

Sleep Stage Classification and Obstructive Sleep Apnea Detection Using Deep Learning

SENEVIRATHNA MUDIYANSELAGE ISURU NIROSHANA

A DISSERTATION

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY


IN COMPUTER SCIENCE AND ENGINEERING

Graduate Department of Computer and Information Systems
The University of Aizu
2021



© Copyright by Senevirathna Mudiyansele Isuru Niroshana
All Rights Reserved.

The thesis titled

 Automatic Sleep Stage Classification and Obstructive Sleep
Apnea Detection Using Deep Learning

by

Senevirathna Mudiyanseelage Isuru Niroshana


is reviewed and approved by:

Chief referee


Senior Associate Professor
ZHU Xin

Zhu Xin 2021/8/13 


Professor
CHEN Wenxi

Wenxi Chen 2021/8/12 

Senior Associate Professor
TRUONG Cong Thang

Thang 2021/8/12 

Associate Professor
OKUYAMA Yuichi

Yuichi 2021/08/12 

The University of Aizu

2021

DEDICATION

THIS THESIS IS DEDICATED TO MY PARENTS AND FIANCEE MR. S.M. CHANDRASEKARA,
MRS. E.A. RAMYA DHAMAYANTHI, AND DR. S.K. MANAWADU

THANK YOU FOR YOUR LOVE AND SUPPORT

AND

TO ALL SRI LANKAN CITIZENS FOR PROVIDING ME FREE EDUCATION AND HEALTH

THANK YOU FOR YOUR ENDLESS SACRIFICES

Contents

LIST OF FIGURES	vi
LIST OF TABLES	vii
LIST OF ABBREVIATIONS	viii
ABSTRACT	1
1 INTRODUCTION	3
1.1 Background and Motivation	3
1.1.1 Polysomnography	5
1.1.2 Sleep stages	6
1.1.3 Sleep stage scoring	10
1.1.4 Hypnogram and sleep cycles	10
1.2 Dissertation outline	12
2 LITERATURE REVIEW AND BACKGROUND	13
2.1 Hand engineered feature extraction methods	14
2.1.1 Feature extraction and classification	14
2.1.2 Support Vector Machine	15
2.1.3 K-Nearest Neighbor	16
2.1.4 Random Forest	16
2.1.5 Linear Discriminant Analysis & Nearest Centers	16
2.1.6 Artificial Neural Network (ANN)	17
2.2 Automatic feature extraction methods involved with sleep stage classification	18
2.2.1 A brief summary of automatic sleep stage approaches	19
2.3 Obstructive sleep apnea syndrome	19
2.3.1 Sleep apnea	19
2.3.2 OSAS and apnea detection)	26
3 AUTOMATIC DETECTION OF SLEEP STAGE USING PSG SIGNALS	33
3.1 Dataset and pre-processing	33
3.2 Evaluation metrics for multi-class scenario	35
3.3 Experiment I (CNN model)	37
3.3.1 Training the network	37
3.3.2 Results	38
3.3.3 Model visualization	38

3.4	Experiment II	39
3.4.1	Feature Learning	39
3.4.2	Sequence Learning	41
3.4.3	Complete Model	41
3.4.4	Multi-step training	42
	Multi-channel models	42
	Single-channel models	43
3.4.5	Implementation	43
3.4.6	Results	43
	Multi-channel models	44
	Single-channel models	45
3.4.7	Overall Performances of Proposed Models	45
3.5	Experiment III	61
3.5.1	Feature learning	62
3.5.2	Sequence learning	64
3.5.3	Complete model	65
3.5.4	Multi-phase training	67
3.5.5	Experimental setup	69
3.5.6	Results	69
	Base model: Experiment I	70
	Experiment II and III	73
3.6	Results comparison of Experiment I, Experiment II, and Experiment III	75
4	OBSTRUCTIVE SLEEP APNEA SYNDROME DETECTION BASED ON FUSED TIME-FREQUENCY SPECTRAL IMAGES	77
4.1	Materials and Methods	77
4.1.1	Dataset	78
4.1.2	Method	79
4.1.3	Preprocessing and image creation	83
	Signal noise and baseline drift	83
	Image creation	85
4.1.4	Proposed model	89
4.1.5	Implementation of model training	90
4.1.6	Evaluation metrics for binary class scenario	93
4.2	Results	94
4.2.1	Robustness evaluation	97
4.2.2	Comparison with existing methods	98
5	DISCUSSION	103
5.1	Sleep stage classification	103
5.2	OSAS detection	104
6	CONCLUSION AND FUTURE WORKS	105
6.1	Main contributions	105
6.2	Sleep stage classification	107
6.3	OSAS detection	108

