will describe in following Part II). Attention is a parallel mechanism, that is, without merging the relative or absolute information, it can be regarded as a bag-of-words model. Similar to [24], we inject the parameterization positional embedding E_{pos} to the patch sequence. An extra class token S_{Cls} is appended which is compressed the global relevance and used for the final classification task. The input sequence of Transformer

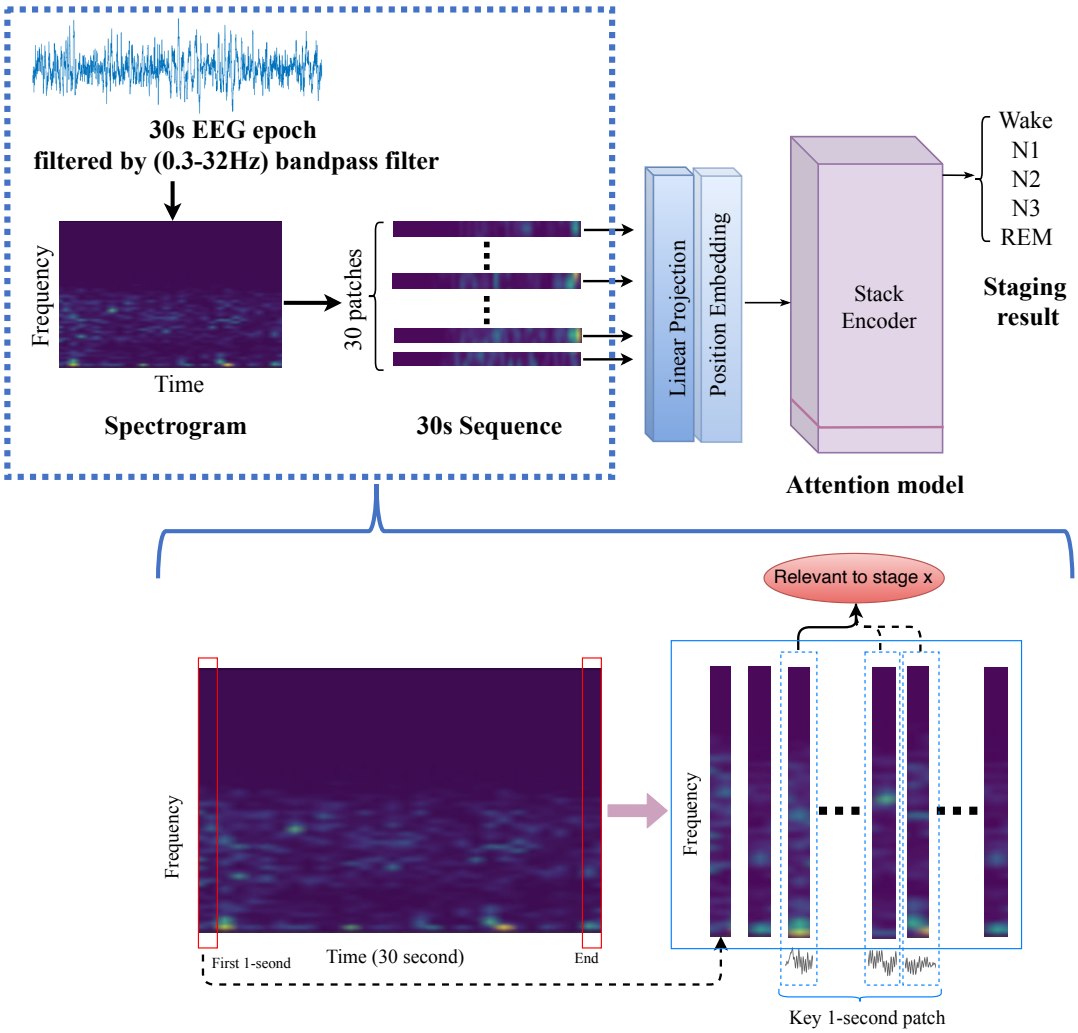
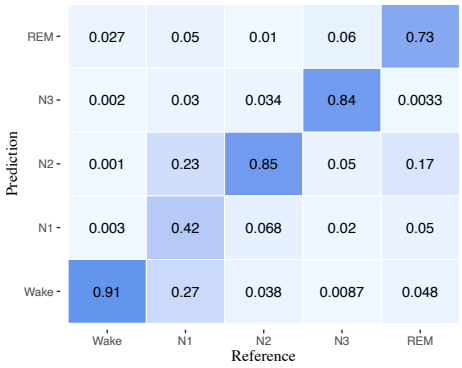


Figure 3.13: The framework of sleep scoring system. Each 30s EEG epoch is first transferred to a time-frequency spectrogram (size: 30×30). Then 30 patches are segmented with a 1-second moving window without overlapping. Each patch as a 30×1 vector is projected to a high dimension by using to a fully connected layer. The generated sequence feeds to the attention model. A softmax layer is finally applied to the trained token to result the classification tasks.

		Wake	N1	N2	N3	REM
Baseline Method [64] (Enrique *et al*, 2020)	Pre	0.91	**0.54**	0.84	0.83	**0.87**
	Re	0.91	**0.22**	**0.91**	0.82	**0.82**
	F1	-	-	-	-	-
Cosine Method	Pre	0.89	0.41	0.79	0.83	0.72
	Re	0.88	0.19	0.86	0.80	0.63
	F1	0.88	0.26	0.82	0.81	0.67
Learnable Method	Pre	**0.91**	0.42	**0.85**	**0.84**	0.73
	Re	**0.93**	0.20	0.89	**0.83**	0.76
	F1	0.92	0.27	**0.87**	**0.84**	0.74

(a)



(b)

Figure 3.14: (a) shows the comparison between our proposed method and baseline model which evaluated the same dataset and EEG channels; (b) exhibits the corresponding confusion matrix for five stages by using the parameterized position embedding.

model is defined as follow:

$$X = [S_{Cls}; E \cdot \widehat{S}] + Parameterize(E_{pos}) \tag{3.3}$$

where $E \in \mathbb{R}^{(F \cdot C) \times D}$ represents the linear projection of flattened patches. The paremeterized embedding $E_{pos} \in \mathbb{R}^{(F \cdot C+1) \times D}$ is then added to extended patch sequence. In this study, we also evaluate the positional embedding method of the original Transformer. The proposed cosine function allows the model to learn to attend by relative positions. It assumes the closer 1-second patches will provide more potential information when one patch is salient for its sequence.

4.5 Experiment and Result

The result of the proposed method by using two positional embedding methods and comparison performance [28] have been shown in the table in Figure 3.14 (a). Comparing the traditional positional embedding (cosine-method), the learnable method has a better overall performance with a 0.76 ACC and 0.73 macro-averaging F1-score.

The confusion matrix in Figure 3.14 (b) provides the details about the model performance and classification result. It can be seen that Wake is clearly sep-

arated without a large number mismatching. Focusing on the N2 and N3, the model is still able to provide an accurate performance that classifies the N2 and N3 from the other stages. However, the challenge of the N1 stage still remains, due to the unbalance training data. For the REM stage, without considering the movement information of the eyes (EOG) [63], the results need more improvements.

Overall, the comparative results between the proposed method and baseline work which experiments with using the same dataset and EEG channels can be seen the table in Figure 3.14 (a). By using the sole Transformer model, our proposed result has a competitive performance for most of the stages. However, the stage N1 and REM staging should improve more in future work.

4.6 Conclusion

Inspired by the patch embedding in visual task, we experimented with the sleep stage classification by applying the time-frequency embedded representation based on the large EEG database. Comparing with the baseline method which was evaluated in the same dataset, the proposed method has a good performance for the Wake, N2, and N3 stages. Moreover, the work proved the transient characteristics of the frequency domain can be parallel captured from different 1-second time positions. The further study will try to overcome the misclassification of the N1 and REM stages.

5 Summary and Discussions

Sleep screening based on the construction of sleep stages is one of the major tool for the assessment of sleep quality and early detection of sleep-related disorders. Due to the inherent variability such as inter-users anatomical variability and the inter-systems differences, representation learning of sleep stages in order to obtain the stable and reliable characteristics is runoff for downstream tasks in sleep science. In this chapter, the thesis describes three EEG-based representation learning for the different sleep stages.

CR1, proposed the inception-inspiring VAE structure (iVAE) to characterize and reconstruct the EEG spectrogram of four sleep stages. The iVAE is

constructed by replacing the main structure in encoder and decoder with the inception-like module, which will capture the multi-scales features of the spectrogram of EEG. By comparing with the vanilla VAE and the CAE, a more accurate reconstruction and a better classification using the latent features of the iVAE can be confirmed. This study also proves the possibility of automate generation of latent representation of sleep stages by using unsupervised generative model and suggests a further optimization of the iVAE structure will benefit the sleep stage characterization. Moreover, we find that comparing with the different feature extraction methods, the symbolic representation (*CR2* proved the feasibility of an improved representation method of sleep stages and physiological signals based on the data-driven method. In *CR3*, we found the re-segmented spectrogram+Transformer framework has a powerful stage scoring ability. This framework should be refined and incorporated into the pipeline of sleep stage classification.

Part II

Accurate Staging Framework and Decision Interpretability

4 | Mechanism-based End-to-End Staging Pipeline

1 Introduction

Deep neural networks (DNNs) have recently achieved success in the sleep stage scoring community [75, 76, 102]. One of the key attributes of DNNs is their ability to automatically reconstruct a more relevant feature space from a large amount of sleep data. Meanwhile, it takes advantage of the adjustment of the parameters of the feature map layers by backward propagation to fit the representation functions of different sleep stages. Although the performance of the automatic scoring algorithm has been greatly improved by deep learning models as well as free access to large-scale sleep databases [4, 9, 72, 80, 92], the optimization of the learning pipeline is still at the midway point. Specifically, the representation strategy for EEG signals and the learning structure of the network need precise optimization to adapt to the intrinsic traits of autonomic sleep scoring based on EEG signals.

With the optimized combination on these two aspects, we hope to improve the overall performance of automatic sleep stage scoring on one hand, and to shed light on clinical/physiological significance by retaining the interpretability of the model as much as possible. In the following two sub-sections, the rationale for why the aforementioned special care should be given will be explained, whereafter the motivation of this work will be elicited.

1.1 Notes on EEG Representation

In the Part I, we have introduced the EEG representation. The end-to-end framework implemented by deep learning models in recent years has made feature generation more flexible and problem-specific [75, 80]. The subject-wise generalization gap, especially in models with complicated representation learning, can be problematic [81]. Therefore, where to draw a line between predefined and adjustable representation generation should be discussed. We can think of predefined-feature-only input and raw EEG input to a learning model as two extremes and assume that a better solution lies between them.

1.2 Notes on Learning Structure

The need to delicately select learning structures is rooted in the origin of the EEG signal and the intrinsic characteristics of neuron activities in the cortex, that is, the spatial heterogeneity of relevant features in the cortex [7], [26], which means that the corresponding features may not always be seen from one lead, while stage-irrelevant features/disturbance may be seen [78]. This situation is even made tougher by the transient and temporally random natures of some dominant characteristics [18, 45]. For instance, sleep spindle is the defining characteristic of N2 stage, however, it is a spontaneously burst and only last for a short duration of 0.5–1.5 seconds. Therefore, in N2 stage, sometimes the stage-irrelevant features dominate the current epoch[2]. As a consequence, the stage-dependent features are not always unequivocal among the same staged EEG epochs. Therefore, in addition to the feature representation, stage-specific feature refinement in the global content is needed.

The sleep community has witnessed significant progress with the merging of deep neural networks in automatic sleep scoring [4, 72, 80]. It is expected that the separated feature engineering study and classification study to be superseded by the end-to-end framework[23, 85].

Due to CNNs becoming the defacto standard for visual classification tasks, CNNs have been used in feature extraction from the frequency or time-frequency domain [4, 80, 92]. By interleaving a collection of multi-size convolutional filters with non-linear activation functions and downsampling operators, proposed

scoring frameworks expect to capture sophisticated local features and attain an inductive representation of the EEG epoch [35]. Since CNN does not consider the local features against the global context, it will results in a local inductive bias [98]. Consequently, it may draw more attention from the classifier onto a pattern of the neighboring area while distracting the classifier from really important but transient information and its relevance. Meanwhile, feature extraction in each EEG epoch is handled by the sharing weights of filters. This mechanism expects key rhythms to occur in a relatively inherent position. Therefore, there is a mismatching between the translation-invariant constraint of CNNs and the temporally random and transient nature of the relevant defining characteristics.

Recurrent neural networks (RNNs), such as long short-term memory (LSTM) and gated recurrent neural networks, have been established as the state-of-the-art (SOTA) in automatic sleep scoring [75, 80, 85, 89]. Different from CNNs, the recurrent-based models pay attention to the information of the global context. By generating a set of hidden states, the decision-making considers the influences of a sequence of previous time steps or future steps (by the Bidirectional LSTM) [52]. Given the issue of transiently bursting rhythms, such as non-periodic sharp-wave ripples and spindle activities [2, 10], RNNs cannot be regarded as the best choice for sleep scoring. Further, the inherently sequential processing flow of RNNs precludes the possibility of parallelization within feature capture [99].

Although some advanced deep frameworks have been proposed, for instance, the sequence-to-sequence framework that can handles a set of consecutive sleep epochs by simulating the transition rule of inter-stage[75], epoch-wise automatic scoring frameworks remain the limitations. It implies that a context-sensitive flexible pipeline that automatically adapts attributes of data itself is a requirement.