ACKNOWLEDGMENTS

110

REFERENCES

124

List of Figures

1.1	Overview of sleep	4
1.2	10-20 Electrode Placement System for EEG data collection	5
1.3	Alpha waves [1]	7
1.4	k-complexes and sleep spindles [2]	8
1.5	Hypnogram and sleep cycles of a healthy person	11
1.6	Hypnogram of a healthy person	11
2.1	Automatic sleep stage classification	14
2.2	ANN architecture, feed-forward network, ‘W’ represent the weight matrices, and ‘x’ and ‘y’ represent inputs and outputs respectively	18
2.3	Sleep apnea syndrome	23
2.4	Prevalence of obstructive sleep apnea	24
2.5	Apnea–hypopnea index (AHI)	25
3.1	Epoch with artifacts due to the moment of the electrodes	47
3.2	Normal epoch (a) original signal (b) decimated signal	48
3.3	Pre-processing and rearranging data	49
3.4	Training and test data	49
3.5	Architecture of proposed method I	50
3.6	Evaluation matrices (a) Normalized Confutation Matrix (b) Per-Class Metric	50
3.7	Receiver operating curve (ROC) for proposed Method I	51
3.8	Reshaped input corresponds to a deep sleep epoch	51
3.9	Feature map corresponds to the max pool layer (L6) of left branch	52
3.10	Feature map corresponds to the max pool layer (L9) of right branch	52
3.11	Feature map corresponds to the convolutional layer (L7) of left branch	53
3.12	Proposed Multi-channel model, f_s is the frequency	54
3.13	Feature extraction section	55
3.14	Input of 1-D time series array in (1).	55
3.15	Visualizations of convolution layers of trained multi-channel model corresponding to a deep sleep epoch (a). activations of the first 5 filters (layer-3, branch-4) (b). activations of the first 8 filters (layer-4, branch-3) (c). activations of the first 8 filters (layer-4, branch-4) (d). activations of the first 8 filters (7 th layer)	56
3.16	Normalized confusion matrix for multi-channel models. (a). <i>mod_D1</i> (model for healthy subjects); (b). <i>mod_D2</i> (model for patients).	57
3.17	Receiver operating characteristic curves for multi-channel models (a). <i>mod_D1</i> (model for healthy subjects) (b). <i>mod_D2</i> (model for patients)	57

3.18	Normalized confusion matrix for single-channel models for healthy data set (a). EEG-O1 model(b). EEG-O2 model (c). EEG-C4 model (d). EEG-C3 model.	58
3.19	Normalized confusion matrix for single-channel models for Patient data set (a). EEG-C4 model(b). EEG-O2 model (c). EEG-O1 model (d). EEG-C3 model.	59
3.20	Architecture of CNN-GRU model.	66
3.21	Number of epochs in each sleep stage.(a). training epochs; (b)test epochs	67
3.22	Accuracy comparison of 4-stage and 5-stage cases for each subject.	69
3.23	per-class interquartile range (IQR) plots for precision, recall, and F1 score for 5-stage and 4-stage classifications obtained via the proposed method for all subjects, (center line: median; box limits: upper and lower quartiles; whiskers: $1.5 \times \text{IQR}$; \times : mean) (a). Precision values (b). Recall (c). F1-score.	71
3.24	Receiver operating characteristic (ROC curve) for \mathcal{M}_{base} (a) 5-stage classifica- tion (b). 4-stage classification.	71
3.25	Hypnogram for the 5-stage case (a). manually scored by a sleep expert(b). au- tomatically scored by the proposed multi-channel model.	72
3.26	Hypnogram for the 4-stage case (a). manually scored by a sleep expert (b). automatically scored by the proposed multi-channel model.	72
3.27	Confusion matrix obtained for test subjects (a). 5-stage classification (b). 4- stage classification.	73
3.28	Overall accuracy and Cohen’s kappa coefficient comparison for all experiments.	74
3.29	Comparison of evaluation metrics for all models for two classification scenar- ios (a). Multi-channel models (b). Single-channel models.	75
3.30	Result comparison Experiment I and Experiment III (5-stage case)	76
3.31	Result comparison Experiment II and Experiment III (4-stage case)	76
3.32	Accuracy comparison Experiment II and Experiment III (4-stage case, single)	76
4.1	AHI and apnea episodes distribution in the dataset	80
4.2	Additional information about the subjects in the PhysioNet Apnea dataset	81
4.3	Proposed 2D-CNN network for OSA event detection. “ $\text{Conv}(k,s,f)$ ” denotes a convolutional layer where k , s , and f are the kernel size, stride size, and number of filters, respectively. “ $\text{Max}(p,s)$ ” denotes a max-pooling layer where p and s are the pool size and stride size, respectively. The values for the filter sizes “ f ” in the four residual blocks are 32, 64, 96, and 128.	84
4.4	Preprocessing ECG segments. (a) Part of an original ECG segment. (b) The denoised and amplitude scaled version.	86
4.5	Image dataset creation	87
4.6	One-minute ECG segments transformed into (128, 128, 3) RGB images. (a) Scalogram image of a normal ECG segment. (b) Spectrogram image of the normal segment. (c) Scalogram image of an apnea ECG segment. (d) Spectro- gram image of the apnea segment.	88
4.7	Fusing the scalogram and spectrogram for an apnea ECG segment. (a) Gray- scaled scalogram and spectrogram images. (b) RGB components of the mod- ified image, where W is the gray scaled values of scalogram, and S is the gray scaled values of spectrogram (c) Fused image. (d) Fused image of a normal ECG segment. (e) Fused image of an apnea ECG segment.	89

4.8	Schematic diagram of the training procedure for the proposed 2D-CNN model with 10-fold cross validation.	91
4.9	Distributions of validation accuracy for TFR images and fused images over 10 folds.	94
4.10	Confusion matrices for per-segment apnea detection, with classwise PR and RE shown in the bottom and right-hand boxes, respectively: (a) Wigner–Ville distribution images, (b) Scalogram images, (c) Spectrogram images, and (d) Fused images.	95
4.11	IQR plots of PR, RE, and F1 for apnea detection obtained across all folds. The center line indicates the median, the box limits indicate the upper and lower quartiles, the whiskers indicate $1.5 \times \text{IQR}$, and \times indicates the mean. The images are Wigner–Ville distribution images (wg), scalogram images (sc), spectrogram images (sp), or fused images (fu).	95
4.12	Accuracy-loss graph of the proposed CNN (for the lowest-performing model).	96
4.13	Overall 10-fold cross-validation results for per-segment apnea detection with Wigner–Ville distribution, scalogram, spectrogram, and fused images. Black lines indicate the corresponding 95 % confidence interval.	97
4.14	Results comparison with existing methods	102

List of Tables

2.1	Research works that used EEG signals to perform sleep stage classification . . .	20
2.2	Research works that used ECG signals to score sleep stages	21
2.3	Research works that used combination of PSG signals to score sleep stage . .	22
2.4	Comparison of automated systems proposed for OSA detection using CinC- 2000 PhysioNet database [3]	28
3.1	Parameters of raw EEG signal	34
3.2	per-class evaluation metrics for multi-channel models	45
3.3	per-class evaluation metrics for single-channel models trained for healthy data set	46
3.4	per-class evaluation metrics for single-channel models trained for patient data set	46
3.5	Overall performances of proposed models	46
3.6	Result comparison	60
3.7	Model specifications	64
3.8	per-class evaluation metrics for proposed classification models	74
4.1	Overall performance in per-segment apnea detection TFR images and fused images	96
4.2	Performance comparison of proposed and previous methods for per-segment apnea detection for the same data set	100

List of Abbreviations

ANN	Artificial Neural Network
ASMR	Autonomous Sensory Meridian Response
AASM	American Academic Sleep Medicine
AUC	Area Under the Curve (ROC)
CNN	Convolution Neural Network
CWT	Continuous Wavelet Transform
DFT	Discrete Fourier Transform
ECG/EKG	Electrocardiogram
EDF	European Data Format
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FFT	Fast Fourier Transform
FPR	False Positive Rate
kNN	k-Nearest Neighbors
LC	Laterality Coefficient
LDA	Linear Discriminant Analysis
MA	Moving Average
ML	Machine Learning
NREM	Non-rapid Eye Movement
OSA	Obstructive Sleep Apnoea
OSAS	Obstructive Sleep Apnoea Syndrome
PCA	Principal Component Analysis
PSG	Polysomnography
REM	Rapid Eye Movement
RNN	Recurrent Neural Network
ROC	Receiver Operating Curve
SVM	Support Vector Machine
SWS	Slow wave sleep
TFR	Time Frequency Representation
TPR	True Positive Rate
WVD	Wigner–Ville Distribution

Sleep Stage Classification and Obstructive Sleep Apnea Detection Using Deep Learning

ABSTRACT

Sleep is a one of a crucial physiological process of human body, which regulates the cellular and molecular mechanisms. Sleep helps in the rejuvenation of the body and is essential for regulating both mental and physical health. Since the sleep directly regulate the physiological functions, the quality of life is strictly associate with the quality of sleep. Recently, many studies have revealed that the sleep culture is significantly changed worldwide in past few decades. These changes strongly corelated with public health consequences as well as social and economic break downs. Therefore, it is very important to improve the diagnosis of sleep disorders through sleep studies to optimize treatments for sleep disorders and to improve the quality of sleep. One of the critical steps of sleep medicine is to identify the sleep stages. The most established method is visual sleep stage scoring using polysomnography. The traditional sleep scoring method can be a tedious and time-consuming process since it needs lot of human interventions. Because of that reason, there is a huge need for improved computer based automatic sleep stage detection method. Especially, there is a huge demand for a system that can assist the sleep technicians to perform the scoring efficiently and effectively. This research project consists of two main contributions. In the first main contribution, there are three experiments focused on automatic sleep stage detection based on overnight polysomnography (PSG) data. The final part is dedicated for detecting Obstructive Sleep Apnea Syndrome (OSAS). In sleep stage detection part, experiment(I) brings a deep learning model to classify sleep stages based on 4 electroencephalogram (EEG) electrodes and two electrooculogram (EOG) for 5-stage sleep classification. In experiment (II), a combination of convolution blocks and a recurrent neural network, is used to score 4-stage sleep stages automatically. Experiment (III) is dedicated to improving the performance of the proposed model in Experiment(I) for 5-stage sleep scoring.