1.3 Goal and Contributions

Given the considerations introduced above, this work proposes a tailored pipeline for autonomic sleep scoring by proposing a novel way of generating features that alleviate the influence of temporally random and transient nature of the EEG features while retaining the resolution in the frequency domain. This method can find its clinically informative explanation, which will be explained in Section

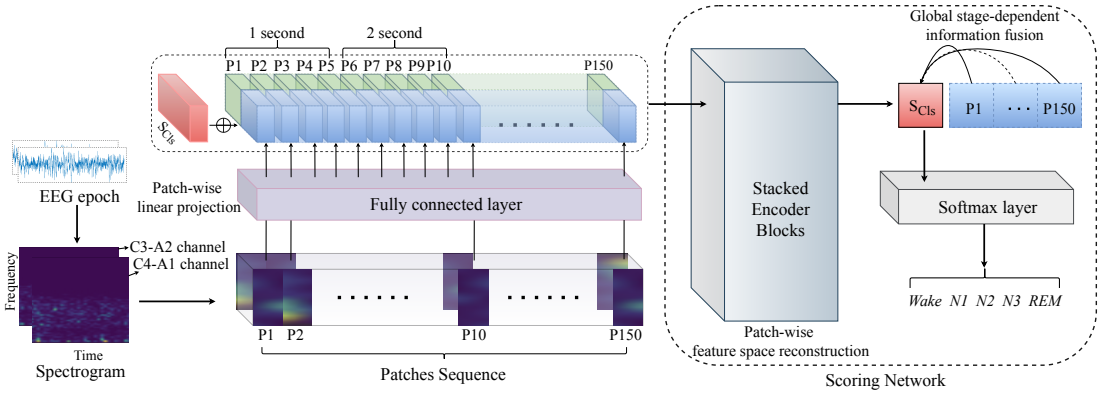


Figure 4.1: Framework of sleep scoring pipeline. Experiment of this work was done with two channel EEGs. Each 30-second EEG epoch is first transferred to time-frequency spectrogram (size: 32×30). Then, spectrograms of the 2 channels are segmented into patch sequence respectively. Each patch as 1-second-1-frequency-band feature vector is projected to high dimension by using patch-wise fully connected layer. Positional embedding procedure adds relative positional information to patch sequence, while extra class indicator is also concatenated. Afterward, augmented patch sequence is fed to scoring module model that contains stacked encoder blocks and the final classification layer. After training, class patch that absorbs intra-patch characteristics is used for stage assignment decision.

3.1 and 3.2.

It can also be seen that epoch-wise contextual relating of temporally local features is necessary. The time-frequency feature representation is first organized as a time series of features of the frequency domain and then input into an elaborate attention-based deep learning model. The attention mechanism is designed to extract the global contextual relevance between units of a time series signal or stream of text. Therefore, it is regarded as a suitable learning scheme in this study.

The attention-based model, including the Transformer [24, 60], has been tried in sleep scoring and relevant problems [15, 53, 80]. Among these preceding studies, Qu *et al.* applied residual blocks to the EEG signal after Hilbert-Transform-like preprocessing and the Transformer on the epoch level for accurate sleep scoring[80]. The results are plausible since the macro sleep structure is relatively constant with cyclic patterns and can be modeled by using sequence-to-sequence strategy. According to the AASM guideline, sleep stages can be generally determined on the basis of intra-epoch signals. Moreover, as has been validated both computationally and experimentally, the structure of shorter signals embeds information related to sleep stages [16, 96].

Specifically, this work transforms two channels of EEG signals into spectrograms, which are then divided into five parts following the five frequency bands as indicated in TABLE 3.1 and the power spectral density is calculated for every 1 Hz (frequency patching). Whereafter, each part is further partitioned temporally into 1-second bins (time patching). Finally, the resultant patches are input into an epoch-wise scoring model. With the attention matrix output by the trained attention-based model, this pipeline is expected to show physiologically interpretable patterns that are important in stage classification. The results also show that our methods reach a new state-of-the-art performance in an experiment with a large clinical dataset (Sleep Heart Health Study) of 5736 patients.

1.4 Chapter Organization

The remainder of this chapter is organized as follows: Section 2 introduces the details of dataset and its preprocessing. Section 3 describes the details that how to represent the stage-dependent characteristics in EEG data, while the subsequent

framework is reported in this Section. Section 4 introduces the experiments of the case study and the details of model parameter settings. Section 4 shows the results of framework performance. Finally, we conclude this chapter and discusses the implications from the findings in Section 5.

2 Database and Preprocessing

The dataset used in this work is from the Sleep Heart Health Study (SHHS). The purpose of SHHS is to test whether sleep-related breathing is associated with an increased risk of coronary heart disease, stroke, all-cause mortality, and hypertension. Access to the SHHS was permitted via the National Sleep Research Resource. The database consists of two rounds of at-home PSG recordings (SHHS-1 and SHHS-2), and only SHHS-1 is used in our work. Nine institutions cooperatively created SHHS-1, for which the full PSG data of 5793 individuals are collected between 1995 and 1998. The participants were restricted to those who met the recruited criteria, including, age (older than 40 years), no sleep-related diseases, etc.

Sleep stages were scored by consensus between two sleep technicians blind to the condition of the participants for six classes (wake, REM, S1, S2, S3, S4) according to the R&K guidelines [1]. Noteworthily, this work merges S3 and S4 into stage N3 in reference to the AASM criteria Due to unscored epochs, invalid labels, and misaligned records, 5736 subjects were selected to construct the experimental dataset in this study. Each recording provided two channels (C4-A1 and C3-A2) of EEG, sampled at 125 Hz.

Considering the issue of imbalance in the dataset, a sample balanced subset called 'healthy-set' was extracted. Here, this subset served in the pre-training phase in our experiments. Since SHHS-1 provides a personal health description of all subjects, the subjects were selected on the basis of inspection of the clinical criteria. The SHHS provides the health status assessments of each participant before and after the at-home PSG experiments. Considering the subject is in relatively good health status can be regarded as a more general sample in the dataset, we screen the subjects and construct a healthy set from below six aspects.

- *Prev_ang:* Number of Angina Episodes Prior to Baseline Angina episodes

prior to baseline PSG;

- *Prev_ chf:* Number of Congestive Heart Failure (CHF) episodes Prior to Baseline Congestive Heart Failure (CHF) episodes prior to baseline PSG;
- *Prev_ mi:* Number of myocardial infarctions (MIs) Prior to Baseline Myocardial infarction (MI) prior to baseline PSG;
- *Prev_ mip:* Number of Procedures Related to Heart Attack Prior to Baseline Myocardial infarction (MI)/procedure prior to baseline PSG;
- *Prev_ revpro:* Number of Revascularization Procedures Prior to Baseline Revascularization procedure prior to baseline PSG;
- *prev_ stk:* Number of Strokes Prior to Baseline Stroke prior to baseline

where *Prev_ ang, Prev_ chf, Prev_ mi, Prev_ mip, Prev_ revpro, and prev_ stk* are variable name in SHHS-1 dataset description [1]. The healthy-set consists of 684 subject-wise recordings with 26080 epochs for each class. A summary of the experimental dataset used in this work is shown in TABLE 4.1.

Table 4.1: SHHS database description

	All	Healthy-set
#Subject	5736	684
Gender	M: 2774 F: 2962	M: 360 F: 324
Age (year)	62.17±11.02	63.14±11.22
#W	1666191 (28.8%)	
#N1	214985 (3.7%)	
#N2	2371496 (40.9%)	26080 (20%)
#N3	732389 (12.6%)	
#REM	809155 (14.0%)	

[1] https://sleepdata.org/datasets/shhs/files/datasets

3 Sleep Mechanism-based Framework

The proposed framework, the overall illustration of which is shown in Figure 4.1, is a sequence-to-target workflow. A spectrogram is transformed into a sequence of patches of five frequency bands, augmented with class tokens, and finally input into our attention-based model for sleep classification. This section is divided into four sub-sections, the first two of which introduce the generation of time-frequency representations (3.1) and a novel framework for feature organization (3.2). The second half focuses on the network architecture, which includes the sequence augmentation and embedding of the input (3.3) and the architecture of the model (3.4).

3.1 Time-Frequency Representation

EEG recordings are typically contaminated with various types of artifacts. An 8th order Butterworth bandpass filter with cutoff frequencies between 0.2 and 32 Hz was applied to all recordings. Whereafter, each epoch is transformed into a spectrogram to manifest its time-frequency characteristics. Given that the advantage of short EEG segments (1-second) in expressing the transient sleep rhythms has been shown [13], a log-power spectrogram is generated for every 1-second EEG signal by using a non-overlapping Hamming window and fast Fourier transform and then calculating the integral in power spectrum every 1 Hz. With its time frequency resolution, the spectrogram possesses sufficient information for sleep stages, whose local or contextual features can find their correspondence in the spectrogram.

Therefore, for one epoch, two spectrograms that correspond to the two EEG channels are generated. The initial feature is denoted as $S \in \mathbb{R}^{F \times T \times C}$, where F denotes the frequency range (0-32 Hz), T denotes the time (30-seconds), and C denotes the channels (two channels).

3.2 Frequency-time patching

In this part, we propose an important processing framework termed *frequency-time patching* for organizing the input spectrogram. That is, the spectrogram is

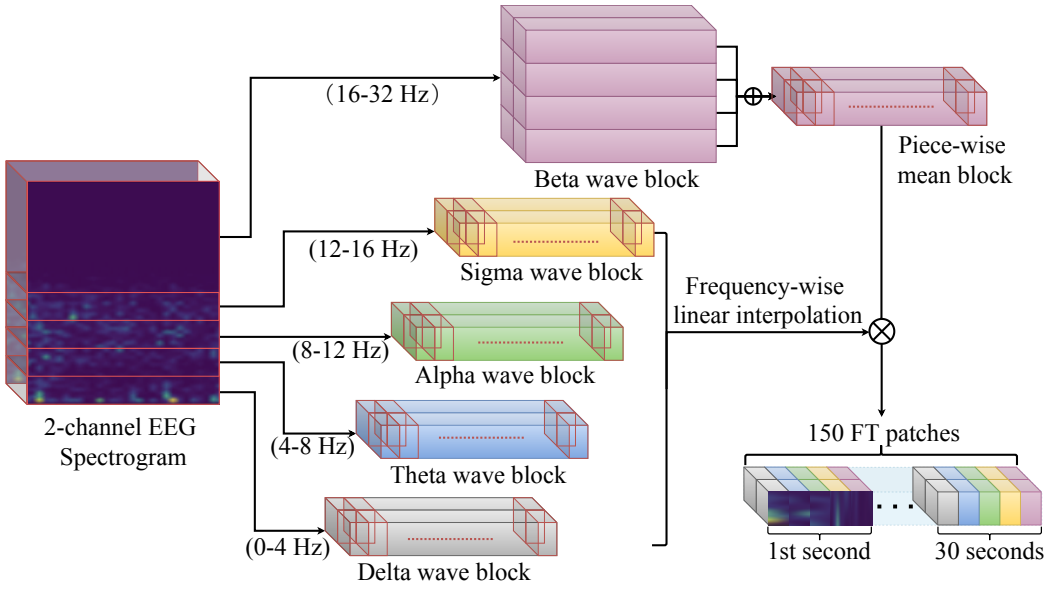


Figure 4.2: Workflow of time-frequency patching

divided into eight frequency bands every 4 Hz in accordance with the five frequency bands of TABLE 3.1, where the *beta* band is further divided into four segments (*frequency patching*). Afterward, time patches are acquired by extracting and rearranging each column (1-sec) of the spectrogram (*time patching*). This framework embodies two scientific findings, that is, the concurrence of sleep stages in one epoch [26] and the fact that most of the clinically crucial features can be represented by the time-frequency features at a 1-sec resolution [18]. With the special structure of the network architecture, which will be introduced in 3.4, the *frequency-time patching* structure will be kept consistent throughout the network to facilitate the discussion of model interpretability. Along with Figure 4.2, *frequency-time patching* is introduced herein.

- Frequency patching: Each S is split into five parts $S = (S_\delta, S_\theta, S_\alpha, S_\sigma, S_\beta)$ in accordance with the five predominant frequency bands (TABLE 3.1) of sleep rhythms. The first four spectrogram parts, S_δ, S_θ, S_α, $S_\sigma \in \mathbb{R}^{(F/8)\times T\times C}$, have the same bandwidth of 4 Hz. A block of the beta band is quartered into four sub-blocks, $S_\beta = (S_{\beta 1}, S_{\beta 2}, S_{\beta 3}, S_{\beta 4})$, where the subscripts $\beta 1 \sim \beta 4$ correspond to the four quarters of the beta band.

- Time patching: Each frequency block as shown in the middle of Figure 4.2 is divided by column to extract frequency-time patches. Therefore, one time patch is S^i ($S^i = (S^i_\delta, S^i_\theta, S^i_\alpha, S^i_\sigma, S^i_\beta)$, where the superscript i denotes the *i-th* second.

- Sub-block averaging: For the frequency-time patches in the Beta band, mean values are calculated and used in the following steps, so that $\bar{S}_\beta = (\bar{S}_{\beta 1}, \bar{S}_{\beta 2}, \bar{S}_{\beta 3}, \bar{S}_{\beta 4})$.

- Spectrogram transformation: For each time patch S^i, the new structure is $S'^i = (S^i_\delta, S^i_\theta, S^i_\alpha, S^i_\sigma, \bar{S}^i_\beta) \in \mathbb{R}^{20}$, resulting in the transformation of the original spectrogram S into $S' \in \mathbb{R}^{20 \times T \times C}$.

- Patches rearrangement: A sequence of frequency-time (FT) patches is generated by column-wise traversal. Therefore, $S_{seq} \in \mathbb{R}^{150 \times D_p}$ ($150 = 5$ frequency bands $\times 30$ sec) is the patch sequence for one epoch, where D_p denotes the dimension of each patch $\in \mathbb{R}^4$.

To validate the effect of the *frequency patching*, we compare its performance with time patches (1-second) input to the same network architecture. The results of the baseline model can be found in TABLE 4.4, the (Time patching + proposed model) pipeline.

3.3 Patches Sequence Embedding

As indicated in above, physiologically, the individual FT patches contain uneven information about sleep stages and the concurrence of patches along with its sequential distance (time difference) is informative. Therefore, the self-attention, is a suitable mechanism for the process of feature extraction, and the attention-based classification model have been used computer vision [20, 24].

To construct our model, an extra class token S_{Cls} is inserted to each sample at the beginning of the FT patches sequence. With the processing and propagation of the input through the model layers, this token are retained as the stage indicator, and will be used in the final stage scoring/classification step. These two steps can be formulated as below:

$$S'_{seq} = Concat(S_{Cls}, E \cdot S_{seq}), \tag{4.1}$$

where E is a patch-wise linear projection that project the two FT patches of two EEG channels of the same time into higher dimension. Consequently, $S'_{seq} \in \mathbb{R}^{(150+1)\times D}$ is the output sequence, where D is the dimension of output of the linear projection.

Because The attention mechanism does not differentiate the position of the keys when calculating the attention values of a query, the position embedding is a routine supplementary. In adapting the mechanism for sleep staging, where the global position could be a nuisance, we make use of the positional encoding technique [24], by which a suitable positioning scheme can be generated by training. After, the position embedding is merged into the S'_{seq} to form final input

$$X = S'_{seq} + Parameterize(E_{pos}), \tag{4.2}$$

where $E_{pos} \in \mathbb{R}^{(150+1)\times D}$ with the same shape of S'_{seq} represents the parameterized embedding sequence.

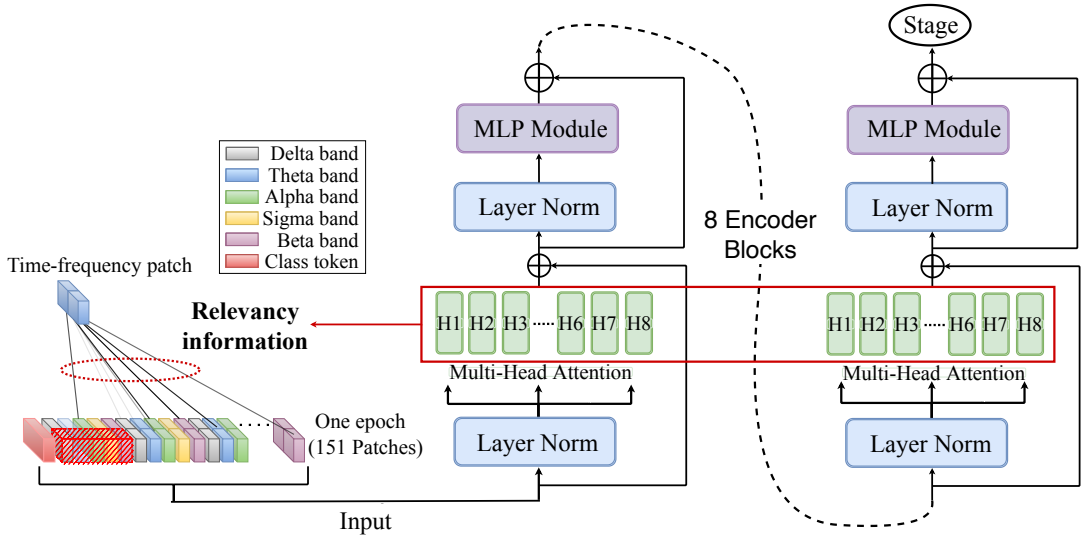


Figure 4.3: Architecture of the stacked encoder blocks

3.4 Scoring network

The frequency-time patches go through the network architecture as shown in Figure 4.3. Augmented by *Multi-heading* (see the red box in Figure 4.3) the attention layer is applied to delineate the contextual relevance of FT patches in the

Multi-layer perceptron (MLP) module, which includes the patch-wise connection (PaC) layer and attention layer (Figure 4.6). The basic encoder blocks of the {*Layer Norm (1)*, *Multi-head Attention*, *Layer Norm (2)*, *PaC layers*} with two residual connections are stacked to construct the network architecture (Figure 4.3). In the PaC layers, information propagation is restricted to the same patch. The purpose is to render the network to calculate the relevance of patches by self-attention only. We give details on the three important manipulations/layers herein.

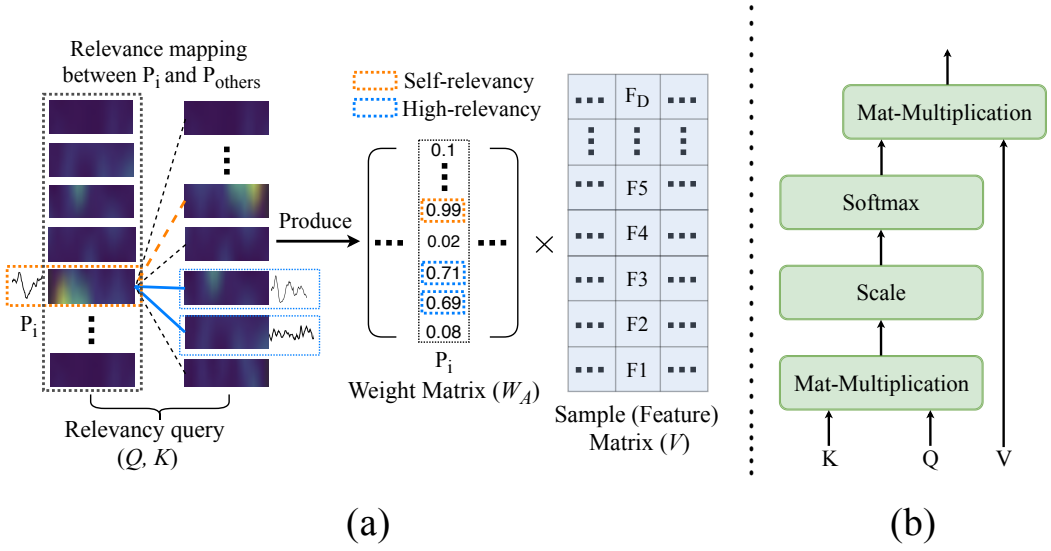


Figure 4.4: Attention mechanism: (a) shows that the attention layer first calculates relevancies among patches and then maps the relevant weight matrix to an input of each attention layer. (b) illuminates the workflow of the attention layer.

Attention mechanism: The attention mechanism associates the individual patches and maps the relevance to the ground truth stage y with three components: the query (Q), key (K), and value (V) matrices, which are the matrix of linear projections produced by the input X.

$$Q, K, V = Linear(X), X \in \mathbb{R}^{(L+1)\times D} \tag{4.3}$$

where, Q, K, $V \in \mathbb{R}^{(H\cdot W+1)\times D}$ Here, the matrix Q (in the case of using mini-batch) represents query that comprises of a query sequence with basic units (FT

patches here). In the case of self-attention, the K is the same as Q, and the attention is utilized to calculated relevance among the patches. The resultant relevance values are further used to calculate V. Ultimately, the weighted value matrix that encompasses the importance of FT patches extracted via a global reference will be fed into a further encoder or final staging module. The mathematical processing of Q, K, V can be summarized as below:

$$W_A = Softmax(\frac{QK^T}{\sqrt{d}}) \tag{4.4}$$

$$AttentionScore = W_A \cdot V \tag{4.5}$$

where layer $\sqrt{d}$ denotes a normalization-like scale that is applied to each Q-K computation, and the softmax layer is used to functionally obtain the weight vector of attention W_A for V.

Multi-head attention: Similar to the way that a CNN increases the number of filters to enrich the expressiveness of the feature space, the attention mechanism can be extended to multi-head attention to prevent losing the manifold expression of the features. At the beginning of each building block, h (the number of the heads) set of Q and K is generated and mapped by the linear projection. Then, the self-attention implements h times in parallel to calculate relevance representations, where each operation is called a "head." Eventually, a linear layer projects their concatenated outputs and summarizes the attention result. The multi-head attention is defined as follows:

$$Z_{Mhead}(Q, K, V) = Concat(head_1, head_2, \cdots, head_h)W^o, \tag{4.6}$$

where $W^o \in \mathbb{R}^{h \cdot D \times (150+1)}$ is a weight matrix. It is used for head-wise attention, while a linear projection is applied after the output of the multi-head attention for each round. Different from the conventional channel-wise output (like the CNNs), the multi-head attention can be regarded as multi-threads to implement the attention in parallel.

MLP&Staging module: As introduced above, the output of the multi-head layer is fed to the PaC layers with GELU being their activation function. Simultaneously, a residual connection skip-connects to the output of the building block to avoid the gradient vanishing.

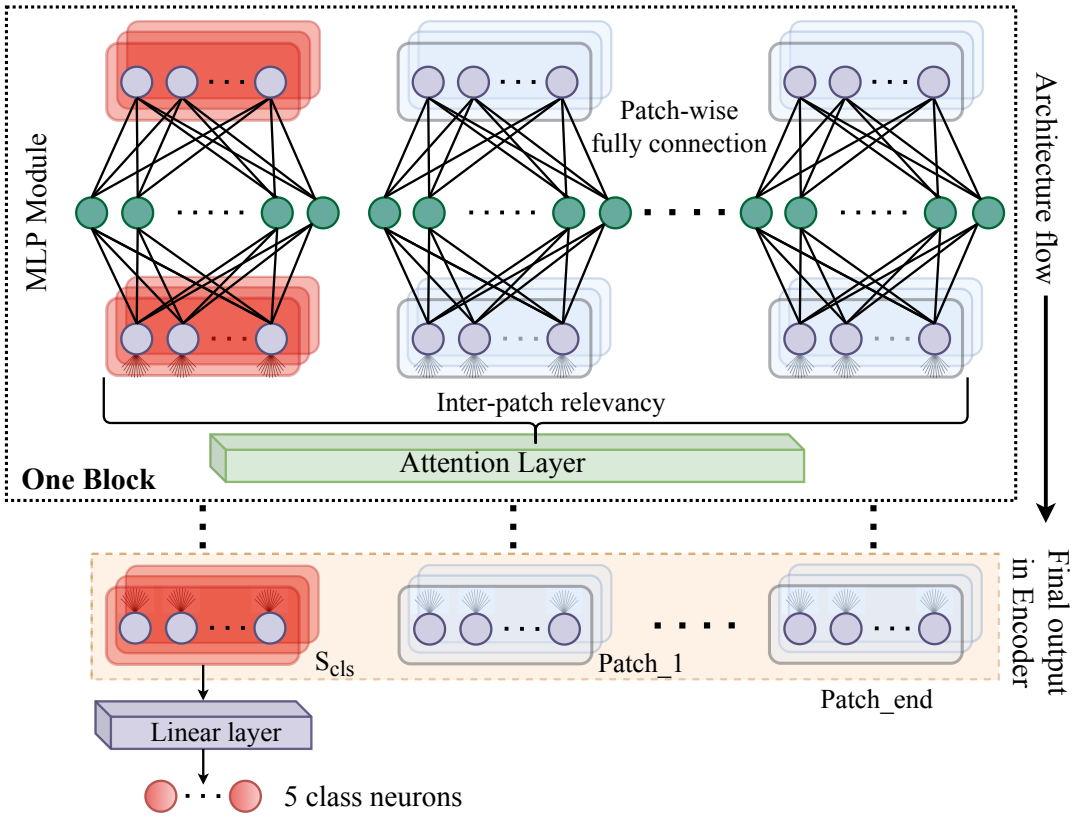


Figure 4.5: Patch-wise MLP architecture and staging module. This MLP can be view as a pre-defined network, that purpose is to reconstruct the feature space of each patch while keep the Independence in each patch.

As the information passes through all stacked blocks, the class patch S'_{Cls} has absorbed information about the relevance of FT patches extracted from the global context and is used solely in the scoring step. As shown in Figure 4.6, a linear projection finally compresses the flattened class token to neurons that have the same number of sleep stages.

$$y' = Linear(LayerNorm(S'_{Cls})), \tag{4.7}$$

where $y' \in \{\text{W, N1, N2, N3, REM}\}$, and S'_{Cls} is also normalized before the final classification layer.

4 Experiment

4.1 Training Strategy

Since a healthy dataset (mentioned in Section II) was extracted, we conducted a pre-training-to-fine-tuning strategy in the training process. Specifically, during the pre-training phase, the healthy-set, which contained balanced samples for each of the stages, was used to optimize the model parameters. Here, we utilized the *Adam* optimizer with a biggish learning rate of 10^{-3} to spur the model to converge fast and adjust the parameters along the broadly right direction [100]. In the fine-tuning phase, the remaining training data was used to further optimize the pre-trained parameters. Moreover, the *AdamW* optimizer [61] is used. The learning rate was set to 10^{-4} to meticulously optimize the cross-entropy loss function.

To alleviate the overvaluation of the performance, we implemented a subject-wise 7-fold cross-validation by splitting the data into seven subject-wise subsets. In one trial, six subsets were used in the training step, while the remaining subset (roughly 800 subjects) was used for testing.