In the second main contribution, an Obstructive Sleep Apnea (OSA) detection method is proposed. Generally, an interruption of airflow resulting from an obstruction in the upper airways is recognized as an Obstructive Apnea event. Overly shallow breathing lasting more than 10 seconds also considers as an OSA episode. Uncontrolled Sleep Apnea causes diabetes, strokes, heart attacks and even a shortened lifespan. It is very important to detect and treat sleep apnea at the early stages to avoid long-term consequences of health especially for obese and older persons.

Like the sleep stage classification case, manual method is slower and needs rigorous intervention of sleep experts. Therefore, a novel method for obstructive sleep apnea detection based on fused time-frequency representation is presented in the second main contribution. A combination of spectral images formed with Continuous Wavelet Transform (CWT) and Short Time Fourier Transform (STFT) is used to classify the OSAS events from a single lead ECG signal. For classifying the apneic and non-apneic events, a residual neural network was designed and implemented.

1

Introduction

1.1 Background and Motivation

Sleep is an unconditional necessity for humans and plays a crucial role in maintaining the stability of biological processes. Lack of sleep may lead to drowsiness, severe fatigue, loss of day-time performance, disturbance in circadian rhythm, impairments of mental activity, low performance of the immune system, reduced cognitive functioning, and other disruptions on biological function causing long-time health risks [4] (see Fig. 1.1). On the other hand, long-term sleep disruptions cause even more health consequences such as increased risk of hypertension, diabetes, obesity or heart attack etc. Various studies show that sleep disorders or abnormalities generally have a strong correlation with depression, diabetes, metabolic syndrome, sudden

death, heart failure and other cardiovascular diseases [5][6]. The average sleep duration of an average person has decreased over the passed century due to the busy lifestyle of today's society. Nowadays, many people tend to miss their sleep due to various reasons like shift work, work more hours, using Internet, and watching TV. Furthermore, around 20% road accidents and injuries are associated with sleepiness of the driver, and most of these cases are reported the in the early morning hours [7]. Sleep loss and sleep related disorders may significantly affect for the economy resulting medical costs, hospital services, sleep diagnostic equipment and sleep medicine. Sleep medicine related awareness among the general is reported low, even though many individuals having sleep problems. It should be noted that the impact of sleep on health and life quality is not recognized as a significant fact in the general public. However, sleep diagnostics is performed for some patients to evaluate the quality of sleep.

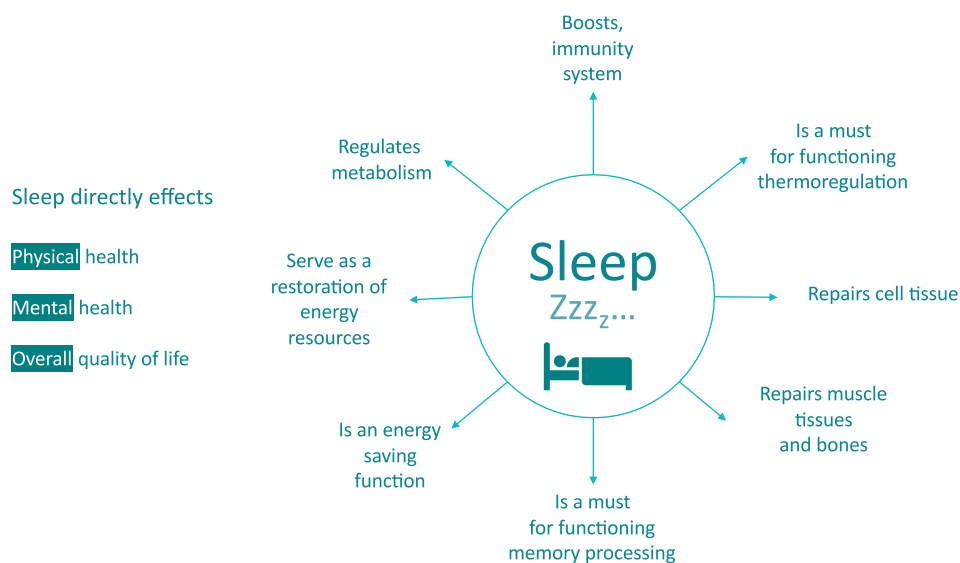


Figure 1.1: Overview of sleep

Sleep scoring is an essential part of sleep studies and diagnosis of sleep disorders. As sleep directly influences our body functions and the quality of life, it is important to improve the diagnosis methods of sleep related diseases. The cornerstone of these procedures is sleep scoring and hypnogram. Hypnogram shows the relative representation of sleep stages throughout sleep and it is useful to save time when evaluating sleep. This thesis focuses on automatizing the process of sleep scoring to save time and avoid subjective mistakes of manual sleep scoring.