4.2 Parameter settings

The details of the parameter settings are shown in TABLE 4.2. To make the utmost of the model, a grid search of hyperparameters was implemented in this work to seek the best combination. Note that the optimal settings (the bold

Table 4.2: Grid search of the model parameter setting in experiments and the optimal combination is in bold.

Parameter	**Value**
#Stacked encoder	{6, **8**, 12}
#Heads (h)	{2, 4, **8**, 12}
Dimension of linear projection of D	{16, **32**, 64}
Normalization-like scale ($\sqrt{d}$)	{2, **4**}
Dimension of MLP output	{64, **128**, 256}
Dropout rate	{0.2, **0.5**, 0.8}
#Training epoch	200
Batch size	32
#Parameters	1.3×10^5

values in TABLE 4.2) were used both for pre-training and the subsequent training. Additionally, a dropout layer was added after each linear projection and attention layer to further avoid the overfitting problem. The model was implemented with the Pytorch v1.4 framework, and all experiments were conducted on a server with an NVIDIA GeForce RTX 2080Ti GPU.

4.3 Baseline Networks

As introduced above, there are special treatments in the feature generation and network structure in this research. The validate their efficacy, seven baseline models with changes in the following three aspects are also constructed for comparison.

Data processing: (i) The argument for using the spectrogram was supported by repeating the process of feature generation that appeared in the preceding works of [85, 92], where the pipeline tried to learn how to generate features with raw EEG signals (see the [Inception + proposed model] pipeline in TABLE 4.4). By using the multi-scale Inception architecture [94], the pipeline is expected to find more latent features in parallel [94]. Specifically, one convolutional layer containing five parallel filters corresponding to the five frequency bands of sleep EEG was constructed. Therefore, the output of this layer can be regarded as the

multi-scale components of the original EEG signal. Details on this framework can be seen in Appendix.

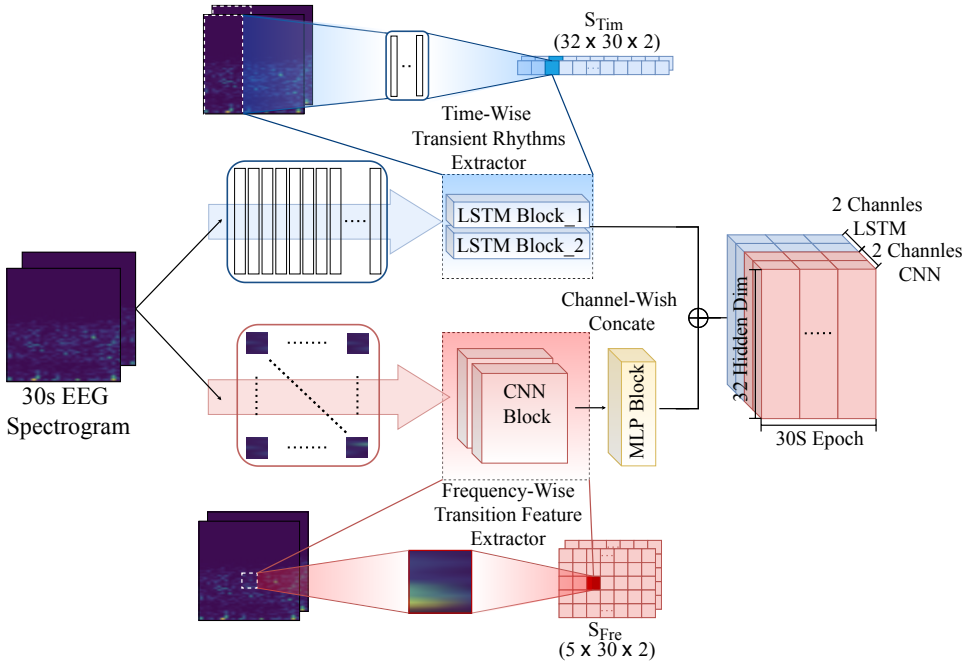


Figure 4.6: Architecture of stage-specific feature mapping network

(*ii*) To evaluate the *frequency-time patching* of the spectrogram, we prepared input that consisted of 1-sec time patches without frequency patching [13]. After, a sequence of 30 time patches of 32 dimensions was embedded to a higher dimension of 64 and fed into the network with the same settings in TABLE III (see the [Time patching+proposed model] pipeline).

Sequential model: The argument over the advantage of using attention is validated by replacing the attention mechanism with RNN networks. Four RNN models using LSTM and Bi-LSTM were used as the replacement (see the four pipelines using LSTM or Bi-LSTM in TABLE 4.4 that cover both the raw EEG input and time-patch input).

Learning strategy: Although the model was pre-trained within a balanced dataset, the imbalance issue existed in fine-tuning training. Since previous work on deep learning has tried to overcome the imbalance problem with weighted loss function, we implemented a class-wise weighted cross-entropy loss function[82]

in another baseline model while all the other aspects are the same (see the [*T patching+proposed model+weighted learning*] pipeline in TABLE 4.4).

For each fold, we counted the proportion of each class (cls) from the total sample T_{cls}. To provide more weight to the class that had fewer samples, the weight vector P_w was derived from the inverse of the class proportion. This step is defined as below.

$$P_w = \frac{\sum_{cls=1}^{N} T_{cls}}{T_{cls}} \tag{4.8}$$

To normalize the sample distributions of different folds to the same distribution, P_w was scaled by its maximum element. Afterwards, the resulting vector W_{wce} as the parameter was transferred to the following loss function.

$$W_{wce} = P_w / Max(P_w) \tag{4.9}$$

Since the maximum element divided by itself was one, W_{wce} in this implementation was registered to a range of $[0, 1]$. The final expression of the loss is described as:

$$\pounds = -W_{wce} \sum_{cls=1}^{N} ylog(y'), \tag{4.10}$$

where y was the ground truth of the input samples, while y' was the predicted label that resulted in the network referring to Eq. 4.7.

Table 4.3: Hyperparameters of the Inception and LSTM/Bi-LSTM used in the baseline models

Inception		LSTM / Bi-LSTM	
#Module	1	#Layers	{4, 6}
Dropout rate	0	Dropout rate	0
Kernel size	{12, 31, 62, 125}	Hidden size	32
Padding size	{105, 0, 0, 0}	Bias	True
Stride	{12, 31, 62, 125}	-	
#Parameters	3.5×10^5	#Parameters	1.2×10^6

4.4 Evaluation metrics

To evaluate the staging performance of each class, three metrics were used in the experiments, i.e., the stage-specific precision (Pre), recall (Re), and F1-score ($F1$). Precision is the proportion of positive prediction that was actually correct, while recall is the proportion of actual positives that were successfully predicted. F1-score reflects the overall metrics based on these two. Moreover, we adopted overall accuracy to evaluate the training effusiveness and Cohen's Kappa coefficient (k) to measure the inter-rater reliability (IRR) [17]. The definitions of precision (Pre), recall (Re), F1-score (F1), accuracy (ACC) and kappa (k) are as follows:

$$Pre_c = \frac{TP}{TP + FP} \tag{4.11}$$

$$Re_c = \frac{TP}{TP + FN} \tag{4.12}$$

$$F1_c = \frac{2 \cdot Pre_c \cdot Re_c}{Pre_c + Re_c} \tag{4.13}$$

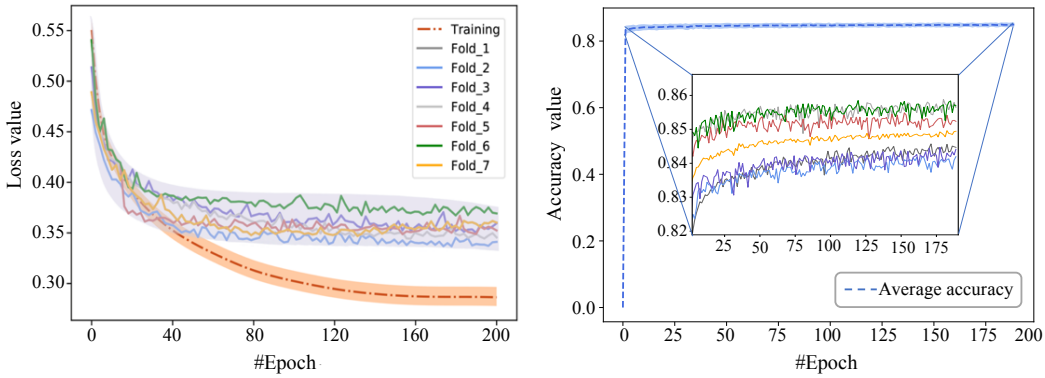
$$ACC = \frac{\sum_{c=1}^{c} TP_c}{N} \tag{4.14}$$

$$k = \frac{N_c - N_e}{N_t - N_e} \tag{4.15}$$

where TP, FP, and FN denote the true positives, false positives, and false negatives of the class (c), respectively. Meanwhile N_c denotes the number of correctly scored stages, N_t denotes the total number of stages while N_e denotes the expected number of agreements for each stage.

4.5 Result

Observation 1: *The proposed model had a small generalization gap.* Figure 4.7 (a) depicts the average training loss and the 7-fold validation loss in the fine-tuning step. Since the cost gap between the training and the validation can be viewed as a loose measure of the *generalization gap*, the small sample-wise gap



(a) Training and validation loss (b) 7-fold accuracy

Figure 4.7: Training effectiveness in the fine-tuning: (a) shows 7-fold training loss and validation loss of each fold. (b) is average accuracy in each fold validation.

(0.06 on average) and the narrow variation among folds suggests the model could be extended to new dataset without a plunge in performance. As the training loss converged gradually, the validation loss showed a similar trend without obvious fluctuation in the latter training epochs, which implies that the over-fitting of the trained model was not severe in the proposed model.

Figure 4.7 (b) shows the trends in the accuracy of the 7-fold validations. the accuracy reached a plateau in a fewer epochs than the decrease of loss and was steady (0.84 – 0.86) in the following training epochs.

Observation 2: *The proposed model is better in generating the stage-dependent features.* Looking closer into the performance of the baseline models in TABLE 4.4, two conclusions can be drawn. The first one is that the models based on the attention mechanism, including the Time patching+Attention and the Inception+Attention pipelines, outperformed the RNN based ones. Given that the best of the RNN models comes from FT Patching+Bi-LSTM, which is different from the proposed pipeline in terms of the model architecture only, it is reasonable to conclude that the proposed model is a better architecture in generating stage-dependent features.

Observation 3: *The proposed frequency-time patching is an ideal representation of sleep stages.* The conclusion is drawn by comparing the performance of the

Table 4.4: Comparison of performance among baseline pipelines and our pipelines. We make the best stage-wise performance of each evaluation metric.

Comparative Domain	Pipeline	Evaluation Metrics	Wake	N1	N2	N3	REM	Overall
Data Processing	*Time patching + proposed model*	*Pre*	0.91	0.42	0.85	0.84	0.73	0.75
		Re	0.93	0.20	0.89	0.83	0.76	0.72
		F1	0.92	0.31	0.87	0.84	0.75	0.74
	Inception + proposed model	*Pre*	0.91	0.38	0.85	0.85	0.75	0.77
		Re	0.89	0.32	0.80	0.84	0.77	0.72
		F1	0.90	0.34	0.82	0.84	0.76	0.74
Model	*Time patching + LSTM*	*Pre*	0.85	0.31	0.80	0.83	0.65	0.69
		Re	0.89	0.18	0.79	0.79	0.64	0.66
		F1	0.87	0.22	0.80	0.82	0.64	0.68
	FT patching + LSTM	*Pre*	0.86	0.32	0.82	0.83	0.66	0.70
		Re	0.88	0.22	0.80	0.80	0.65	0.67
		F1	0.87	0.27	0.81	0.82	0.65	0.69
	Time patching + Bi-LSTM	*Pre*	0.84	0.31	0.79	0.83	0.67	0.68
		Re	0.89	0.23	0.86	0.80	0.63	0.68
		F1	0.87	0.26	0.83	0.80	0.65	0.68
	FT patching + Bi-LSTM	*Pre*	0.85	0.33	0.82	0.84	0.68	0.70
		Re	0.89	0.24	0.87	0.81	0.63	0.69
		F1	0.87	0.29	0.85	0.83	0.66	0.70
Learning Strategy	*FT patching + proposed model + weighted learning*	*Pre*	0.90	0.40	0.81	0.82	0.76	0.74
		Re	0.91	0.30	0.85	0.80	0.75	0.72
		F1	0.90	0.35	0.83	0.81	0.75	0.73
Proposed Pipeline	*FT patching + proposed model*	*Pre*	**0.93**	**0.42**	**0.87**	**0.89**	**0.80**	**0.78**
		Re	**0.93**	**0.33**	**0.90**	**0.84**	**0.79**	**0.76**
		F1	**0.93**	**0.38**	**0.89**	**0.87**	**0.80**	**0.77**

model with different inputs, namely the time patching and FT patching of the spectrogram and the raw EEG signal input to the Inception module. Among the three pipelines (weighted learning excluded here), the retainment of the frequency-band information, that is the Inception+proposed model and the FT patching+proposed pipelines, showed its effectiveness in a comparison of the overall performances with the time patching+proposed pipelines. Given that the design of the Inception module serves the same purpose of retaining the resolution in the frequency domain, the combination of spectrogram and FT patching is more appropriate than the data-driven approach for feature generation. This situation may be caused by the nature of the high randomness in EEG signals and can be mitigated by transformation to the frequency domain and the following integral process each 1 Hz.

Observation 4: *The proposed method reached the new SOTA for most of the stages.* We compared the performances of the proposed method with related works that experimented with the same or different database (SHHS) in TABLE 4.5. We can observe that the classification of the wake stage had the best perfor-

Table 4.5: Performance obtained by proposed pipeline and existing works using same SHHS database.

	Method	# Experimental SHHS subjects		Wake	N1	N2	N3	REM	k
Proposed pipeline	*EEG + proposed model*	5736	*Pre*	**0.93**	0.42	**0.87**	**0.89**	0.80	0.80
			Re	**0.93**	0.33	**0.90**	0.84	0.79	
			F1	**0.93**	0.38	**0.88**	**0.87**	0.80	
Enrique Fernandez *et al.*, 2021 [29]	*EEG&EMG + Separable CNN*	5793	*Pre*	0.89	0.57	0.85	0.88	0.83	0.80
			Re	**0.93**	0.23	0.89	0.77	0.85	
			F1	0.91	0.40	0.87	0.83	0.84	
Huy Phan *et al.*, 2021 [77]	*EMG, EOG, EEG + GRU, LSTM*	5791	*Pre*	-	-	-	-	-	**0.85**
			Re	-	-	-	-	-	
			F1	0.92	0.50	0.88	0.85	0.88	
Emadeldeen Eldele *et al.*, 2021 [27]	*EEG + CNN*	329	*Pre*	0.90	0.30	0.87	0.87	0.80	0.81
			Re	0.83	0.36	0.86	**0.87**	0.83	
			F1	0.86	0.33	0.87	0.87	0.82	
Shreyasi Pathak *et al.*, 2021 [73]	*EEG, EOG, EMG + CNN, bi-LSTM*	5793	*Pre*	0.92	0.31	0.83	0.84	**0.88**	0.79
			Re	0.92	0.50	0.84	0.67	**0.89**	
			F1	0.92	0.40	0.84	0.76	**0.89**	
Hogeon Seo *et al.*, 2020 [83]	*EEG + RCNN*	5791	*Pre*	0.92	0.42	0.85	0.85	0.87	0.80
			Re	0.88	0.47	**0.90**	0.86	0.86	
			F1	0.90	0.45	0.87	0.85	0.86	
Niranjan Sridhar *et al.*, 2020 [87]	*ECG + CNN*	561	*Pre*	0.86		0.74	0.68	0.76	0.66
			Re	0.80		0.82	0.49	0.81	
			F1	0.82		0.78	0.57	0.78	
Qiao Li *et al.*, 2018 [56]	*ECG + CNN*	5793	*Pre*	0.85		0.62	0.59	0.60	-
			Re	0.80		0.74	0.54	0.73	
			F1	0.82		0.67	0.56	0.65	
Siddharth Biswal *et al.*, 2018 [9]	*EEG, Resp, EMG + RCNN*	15804	*Pre*	0.90	**0.69**	0.84	0.80	0.79	0.73
			Re	0.81	**0.67**	0.78	0.76	0.74	
			F1	0.85	**0.68**	0.81	0.78	0.76	
Foroozan Karimazadeh *et al.*, 2018 [46]	*EEG + KNN*	140	*Pre*	0.89	0.55	0.75	0.84	0.86	-
			Re	0.81	0.58	0.68	0.54	0.78	
			F1	0.85	0.56	0.71	0.66	0.81	

mance when compared with the other works. Meanwhile, the proposed pipeline also outperformed the other works in terms of stages N2 and N3, with the F1 score being 0.88 and 0.87, respectively. Pathak's paper [73], reach the highest performance for the REM stage by fusing the EOG and EMG signals, which are clinically important for REM stage. As we will mention in the Discussion, although the fusing of EMG gets the new best performance of REM using our pipeline, this paper focus on the EEG-only situation. In a similar way, Biswal *et al.* [9] gets the best performance for N1 by fusing the EEG, waveforms of the chest belt (respiration) and EMG and capturing the intra- and inter-epoch context using the Bidirectional LSTM. We will further discuss the results in in TABLE 4.5 in Discussion.

(a) EEG only (b) EEG + EOG

Figure 4.8: (a) Confusion matrix of the proposed pipeline with EEG as the input. (b) Confusion matrix of the proposed pipeline with EEG and EOG as the input.

5 Summary and Discussions

As introduced above, by proposing a new feature processing framework for EEG signal called FT patching and associating the FT patches through the multiplications of the *Query*, *Key*, and *Value* matrices, the attention scores were generated as alternatives of the conventional features. The proposed model attained the best performance with a lighter network architecture compared with the baseline

models (see *#parameters* in TABLE 4.2 and 4.3).

One of the main purpose of this work is to push the epoch-wise automatic stage scoring algorithm with EEG signal to a new level. From the comparison with the preceding researches, it is reasonable to conclude that this purpose is fulfilled. Of note, the N1 stage stand in the midway of a dynamic process from conscious to a real sleep stage. According to the definition of N1, comparison with the preceding epoch, i.e., decrease of *Alpha* band component, is required. However, such kind of information is inevitably lacking in epoch-wise classification. For this reason, in the research that also takes in the inter-epoch relation [83], significant improvement on the Recall can be seen.

Besides, it is well acknowledged that EOG signal is indispensable in identifying the REM stage, and the results of the Pathak *et al* [73] can be regarded as experimental evidence. By taking the EOG signal as part of the input, the identification of REM reach the best performance. We have tried to include the EOG in our pipeline, and the results get the highest record for the REM stage with 0.91 precision, 0.90 Recall and 0.91 F1 score. The corresponding confusion matrix has shown in Figure 4.8 (b). In the main body of this work the results are discussed under the restriction of EEG signal in pursuing a potential extension to home use, where the simplicity on the sensor attachment is advantageous.

5 | Interprebility in Model Decision

1 introduction

Despite all this progress, we have not yet seen automatic sleep staging widely adopted clinically yet. Unofficial communications with leading sleep experts point to the scepticism of deep learning models being a black box, which is a common criticism when it comes to the application of artificial intelligence in healthcare and medicine [78]. We argue that two overarching obstacles need to be addressed for a machine scoring system to work alongside practitioners in an interactive and collaborative manner: (1) interpretability, and (2) uncertainty quantification. Interpretability is the ability of a model to explain how its decision is made given a certain input, to be understood by a human. Inspired by the way a sleep expert performs manual scoring, interpretability in automatic sleep scoring is reasonably about (but not limited to) what features the model learns from the input signal, whether these features are relevant to and underpin the sleep stages, and how the decision on a target epoch is made under the influence of its neighboring epochs. Interpretability is particularly important due to the fact that sleep stages are ambiguous and even different human experts tend to disagree at a certain extend. Also, due to this ambiguity, quantifying uncertainty in the model's decisions is equally important. Simply put, we are in need of a simple and concrete metric, ideally a single number, for quantifying the model's uncertainty. Using this metric, epochs that are scored with low confidence by the model can be deferred to sleep experts for further inspection.

1.1 Chapter Organization

The remainder of this chapter is structured as follows: Section 2 describes model interpretability methods and tools in computer vision and sequential model (Transformer). Section 3 introduces the method of our visualization. Section 4 shows the results of model classification decision. Section 5 further discusses the model performance by visualizing the decision-making.

2 Preliminary

Explainability in computer vision: Many methods were suggested for generating a heatmap that indicates local relevancy, given an input image and a CNN. Most of these methods belong to one of two classes: gradient methods and attribution methods. Gradient based methods are based on the gradients with respect to the input of each layer, as computed through back-propagation. The gradient is often multiplied by the input activations, which was first done in the gradient-input method. Integrated Gradients [91] also compute the multiplication of the inputs with their derivatives. However, this computation is done on the average gradient and a linear interpolation of the input. SmoothGrad, visualizes the mean gradients of the input, and performs smoothing by adding to the input image a random Gaussian noise at each iteration. The FullGrad method offers a more complete modeling of the gradient by also considering the gradient with respect to the bias term, and not just with respect to the input. We observe that these methods are all classagnostic: at least in practice, similar outputs are obtained regardless of the class used to compute the gradient that is being propagated.

Explainability for Transformers: There are not many contributions that explore the field of visualization for Transformers and, as mentioned, many contributions employ the attention scores themselves. This practice ignores most of the attention components, as well as the parts of the networks that perform other types of computation. A self-attention head involves the computation of queries, keys, and values. Reducing it only to the obtained attention scores (inner products of queries and keys) is myopic. Other layers are not even considered. Our method, in contrast, propagates through all layers from the decision back to the

input. LRP was applied for Transformers based on the premise that considering mean attention heads is not optimal due to different relevance of the attention heads in each layer [11]. However, this was done in a limiting way, in which no relevance scores were propagated back to the input, thus providing partial information on the relevance of each head. We note that the relevancy scores were not directly evaluated, only used for visualization of the relative importance and for pruning less relevant attention heads.

3 Method

3.1 Attention Visualization

The multi-head attention relies heavily on the multiplication operation in the attention calculation, and the relevance scores of the resultant attention matrices might play different roles in the network. Unlike the conventional gradient-based visualization [91], we use an attention-oriented visualization similar to the works of Chefer et al. [11] to highlight the FT patches that the model is attending to by inferring both the gradient and the relevance from the final classification decision for each attention layer.

Hence, the output of the visualization is ideally reconstructed as a spectrogram-like attention graph ($\hat{V}$). That is, the size of the attention graph is maintained with the 1-channel processed spectrogram S' and is defined as:

$$\hat{V} = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot \ldots \cdot \bar{A}^{(B)}, \tag{5.1}$$

where $\hat{V} \in \mathbb{R}^{F' \times T}$ consists of a set of sub-graphs of B encoder modules. Since each row of W_A in Eq. 4.4 is normalized to the attention coefficients of each embedded patch with respect to the others, W_A can be treated as an attention map. Each sub-graph $\bar{A}^{(b)}$ in encoder b has a gradient of the attention map $\bigtriangledown W_A^{(b)}$ and its relevance diffusion $R^{(n_b)}$, which can be formulated in:

$$\bar{A}^{(b)} = I + Mean_h(\bigtriangledown W_A^{(b)} \odot R^{(n_b)}), \tag{5.2}$$

where $\odot$ is the Hadamard product. With to the multi-head mechanism, a mean operation is applied across the h dimension. In addition, the identity matrix (I) is used to avoid the self-inhibition of each patch [11].

The process of relevance propagation starts from the class y of the staging module and iteratively diffuses to the each layer $L^{(n)}$, where $n \in (1, \cdots N)$ is the layer index for the whole Transformer, here, the staging module, i.e., a linear projection in Eq. 4.7 is defined as $L^{(1)}$. Suppose $L^{(n}(X^{(n)}, Y^{(n)})$ describes the layer n function to the corresponding input $X^{(n)}$ and the weights $W^{(n)}$; the relevance propagation is similar to the chain rule and follows the generic Deep Taylor Decomposition [66]:

$$R_j^{(n)} = \sum_i X_j^{(n)} \frac{\partial L_i^{(n)}(X^{(n)}, W^{(n)})}{\partial X_j} \frac{R_i^{(n-1)}}{L_i^{(n)}(X^{(n)}, W^{(n)})}, \tag{5.3}$$

where the subscript j denotes the elements in $R^{(n)}$. Because $L^{(n)}$ corresponds to the first layer of the network, the index i represents the elements in $R^{(n-1)}$. Moreover, this relevance propagation will stop at the first layer of the block for each round b.

3.2 Uncertainty Quantification

To reveal the decision process of the model for different channels, each sample generates two attention graphs. Unit resolution of the attention graph corresponds to 1 FT patch and manifests the intensity it is attended to throughout the pipeline.

Meanwhile, an entropy-based statistical analysis is utilized for each two attention graphs to quantify the causality between the attention visualization and the model decision. Considering the transient attribute of stage-dependent features, the attention intensities in one frequency band distributed homogeneously might contain more stage-dependent information. Otherwise, the band within continuous lower intensity or highlighting should lead to a lower sample entropy value.

4 Result

Attention map of the proposed FT patching is more sensitive and perceivable. To discuss the interpretability of the pipeline proposed in Section 3.1, in Figure 5.1, we attempted to visualize the attention scores of two different input

sequences, which were FT patches and time patches. For simplicity and clarity, we normalized the intensity of each reconstructed spectrogram-like attention map. Gradient-based visualization (GbV) using Grad-CAM [column (e) in Figure 5.1] was also generated for a direct comparison.

Spectrogram & FT patches: While most of the bright spots (higher power of FT patches) in the spectrograms can find their correspondences in the attention maps of FT patches, the latter contain more clues. For instance, regarding the wake stage, the bright patch around 2 sec of the spectrogram (C4-A1 channel) of the *Alpha* band finds its correspondence in the FT patch map (see the stride in the ellipse), and the bright patches of the *Alpha* band can also be found around 22 sec in C3-A2. The re-organization of the FT patches can also be seen in stage N1, where the last patches of the *Theta* band were closely attended to (ellipse of C4-A1). The fusion of the two spectrograms can also be seen around 2–5 sec, where C4-A1 is given more attention than the counterpart in C3-A2 (see the two red boxes of N1). When the stage turned to deep sleep (N2, and N3), it can be seen that the patches in the *Delta* band gradually became patches in the spot light. For the N2 stage, most of the K-complex shown in both the EEG signal and the spectrogram was closely attended to. Besides, a spindle-alike patch (see the red box in C3-A2 channel) was given high attention as well. Regarding the N3 stage, while the spectrograms were very similar to those of N2, the attention maps of the FT patch features themselves were generally bright patches of the *Delta* band.