1.1.1 Polysomnography

Sleep functions are not fully discovered yet. But the measure of how well a person sleeps significantly affects medical treatments, diagnosis, and clinical follow-ups. In order to evaluate the sleep quality, the patient needs to spend a night in a designated sleep center and record a polysomnography (PSG) [8]. The overnight recording is then analyzed by a physician or sleep experts and utilized as the golden standard for clinical diagnosis of sleep disorders [9].

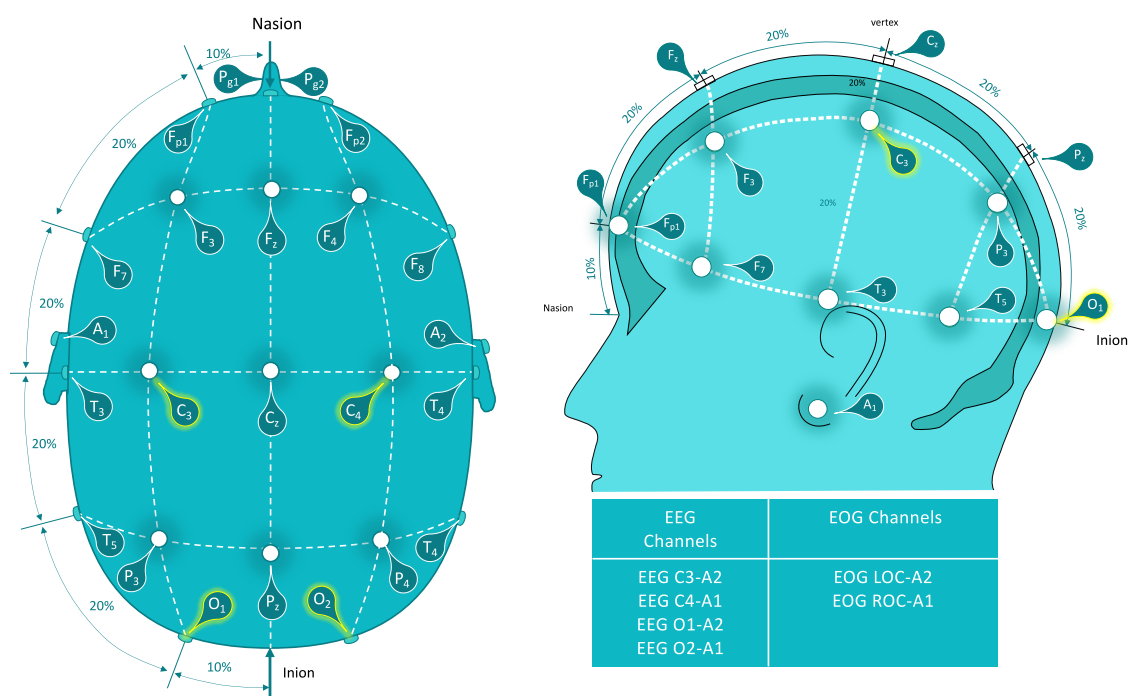


Figure 1.2: 10-20 Electrode Placement System for EEG data collection

Polysomnography was first developed in 1960s, and became a golden standard and the most fundamental approach in studying a sleep behavior of a subject. [10]. Typically, physiological changes are collected using non-invasive surface electrodes such as electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG) during the sleep. These three are the classical physiological signals which are used for the fundamental sleep analysis. Fig.1.2 shows 10-20 Electrode Placement System for EEG data collection. In order to obtain a polysomnogram a whole night recording is performed in a sleep laboratory. Besides the EEG, EOG and EMG signals, signals related to breathing - nasal/oral airflow, thoracic effort and ab-

dominal effort, blood oxygenation, electrocardiography (ECG) and leads are also used to collect data based on the situation.

EEG sensors records the brain activity produced by summation of electrical signals generated by millions of brain neurons. Electroencephalogram signals usually categorized into different frequency bands: delta (0.5 - 2Hz), theta (3 - 7 Hz), alpha (8 - 12 Hz) and beta (12 - 20 Hz). According to many studies, spectral powers of different frequency bands highly correlate with different sleep stages and therefore widely used PSG analysis. Electromyogram (EMG) is the recording of electrical activity produced by skeletal, and electromyogram EOG is a recording of the voltages generated by eyes movements.