Attention map of FT patches & GbV: The GbV suggests different FT patches [bright spots in Figure 5.1 (e)] that contributed to the identification of the sleep stages. However, some of the highlighted FT patches were not authentic. The solely bright stride in the *Delta* band of C4-A1 for the wake stage and the bright strides in the *Delta* band of the two channels of the N1 stage were not concretely supported by clinical findings. For direct visualization, it is reasonable to conclude that an attention map of FT patches results in a distinct view of the input that is more informative than a network without the attention mechanism.

Attention map of FT patches & time patches: The homogeneity of the time patching in the frequency domain drove the attention to inevitably focus on the time domain solely. Referring to the corresponding spectrogram, the

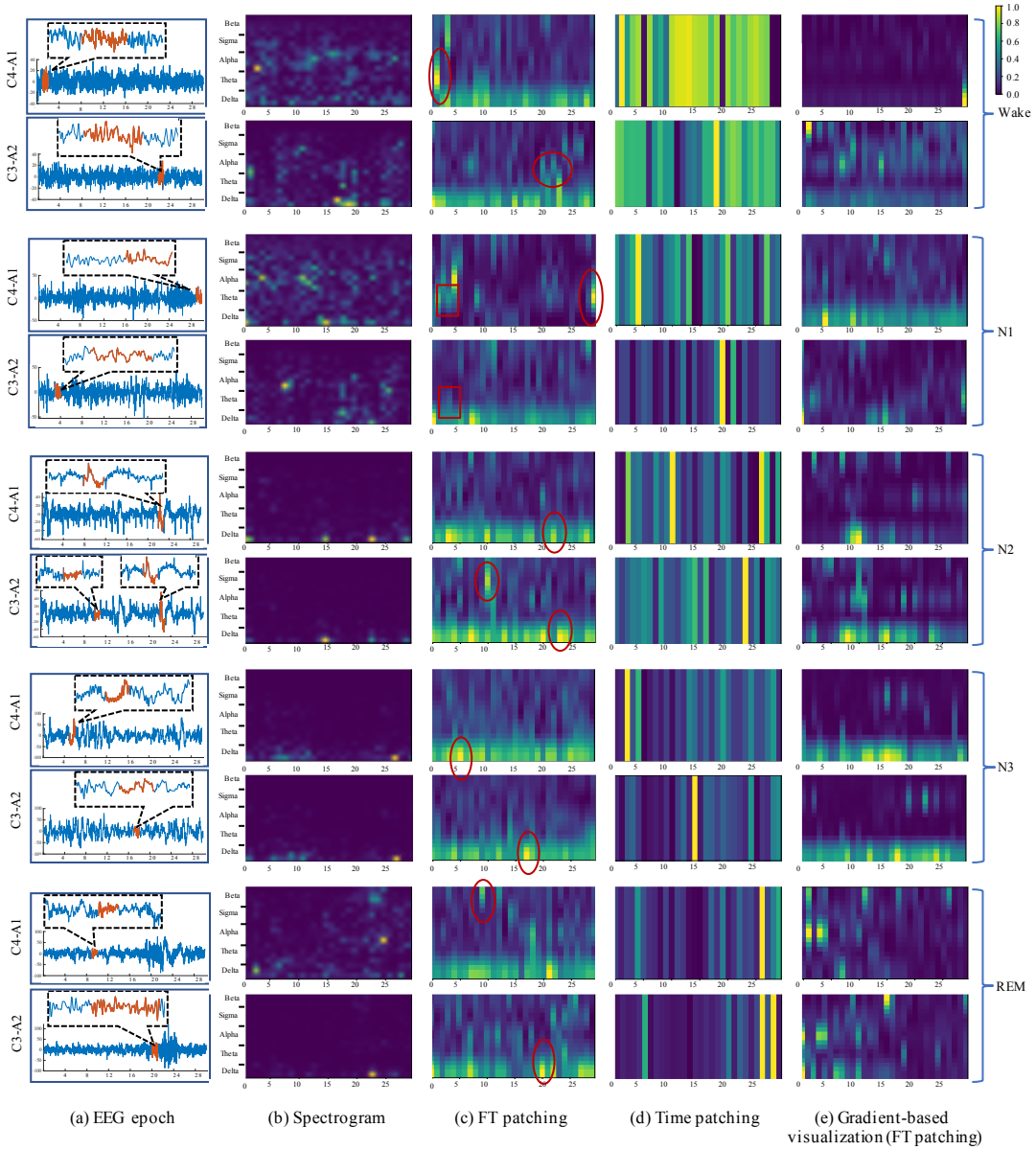


Figure 5.1: Visualization of different pipelines.

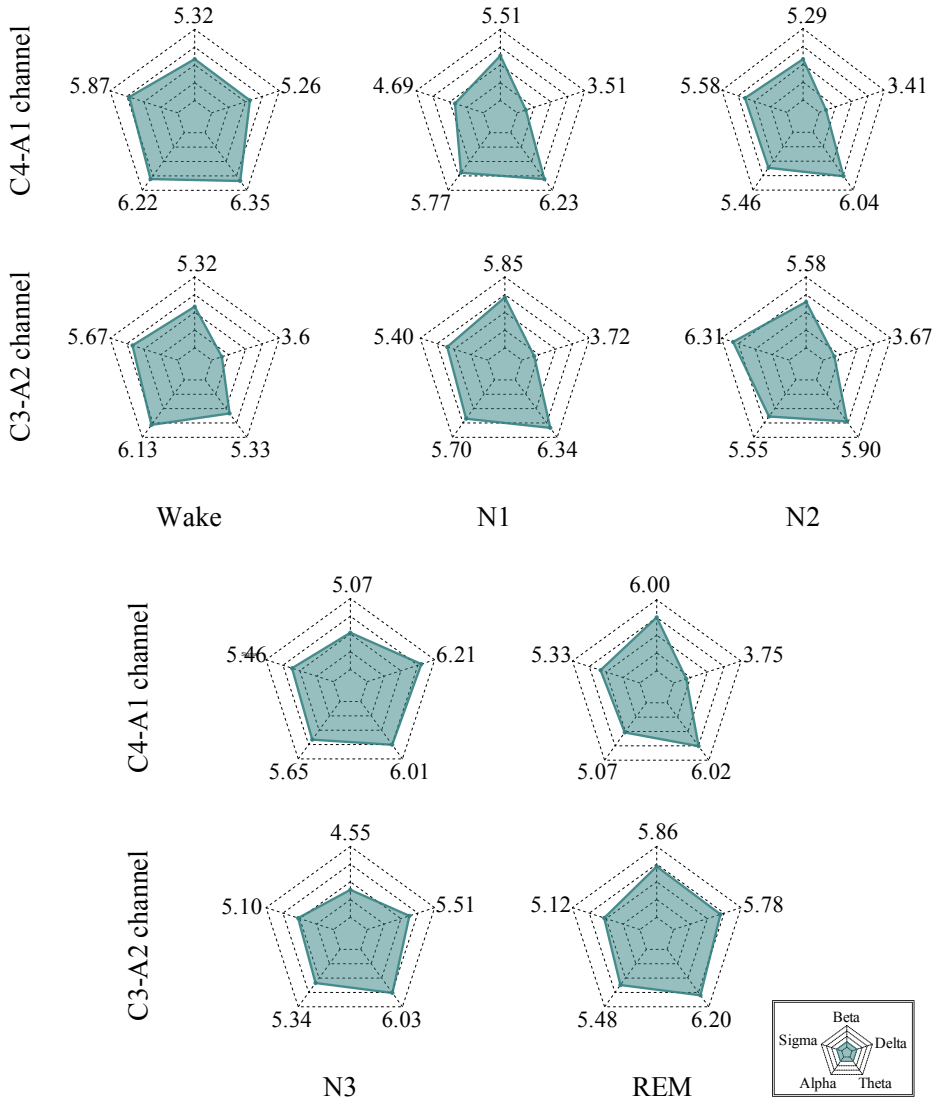


Figure 5.2: Visualization of the attention derived values of each frequency band of the proposed pipeline.

bright time patches shown in Figure 5.1 (d) find themselves unobtrusive against a generally bright background, e.g., the 2-sec time patch of C4-A1 of the wake, or without concrete clinical support, e.g., the bright time patches of the two channels of the N3 stage. Some other samples can be found in Appendix.

Another view of the attention maps of the FT patches can be seen in radar graphs (Figure 5.2), which sums up the sample entropy of the attention intensities for the FT patches in the five frequency bands. In stage wake sample, the entropy of *Alpha* band reaches relatively higher values, that is, 6.21 and 5.51 in C4-A1 and C3-A2. From wake to N1, the dominant *Alpha* band attenuated in N1 accompanying the increase in the *Theta* band. As the sleep went deeper, the attention given to the *Alpha* band turned stable gradually at a relatively low state, while the *Delta* band came into the foreground. Although the *Theta* dominates the REM case in appearance, the *Beta* frequency band indicates certain quantities of information. That result meets the sleeping truth that the brain becomes active again and starts to dream in REM.

5 Summary and Discussions

The augmented class patches that were trained in the model absorbs the pairwise relevance of patches and conclude the patches each class needs to attend to. As mentioned in Section 4 direct connection between the clinical standard and the attention map can be seen in our pipeline. We consider this exposure an important enhancement of the interpretability of an autonomic scoring algorithm and will facilitate clinical/physiological discussion.

By visually comparing the manual annotations of the sleep stages with the autonomic scoring of our pipeline (Figure 5.3), it can be seen that the misclassification tends to occur in between {N2, N3} pair and {Wake, REM} pair (The counting results have shown in TABLE 5.1). Moreover, misclassification often occurs when the sleep stages transition frequently in a relatively short interval (see the red dots in the middle hypnogram of Figure 5.3). This situation may be caused by the incompleteness of sleep relevant information of the EEG signal compared with polysomnography used in manual annotation. In contrast, our model can recognize the transitions of stage with relatively low frequency accurately.

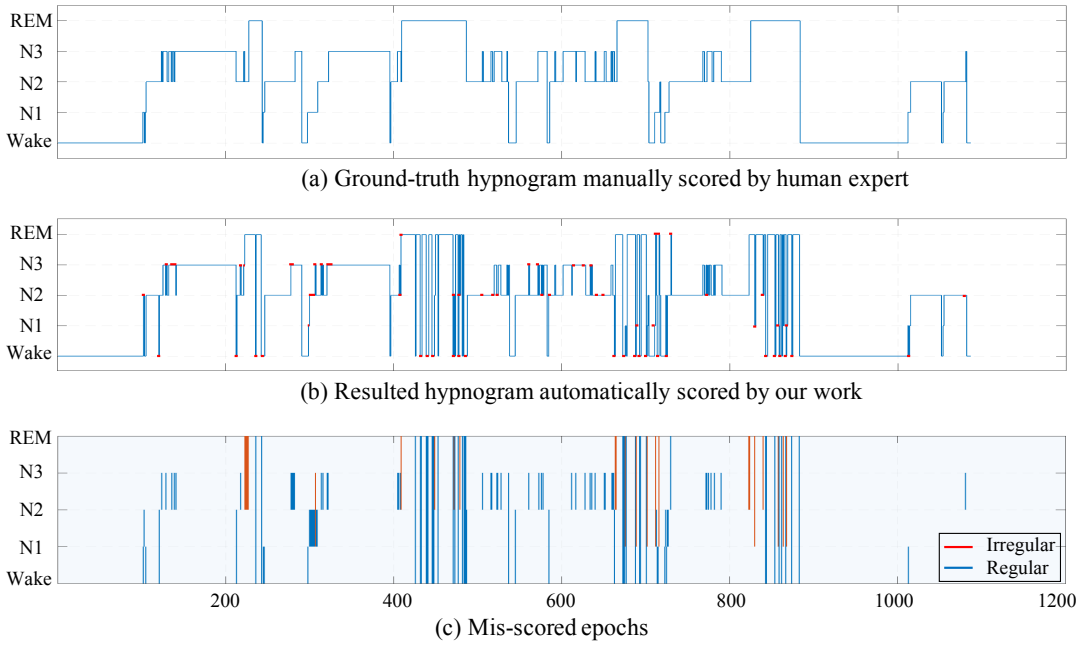


Figure 5.3: Examples of hypnogram manually scored by human expert (a) and hypnogram automatically scored by our method (b) for one subject from SHHS dataset. Misclassification is marked in red. The sticks in the bottom figure (c) mark the wrong labels. Blue sticks represent the regular sleep stage transitions that can not be detected; while the red sticks represent the falsely detected irregular transitions.

Table 5.1: Different types of misclassifications and their counting results. Each pair illuminates the two direction inter-epoch transitions, e.g., {Wake, N1}: $Wake \rightarrow N1$ and $N1 \rightarrow Wake$.

Regular pairs	#Pair	Irregular pairs	#Pair
{Wake, N1}	12	{N1, REM}	9
{Wake, N2}	6	{N2, REM}	18
{Wake, N3}	1	-	-
{N1, N2}	14	-	-
{N2, N3}	58	-	-
{Wake, REM}	35	-	-

Furthermore, irregular misclassification pairs that the inter-epoch transitions violate the regular sleep cyclic pattern can be seen, occupying about 18% of the total misclassification. For instance, our pipeline may output assignments of {N1, REM} ($N1 \rightarrow REM$ or $REM \rightarrow N1$), a sharp change of stage that skips N2 and N3 stages. Noteworthily, for the irregular pairs of {N2, REM}, N2 sometimes changes to REM without the deep sleep phase may happen occasionally (around 200 in Figure 5.3). However once the body becomes stable at REM stage, this kind of change seldom happen. Given the issue mentioned above, introducing of the constraint on inter-epoch transition is considerable in future works.

6 | Conclusion

Sleep is a crucial physiological function of humans. Lacking adequate sleep might leads to the risk of many sleep-related disorders, such as sleep apnea syndrome, schizophrenia, depression, and other neural abnormalities. The gold standard of sleep construction is re-defined as five different stages according to the American Academy of Sleep Medicine. Screening the sleep stages incorporating electroencephalogram is the major tool in the assessment of the sleep quality and diagnosis of sleep-related disorders, such as sleep apnea syndrome, depression, schizophrenia, insomnia, narcolepsy, and other neural abnormalities. Recent advances in portable monitoring technology have increased access to sleep screening, yet the gold-standard in-lab multi-lead EEG capturing from the overnight polysomnography still require manual scoring by sleep experts. This laborious manual process is a major obstacle for advancing our understanding about sleep and more importantly, for deploying sleep-scientific findings into neuroscientific and pathological problems.

To spur the use of automatic sleep stage scoring alternative in the real clinical setting, this thesis explores the representation learning of sleep stage-dependent characteristics and proposes a context-sensitive flexible pipeline that automatically adapts attributes of data itself. Meanwhile, we visualize the stage scoring process of the model decision with the Layer-wise Relevance Propagation method, which shows that the proposed pipeline is more sensitive perceivable in the decision-making process than the baseline pipelines.

1 Contributions

The outcomes drawn from this thesis could be beneficial for both medical practitioners and researchers. Below, the contribution along with suggestions are summarized for each part as follows:

- The proposed model had a small generalization gap. As the training loss converged gradually, the validation loss showed a similar trend without obvious fluctuation in the latter training epochs, which implies that the over-fitting of the trained model was not severe in the proposed model.

- The proposed model is better in generating the stage-dependent features. Given that the comparative models comes from popular sequential model (LSTM/Bi-LSTM), which is different from the proposed pipeline in terms of the model architecture only, it is reasonable to conclude that the proposed model is a better architecture in generating stage-dependent features.

- Results indicate that the proposed frequency-time patching is an ideal representation of sleep stages. Given that the design of the Inception module serves the same purpose of retaining the resolution in the frequency domain, the combination of spectrogram and frequency-time patching is more appropriate than the data-driven approach for feature generation. This situation may be caused by the nature of the high randomness in EEG signals and can be mitigated by transformation to the frequency domain and the following integral process each 1 Hz.

- Attention map of the proposed frequency-time patching is more sensitive and perceivable. Results indicate that while most of the higher power of frequency-time patches in the spectrograms can find their correspondences in the attention maps of frequency-time patches, the latter contain more clues.

- The gradient-based visualization suggests different frequency-time patches that contributed to the identification of the sleep stages. For direct visualization, it is reasonable to conclude that an attention map of patches results

in a distinct view of the input that is more informative than a network without the attention mechanism.

- Without using the EOG signal, although the *Theta* dominates the REM case in appearance, the *Beta* frequency band indicates certain quantities of information. That result meets the sleeping truth that the brain becomes active again and starts to dream in REM..

2 Opportunities for Future Work

We believe that this thesis makes a major contribution to improving the automatic sleep stage scoring alternative. However, there are many open challenges for future work. Below, we outline a list of potential opportunities.

One of the main purpose of this paper is to push the epoch-wise automatic stage scoring algorithm with EEG signal to a new level. From the comparison with the preceding researches, it is reasonable to conclude that this purpose is fulfilled. Of note, the N1 stage stand in the midway of a dynamic process from conscious to a real sleep stage. According to the definition of stage N1, comparison with the preceding epoch, i.e., decrease of *Alpha* band component, is required. However, such kind of information is inevitably lacking in epoch-wise classification. For this reason, in the research that also takes in the inter-epoch relation, significant improvement on the Recall can be seen.

It is well acknowledged that EOG signal is indispensable in identifying the REM stage. By taking the EOG signal as part of the input, the identification of REM reach the best performance. We have tried to include the EOG in our pipeline, and the results get the highest performance for the REM stage. In the main body of this paper the results are discussed under the restriction of EEG signal in pursuing a potential extension to home use, where the simplicity on the sensor attachment is advantageous.

In spite of the difficulties mentioned above, the pipeline still get the best performance for the Wake and deep sleep stages with the least information in terms of information sources (EEG only versus senors fusion) and abundance (epoch-wise versus inter-epoch).

The augmented class patches that were trained in the model absorbs the pairwise relevance of patches and conclude the patches each class needs to attend to. We consider this exposure an important enhancement of the interpretability of an autonomic scoring algorithm and will facilitate clinical/physiological discussion.

References

[1] Rechtschaffen A and Kales A. A manual of standardized terminology, techniques, and scoring system for sleep stages of human subjects. Public Health Service, NIH Publication No. 204.Washington, DC: US Government Printing Office, 1968.

[2] Antoine Adamantidis, Carolina Gutierrez Herrera, and Thomas Gent. Oscillating circuitries in the sleeping brain. Nature Reviews Neuroscience, 20, 10 2019. doi: 10.1038/s41583-019-0223-4.

[3] D Aeschbach, T.T Postolache, L Sher, J.R Matthews, M.A Jackson, and T.A Wehr. Evidence from the waking electroencephalogram that short sleepers live under higher homeostatic sleep pressure than long sleepers. Neuroscience, 102(3):493–502, 2001. ISSN 0306-4522. doi: https://doi.org/10.1016/S0306-4522(00)00518-2. URL https://www.sciencedirect.com/science/article/pii/S0306452200005182.

[4] Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakhan, Busarakum Chaitusaney, Nattapong Jaimchariyatam, Ekapol Chuangsuwanich, Wei Chen, Huy Phan, Nat Dilokthanakul, and Theerawit Wilaiprasitporn. Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning(provide datasets). IEEE Journal of Biomedical and Health Informatics, PP, 11 2020. doi: 10.1109/JBHI.2020.3037693.

[5] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35:1798–1828, 08 2013. doi: 10.1109/TPAMI.2013.50.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A

review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[7] Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Carole L. Marcus, and Bradley V. Vaughn. The AASM Manual for the Scoring of Sleep and Associated Events. American Academy of Sleep Medicine, 53(9): 1689–1699, 2013. ISSN 1098-6596.

[8] Arnab Bhattacharjee, Suvasish Saha, Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad. Sleep apnea detection based on rician modeling of feature variation in multiband eeg signal. IEEE Journal of Biomedical and Health Informatics, 23(3):1066–1074, 2019. doi: 10.1109/JBHI.2018.2845303.

[9] Siddharth Biswal, Haoqi Sun, Balaji Goparaju, M Brandon Westover, J. Sun, and Matt Bianchi. Expert-level sleep scoring with deep neural networks. Journal of the American Medical Informatics Association : JAMIA, 25, 11 2018. doi: 10.1093/jamia/ocy131.

[10] Margaret F. Carr, M. Karlsson, and L. Frank. Transient slow gamma synchrony underlies hippocampal memory replay. Neuron, 75:700–713, 2012.

[11] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. arXiv preprint arXiv:2012.09838, 2020.

[12] Z. Chen, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang. ivae: An improved deep learning structure for eeg signal characterization and reconstruction. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1909–1913, Seoul, Korea (South), dec 2020. IEEE Computer Society. doi: 10.1109/BIBM49941.2020.9313107.

[13] Z. Chen, K. Odani, P. Gao, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang. Feasibility analysis of transformer model for eeg-based sleep scoring. In 2021 IEEE International Conference on Biomedical and Health Informatics (BHI'21), Virtual, July 2021.

[14] Zheng Chen, Naoaki Ono, Wei Chen, Toshiyo Tamura, MD Altaf-Ul-Amin, Shigehiko Kanaya, and Ming Huang. The feasibility of predicting impending malignant ventricular arrhythmias by using nonlinear features of short heartbeat intervals.

Computer Methods and Programs in Biomedicine, 205:106102, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106102.

[15] Zhenghua Chen, Min Wu, Wei Cui, Chengyu Liu, and Xiaoli Li. An attention based cnn-lstm approach for sleep-wake detection with heterogeneous sensors. IEEE Journal of Biomedical and Health Informatics, pages 1–1, 2020. doi: 10.1109/JBHI.2020.3006145.

[16] Julie Anja Engelhard Christensen, Rick Wassing, Yishul Wei, Jennifer R. Ramautar, Oti Lakbila-Kamal, Poul Jørgen Jennum, and Eus J. W. Van Someren. Data-driven analysis of eeg reveals concomitant superficial sleep during deep sleep in insomnia disorder. Frontiers in Neuroscience, 13:598, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00598.

[17] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

[18] Vincenzo Crunelli, Magor Lorincz, William Connelly, François David, Stuart Hughes, Regis Lambert, Nathalie Leresche, and Adam Errington. Dual function of thalamic low-vigilance state oscillations: Rhythm-regulation and plasticity. Nature Reviews Neuroscience, 01 2018. doi: 10.1038/nrn.2017.151.

[19] Mengxi Dai, Dezhi Zheng, Rui Na, Shuai Wang, and Shuailei Zhang. Eeg classification of motor imagery using a novel deep learning framework. Sensors, 19:551, 01 2019. doi: 10.3390/s19030551.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[21] Mohammed Diykh, Yan Li, and Peng Wen. Eeg sleep stages classification based on time domain features and structural graph similarity. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24:1–1, 11 2016. doi: 10.1109/TNSRE.2016.2552539.

[22] Mohammed Diykh, Yan Li, and Shahab Abdulla. Eeg sleep stages identification based on weighted undirected complex networks. Computer Methods and Programs in Biomedicine, 184:105116, 10 2019. doi: 10.1016/j.cmpb.2019.105116.

[23] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. Mixed neural network approach for temporal sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(2):324–333, 2018. doi: 10.1109/TNSRE.2017.2733220.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

[25] M. Dresler, V.I. Spoormaker, P. Beitinger, M. Czisch, M. Kimura, A. Steiger, and F. Holsboer. Neuroscience-driven discovery and development of sleep therapeutics. Pharmacology & Therapeutics, 141(3):300–334, 2014. ISSN 0163-7258. doi: https://doi.org/10.1016/j.pharmthera.2013.10.012.

[26] Alexander Ecker, Philipp Berens, Georgios Keliris, Matthias Bethge, Nikos Logothetis, and Andreas Tolias. Decorrelated neuronal firing in cortical microcircuits. Science (New York, N.Y.), 327:584–7, 01 2010. doi: 10.1126/science.1179867.

[27] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 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, 2021. doi: 10.1109/TNSRE.2021.3076234.

[28] Enrique Fernandez-Blanco, Daniel Rivero, and Alejandro Pazos. Eeg signal processing with separable convolutional neural network for automatic scoring of sleeping stage. Neurocomputing, 410:220–228, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.05.085. URL https://www.sciencedirect.com/science/article/pii/S092523122030936X.

[29] Enrique Fernandez-Blanco, Carlos Fernandez-Lozano, Alejandro Pazos, and Daniel Rivero. Ensemble of convolution neural networks on heterogeneous signals for sleep stage scoring. CoRR, abs/2107.11045, 2021.

[30] Raffaele Ferri, Mauro Manconi, Giuseppe Plazzi, Oliviero Bruni, Stefano Vandi, Pasquale Montagna, Luigi Ferini-Strambi, and Marco Zucconi. A quantitative statistical analysis of the submentalis muscle emg amplitude during sleep in normal

controls and patients with rem sleep behavior disorder. Journal of sleep research, 17:89–100, 04 2008. doi: 10.1111/j.1365-2869.2008.00631.x.

[31] Stuart Fogel and Carlyle Smith. The function of the sleep spindle: A physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. Neuroscience and biobehavioral reviews, 35:1154–65, 12 2010. doi: 10.1016/j.neubiorev.2010.12.003.

[32] Glass L Goldberger A, Amaral L and Stanley H.E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation, 23(101):215–220, 2000.

[33] Luis Herrera, Antonio Mora, Carlos Fernandes, Daria Migotina, Alberto Guillén, and Agostinho Rosa. Symbolic representation of the eeg for sleep stage classification. International Conference on Intelligent Systems Design and Applications, ISDA, pages 253–258, 11 2011. doi: 10.1109/ISDA.2011.6121664.

[34] J Hobson and Edward Pace-Schott. The cognitive neuroscience of sleep: neuronal systems, consciousness and learning. Nature reviews. Neuroscience, 3:679–93, 10 2002. doi: 10.1038/nrn915.

[35] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8):2011–2023, 2020. doi: 10.1109/TPAMI.2019.2913372.

[36] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition, 83:134–149, 2018.

[37] Moody GB Ichimaru Y. Development of the polysomnographic database on cd-rom. Psychiatry and Clinical Neurosciences, 53(4):175–177, 1999.

[38] Anas Imtiaz and Esther Rodriguez-Villegas. A low computational cost algorithm for rem sleep detection using single channel eeg. Annals of biomedical engineering, 42, 08 2014. doi: 10.1007/s10439-014-1085-6.

[39] Pankaj Jadhav, Gaurav Rajguru, Debabrata Datta, and Siddhartha Mukhopadhyay. Automatic sleep stage classification using time–frequency images of cwt and transfer learning using convolution neural network. Biocybernetics and

Biomedical Engineering, 40(1):494–504, 2020. ISSN 0208-5216. doi: https://doi.org/10.1016/j.bbe.2020.01.010.