A doctor, or a sleep physician can diagnose sleep related disorders or abnormalities using the polysomnography records. Normally a polysomnography record is used to evaluate symptoms of sleep apnea. More detail of this disorder is further discussed in Chapter 2. Beside the sleep apnea, narcolepsy (involves extreme drowsiness and 'sleep attack' during the day time), sleep-related seizure disorders, restless legs syndrome (involves uncontrolled flexing and extension of the legs during the sleep), REM sleep behavior disorder (acting out dreams during sleep, unusual behaviors during sleep), chronic insomnia (difficulty falling asleep or maintaining the sleepiness). Polysomnography testing is typically conducted in sleep laboratories established in sleep clinics and hospitals. PSG is a non-invasive clinical procedure that monitors various sensory data including breathing and cardiac parameters which further used to analyze the health condition of a person [3].

1.1.2 Sleep stages

Sleep stages are scored in 30 second sequential time segments starts from the beginning of the PSG recording. The epochs are then assigned sleep stage based on the rules and regulations defined in sleep stage criteria mentioned above. If more stages are seen during a single epoch, the stage occupies the largest portion in the epoch is assigned.

- Stage ' W ' is the state of wake. It ranges from full vigilance to early stages very light sleepiness

- Stage ' N1 ' is phase that the human body transition in between wakefulness and sleep
- Stage ' N2 ' is main body of light sleep. Memory consolidation is taken place, and muscle activity decreases. The awareness of the outside world starts to fade gradually
- Stage ' N3 ' is also known as deep sleep or slow-wave sleep. In this stage it is difficult to awake and less responsive outside simulations
- Stage ' R ' is the phase where most vivid dreams happen in. Body does not move, and the eyes move rapidly, while chin muscle tone activity stays low
- **Wake (W)**

Based on ASSM manual, at least 50% presence of alpha waves in EEG recordings is the most significant discriminative observation for this sleep epoch. Alpha waves are normally in the range of frequencies from 8 to 13 as shown Hz Fig.1.3 Alpha waves [1]. Even without the alpha rhythm is presented, Wake stage is classified, if any of the following observations are present.

- Eye blinks at a frequency of 0.5 - 2 Hz
- Reading eye movements
- Irregular conjugate rapid eye movements associated with normal or high chin muscle tone

Furthermore, alpha rhythm is presented only with closed eyes. With open eyes alpha rhythm is replaced by low-amplitude mixed-frequency EEG pattern.



Figure 1.3: Alpha waves [1]

- **N1 stage**
An epoch is scored as NREM 1, if it generates alpha rhythm and attenuated and replaced by low amplitude, mixed frequency activity for more than 50% of the epoch.

In epochs, where there is no alpha rhythm presented, the stage is recognized by presence of low-amplitude mixed-frequency waves in the range 4-7 Hz with slow eye movements, or vertex waves. If there are no k-complexes associated with external arousal are presented, the epoch is still marked as N1 [2].

- **N2 stage**

For stage N2, presence of k-complexes and sleep spindles within frequencies of 13-16 Hz is expected for stage N2.

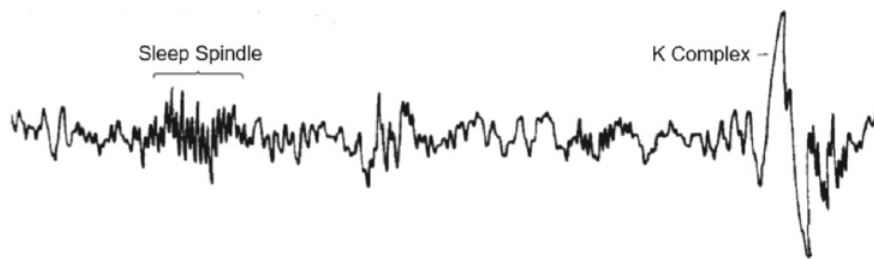


Figure 1.4: k-complexes and sleep spindles [2]

K-complex consists of a negative short wave followed by a positive component standing out from a background EEG as shown in Fig.1.4 [1]. The length of one k-complex is normally less than 0.5 s. Sleep spindles are a train of waves with frequency 11-16 Hz . The duration is about 0.5 s. The same activity as on EEG can be found on EOG. EMG is elevated, but still lower than in stage W. Stage N2 ends when it changes into other stages (N3, W, or REM). It changes into N1 or W arousal observed in EEG, or the conditions of stage N2 are not observed. Major Body Movement (MBM) also can be presented during N2 stage. MBMs are normally followed by slow eye movements and low-amplitude mixed-frequency waves without k-complexes or sleep spindles caused by arousal. If there are no eye movements observed after MBM, the stage is continued to be categorized as N2[2].

- **N3 stage**

Presence of slow waves occurring at least in 20 % of the epoch, can be observed in an N3 epoch. Slow waves have frequencies of 0.5-3 Hz and the amplitude of slow waves has to be at least $75 \mu V$. Sleep spindles are not much considered in marking stage N3. The activity of EOG signal is similar to EEG and the EMG amplitude is changing and relatively low [2].