[40] Ziyu Jia, Xiyang Cai, Gaoxing Zheng, Jing Wang, and Youfang Lin. Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. IEEE Transactions on Artificial Intelligence, 1(3):248–257, 2020. doi: 10.1109/TAI.2021.3060350.

[41] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 1324–1330. International Joint Conferences on Artificial Intelligence Organization, 7 2020. URL https://doi.org/10.24963/ijcai.2020/184. Main track.

[42] Dihong JIANG, Ya-nan LU, Yu MA, and Yuanyuan WANG. Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement. Expert Systems with Applications, 121, 12 2018. doi: 10.1016/j.eswa.2018.12.023.

[43] Dihong JIANG, Ya nan LU, Yu MA, and Yuanyuan WANG. Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement. Expert Systems with Applications, 121:188–203, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.12.023. URL https://www.sciencedirect.com/science/article/pii/S0957417418307917.

[44] William Joiner. Unraveling the evolutionary determinants of sleep. Current Biology, 26, 10 2016. doi: 10.1016/j.cub.2016.08.068.

[45] Alexander Kaplan, Alexander & Andrew Fingelkurts, Sergey Borisov, and Boris Darkhovsky. Nonstationary nature of the brain activity as revealed by eeg/meg: Methodological, practical and conceptual challenges. Signal Processing, 85:2190–2212, 11 2005. doi: 10.1016/j.sigpro.2005.07.010.

[46] Foroozan Karimzadeh, Reza Boostani, Esmaeil Seraj, and Reza Sameni. 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, 2018. doi: 10.1109/TNSRE.2017.2775058.

[47] Diederik Kingma and Max Welling. Auto-encoding variational bayes. 12 2014.

[48] Dani Kiyasseh, Girmaw Abebe, Nhan Le Nguyen Thanh, Tan Van, Louise Thwaites, Tingting Zhu, and David Clifton. Plethaugment: Gan-based ppg augmentation for medical diagnosis in low-resource settings. IEEE Journal of Biomedical and Health Informatics, PP:1–1, 04 2020. doi: 10.1109/JBHI.2020.2979608.

[49] Henriette Koch, Julie Christensen, Rune Frandsen, Marielle Zoetmulder, Lars Arvastson, Soren Christensen, Poul Jennum, and Helge Sorensen. Automatic sleep classification using a data-driven topic model reveals latent sleep states. Journal of neuroscience methods, 235, 07 2014. doi: 10.1016/j.jneumeth.2014.07.002.

[50] James Krueger, Marcos Frank, Jonathan Wisor, and Sandip Roy. Sleep function: Toward elucidating an enigma. Sleep Medicine Reviews, 28, 08 2015. doi: 10.1016/j.smrv.2015.08.005.

[51] Andrew Krystal, Jack Edinger, William Wohlgemuth, and Gail Marsh. Nrem sleep eeg frequency spectral correlates of sleep complaints in primary insomnia subtypes. SLEEP, 25:626–636, 09 2002. doi: 10.1093/sleep/25.6.626.

[52] Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 641–652, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1054.

[53] Martin Langkvist and Amy Loutfi. A deep learning approach with an attention mechanism for automatic sleep stage classification. CoRR, abs/1805.05036, 2018. URL http://arxiv.org/abs/1805.05036.

[54] Xiang Li, Zhigang Zhao, Dawei Song, Yazhou Zhang, Chunyang Niu, Junwei Zhang, Jidong Huo, and Jing Li. Variational autoencoder based latent factor decoding of multichannel eeg for emotion recognition. pages 684–687, 11 2019. doi: 10.1109/BIBM47256.2019.8983341.

[55] Xiaojin Li, Licong Cui, Shiqiang Tao, Jing Chen, Xiang Zhang, and Guo-Qiang Zhang. Hyclasss: A hybrid classifier for automatic sleep stage scoring. IEEE Journal of Biomedical and Health Informatics, PP:1–1, 02 2017. doi: 10.1109/JBHI.2017.2668993.

[56] Liu C Li Q, Nemati S Shashikumar SP, and Clifford GD. Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. Physiol Meas, 2018.

[57] Wootaek Lim, Sangwon Suh, and Youngho Jeong. Weakly labeled semi-supervised sound event detection using crnn with inception module. 11 2018.

[58] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. pages 2–11, 01 2003. doi: 10.1145/882082.882086.

[59] Gi-Ren Liu, Yu-Lun Lo, Yuan-Chung Sheu, and Hau-Tieng Wu. Diffuse to fuse eeg spectra – intrinsic geometry of sleep dynamics for classification. Biomedical Signal Processing and Control, 55, 02 2018. doi: 10.1016/j.bspc.2019.101576.

[60] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. CoRR, abs/2103.14030, 2021. URL https://arxiv.org/abs/2103.14030.

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

[62] Yan Ma, Wenbin Shi, Chung-Kang Peng, and Albert C Yang. Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches. Sleep Medicine Reviews, 37, 01 2017. doi: 10.1016/j.smrv.2017.01.003.

[63] Alexander Malafeev, Dmitry Laptev, Stefan Bauer, Ximena Omlin, Aleksandra Wierzbicka, Adam Wichniak, Wojciech Jernajczyk, Robert Riener, Joachim Buhmann, and Peter Achermann. Automatic human sleep stage scoring using deep neural networks. Frontiers in Neuroscience, 12:781, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00781.

[64] Lisa Marshall, Halla Helgadóttir, Matthias Mölle, and Jan Born. Boosting slow oscillations during sleep potentiates memory. Nature, 444:610–3, 12 2006. doi: 10.1038/nature05278.

[65] Jodi Mindell and Ariel Williamson. Benefits of a bedtime routine in young children: Sleep, development, and beyond. Sleep Medicine Reviews, 40, 11 2017. doi: 10.1016/j.smrv.2017.10.007.

[66] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65:211–222, 05 2017. doi: 10.1016/j.patcog.2016.11.008.

[67] Samaneh Nasiri and Gari D. Clifford. Attentive adversarial network for large-scale sleep staging. In Proceedings of the 5th Machine Learning for Healthcare Conference, volume 126 of Proceedings of Machine Learning Research, pages 457–478, Virtual, 07–08 Aug 2020. PMLR.

[68] Hong-Viet Ngo, Thomas Martinetz, Jan Born, and Matthias Mölle. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. Neuron, 78, 04 2013. doi: 10.1016/j.neuron.2013.03.006.

[69] Buckner R.L Nishida M, Pearsall J and Walker M.P. Rem sleep, prefrontal theta, and the consolidation of human emotional memory. Cerebral Cortex, 19(5):1158–1166, 2009.

[70] Shigeru Nonoue, Midori Mashita, Shingo Haraki, Akira Mikami, Hiroyoshi Adachi, Hirofumi Yatani, Atsushi Yoshida, Masako Taniike, and Takafumi Kato. Inter-scorer reliability of sleep assessment using eeg and eog recording system in comparison to polysomnography. Sleep and Biological Rhythms, 15:39–48, 09 2016. doi: 10.1007/s41105-016-0078-2.

[71] Edward Pace-Schott and J Hobson. Pace-schott ef, hobson ja. the neurobiology of sleep: genetics, cellular physiology and subcortical networks. nat rev neurosci 3: 591-605. Nature reviews. Neuroscience, 3:591–605, 09 2002. doi: 10.1038/nrn895.

[72] Amiya Patanaik, Ju Lynn Ong, Joshua Gooley, Sonia Ancoli-Israel, and Michael Chee. An end-to-end framework for real-time automatic sleep stage classification. Sleep, 41, 03 2018. doi: 10.1093/sleep/zsy041.

[73] Shreyasi Pathak, Changqing Lu, Sunil Belur Nagaraj, Michel van Putten, and Christin Seifert. Stqs: Interpretable multi-modal spatial-temporal-sequential model for automatic sleep scoring. Artificial Intelligence in Medicine, 114:102038, 2021. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2021.102038.

[74] Musa Peker. A new approach for automatic sleep scoring: Combining taguchi based complex-valued neural network and complex wavelet transform. Computer Methods and Programs in Biomedicine, 129:203–216, 2016. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2016.01.001. URL https://www.sciencedirect.com/science/article/pii/S016926071600002X.

[75] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Chén, and Maarten de Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27:400 – 410, 01 2019. doi: 10.1109/TNSRE.2019.2896659.

[76] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. Joint classification and prediction cnn framework for automatic sleep stage classification. IEEE Transactions on Biomedical Engineering, 66(5):1285–1296, 2019. doi: 10.1109/TBME.2018.2872652.

[77] Huy Phan, Oliver Y. Chen, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3070057.

[78] Huy Phan, Kaare B. Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. CoRR, abs/2105.11043, 2021. URL https://arxiv.org/abs/2105.11043.

[79] A P Prathosh, Varun Srivastava, and Mayank Mishra. Adversarial approximate inference for speech to electroglottograph conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, PP:1–1, 09 2019. doi: 10.1109/TASLP.2019.2942140.

[80] Wei Qu, Zhiyong Wang, Hong Hong, Zheru Chi, David Dagan Feng, Ron Grunstein, and Christopher Gordon. A residual based attention model for eeg based

sleep staging. IEEE Journal of Biomedical and Health Informatics, 24(10):2833–2843, 2020. doi: 10.1109/JBHI.2020.2978004.

[81] Borja Rodríguez Gálvez, Ragnar Thobaben, and Mikael Skoglund. The convex information bottleneck lagrangian. Entropy, 22(1), 2020. ISSN 1099-4300. doi: 10.3390/e22010098. URL https://www.mdpi.com/1099-4300/22/1/98.

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[83] Hogeon Seo, Seunghyeok Back, Seongju Lee, Deokhwan Park, Tae Kim, and Kyoobin Lee. 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, 2020. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2020.102037.

[84] John Shambroom, Stephan Fábregas, and Jack Johnstone. Validation of an automated wireless system to monitor sleep in healthy adults. Journal of sleep research, 21:221–30, 08 2011. doi: 10.1111/j.1365-2869.2011.00944.x.

[85] Isuru Niroshana S.M., Xin Zhu, Ying Chen, and Wenxi Chen. 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, 2019. doi: 10.1109/ICAwST.2019.8923359.

[86] Arnaud Sors, Stphane Bonnet, Sebastien Mirek, Laurent Vercueil, and Jean-Franois Payen. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomedical Signal Processing and Control, April 2018. doi: 10.1016/j.bspc.2017.12.001.

[87] Niranjan Sridhar, Philip Stephens Ali Shoeb, David Ben Shimol Alaa Kharbouch, and Joshua Burkart. Deep learning for automated sleep staging using instantaneous heart rate. npj Digital Medicine, (1), 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0291-x.

[88] Robert Stickgold, J.A. Hobson, Roar Fosse, and Magdalena Fosse. Sleep, learning, and dreams: Off-line memory reprocessing. Science (New York, N.Y.), 294:1052–7, 12 2001. doi: 10.1126/science.1063530.

[89] Chenglu Sun, Chen Chen, Wei Li, Jiahao Fan, and Wei Chen. A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning. IEEE Journal of Biomedical and Health Informatics, PP:1–1, 08 2019. doi: 10.1109/JBHI.2019.2937558.

[90] Youqiang Sun, Jiuyong Li, Jixue Liu, Bingyu Sun, and Christopher Chow. An improvement of symbolic aggregate approximation distance measure for time series. Neurocomputing, 138:189–198, 08 2014. doi: 10.1016/j.neucom.2014.01.045.

[91] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 3319–3328. JMLR.org, 2017.

[92] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Transactions on Neural Systems and Rehabilitation Engineering, PP, 03 2017. doi: 10.1109/TNSRE.2017.2721116.

[93] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. pages 1–9, 06 2015. doi: 10.1109/CVPR.2015.7298594.

[94] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

[95] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI Conference on Artificial Intelligence, 02 2016.

[96] Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, Robert Thomas, and Arthur Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep. Sleep Medicine, 2(6):537–553, 2001. ISSN 1389-9457. doi: https://doi.org/10.1016/S1389-9457(01)00149-6.

[97] Orestis Tsinalis, Paul Matthews, and Yike Guo. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Annals of biomedical engineering, 44, 10 2015. doi: 10.1007/s10439-015-1444-y.

[98] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? CoRR, abs/2105.07197, 2021. URL https://arxiv.org/abs/2105.07197.

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[100] Sinong Wang, Madian Khabsa, and Hao Ma. To pretrain or not to pretrain: Examining the benefits of pretrainng on resource rich tasks. pages 2209–2213, 01 2020. doi: 10.18653/v1/2020.acl-main.200.

[101] Wei Wu, Zhe Chen, Xiaorong Gao, and Emery Brown. A hierarchical bayesian approach for learning sparse spatio-temporal decomposition of multichannel eeg. NeuroImage, 56:1929–45, 03 2011. doi: 10.1016/j.neuroimage.2011.03.032.

[102] Ziliang Xu, Xuejuan Yang, Jinbo Sun, Peng Liu, and Wei Qin. Sleep stage classification using time-frequency spectra from consecutive multi-time points. Frontiers in Neuroscience, 14, 01 2020. doi: 10.3389/fnins.2020.00014.

[103] Özal yıldırım, Ru San Tan, and U Rajendra Acharya. An efficient compression of ecg signals using deep convolutional autoencoders. Cognitive Systems Research, 52:198–211, 07 2018. doi: 10.1016/j.cogsys.2018.07.004.

[104] Junming Zhang, Ruxian Yao, Wengeng Ge, and Jinfeng Gao. Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel eeg. Computer Methods and Programs in Biomedicine, 183:105089, 09 2019. doi: 10.1016/j.cmpb.2019.105089.