- **R stage**

Rapid eye moment is normally captured from EOG signal and is the basic observation of the REM. Irregular and sharp waves with initial deflection with a duration of less than 500 ms can also be observed. Besides that, EEG saw tooth waves (in frequencies of 2-6 Hz) can also be appeared. Chin EMG muscle tone is quite low in this stage compared to other stages. Transient muscle activity, (< 0.25 s), can be observed on chin EMG or anterior tibialis EMG. In stage R low-amplitude mixed-frequency waves on EEG likely to be presented. An epoch is still can be recognized as stage R even if no REM occur, when low amplitude mixed-frequency EEG signal persists, while EMG is at its minimum, and no k-complexes or sleep spindles are present [2].

Although, the visual inspection method (manual sleep stage scoring) is the most practiced method in sleep medicine, visual inspection is a laborious and complex task. The manual sleep scoring is associated with expert human interventions. Though sleep stage guidelines are well-defined and standardized, [11] sleep stage scoring involves some vagueness. Such kind of vagueness can be identified as the shakiness of individual interpretation on the sleep rules. Therefore expert-based staging might be bias to some extent. Given the equitable inter-rater disagreement between sleep experts in manual sleep scoring, it is extremely important to develop a reliable automatic sleep stage classification system. The inter rater agreement is only 76.8 % [12], and the inter-scorer agreement in a large group is approximately 83 % [13]. However, an automatic sleep stage scoring system is very practical and efficacious as a computerized assistant. And Also, it would be very useful for sleep technicians and exhibits the potential to decrease the cost of a

sleep study.

1.1.3 Sleep stage scoring

Sleep stages defined based on different physiological and neuronal activities. Sleep scoring, or sleep staging, is the process of classifying these stages is performed by a trained human expert based on visual interpretation of the PSG signals. This process is very critical since sleep stage scoring are the base for further examination. Clinically, sleep technicians follow established guidelines, such as Rechtschaffen and Kales (R & K) criteria (published in 1968 by Rechtschaffen and Kales,) to score the sleep stages manually [14]. In R & K manual, 6 sleep stages can be identified, namely Stage Wake, S1, S2, S3, S4 and (Rapid eye movement) REM. Even though R & K manual is considered as a widely used standard for analyzing human sleep for 40 years, it should be noted that the manual has been criticized for leaving space for subjective interpretation[15]. Lately, American Academy of Sleep Medicine (AASM) has modified the R & K standard guidelines in 2007. According to AASM manual, there are five sleep stages defined namely, stage W, Non-Rapid Eye Movement (stage N1, N2 and N3) and Rapid Eye Movement (stage R).

1.1.4 Hypnogram and sleep cycles

Hypnogram is graphical representation of stages of sleep as a function of time [16, 17]. After determining sleep stages, a visual depiction of the behavior of sleep stage is represented in hypnogram. Furthermore, physicians can examine succession of stages over the night. A hypnogram normally consists of 5 to 6 sleep cycles. Generally, a sleep cycle lasts about 90 to 110 minutes, and during that time the body moves through five stages of sleep. The first sleep cycle normally has a shorter REM sleep and longer deep sleep. But later in the night, REM periods tends to lengthen and deep sleep time gradually decreases.

Fig. 1.5 [18] and Fig. 1.6 (image source : <http://www.ceams-carsm.ca/en/normal-sleep>) illustrate hypnogram of a healthy person. However, in reality, hypnograms differs from person to person. Hypnogram is a very useful in recognizing persons who suffering from sleep

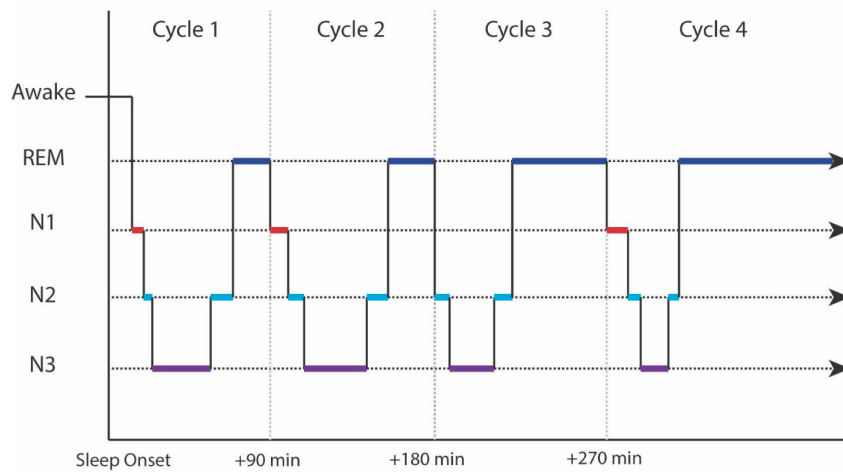


Figure 1.5: Hypnogram and sleep cycles of a healthy person

disorders.

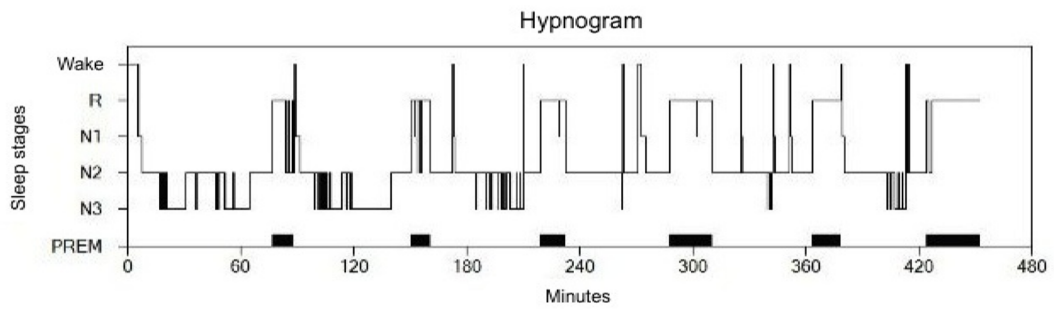


Figure 1.6: Hypnogram of a healthy person

1.2 Dissertation outline

The contents of the rest of the thesis is listed and summarized as follows:

- **Chapter 2** *Literature Review and Background*. This chapter includes a detailed literature review about machine learning techniques and automatic feature extraction methods, a summary of automatic sleep stage approaches, and a literature review on obstructive sleep apnea syndrome is presented.
- **Chapter 3** *Automatic sleep stage detection using raw PSG signals based on deep learning models*. This chapter describes the three methods proposed for sleep stage classification based on deep learning models with two data sets collected for this work. Three experiments performed on automatic sleep stage classification is presented in this chapter.
- **Chapter 4** *Obstructive sleep apnea syndrome detection based on fused Time-Frequency spectral images*. A novel method for obstructive sleep apnea detection based on fused time-frequency representation is presented in this chapter. A combination of spectral images formed with continuous wavelet transform (CWT) and Short Time Fourier Transform (STFT) is used to classify the OSA events from single lead ECG signal.
- **Chapter 5** *Discussion*: This chapter discusses the overall contribution and performances of the proposed methods.
- **Chapter 6** *Conclusion and future works*. This chapter conclude the whole study and future works are suggested based on the conclusions of this study.

2

Literature Review and Background

Recently, several studies focused on developing an automatic sleep stage scoring mechanism indicate that the need for computer assistance for sleep stage scoring is realized and broadened over time. Mainly, the studies can be grouped into two main genres based on the feature extraction techniques. The first type relies on hand-engineered features where prior knowledge of signals is crucial to extract the most significant features correlated with sleep related information. The other studies rely on automated feature extraction, such as deep learning algorithms, where pertained features are used to perform the classification.

2.1 Hand engineered feature extraction methods

As mentioned in chapter 1 Sleep medicine uses polysomnography (PSG) to record biological signals for the analysis of sleep related disorders. Fig.2.1 shows the classical procedure for developing automatic sleep stage scoring systems. The first step is the PSG data acquisition from interested subjects. Secondly the acquired data are pre-processed. The presence of artifacts may cause the misinterpretation of sleep related information such as sharp vertex waves, and K-complexes etc. Noisy PSG data causes significant accuracy drops in automatic sleep scoring systems. Therefore, the preprocessing step is an essential part for prior to any further analysis [19].

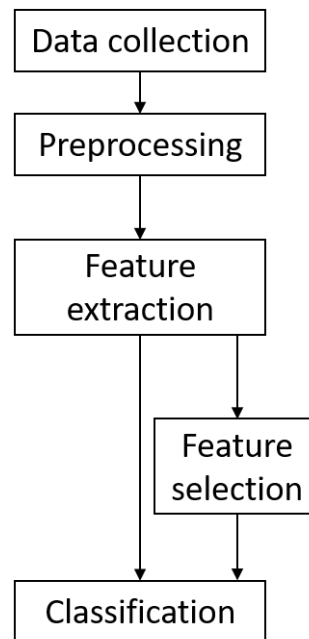


Figure 2.1: Automatic sleep stage classification

2.1.1 Feature extraction and classification

Extracting most informative, discriminative and independent features is the key for any successful classifier[20]. Features can be identified as time domain, frequency domain, time-frequency domain and nonlinear time domain features such as statistical parameters (mean, standard deviation, skewness, kurtosis etc), zero crossing rate, and Hjorth parameters represent

the morphological characteristics of a signal [21][22]. Time domain features are widely used in real-time applications. Frequency domain features are one of the most frequently utilized type of features in EEG signal based research problems. To obtain spectral characteristics of the time domain, the signal should be transferred to the frequency domain using Fourier transform (FT). Another way of obtaining meaningful features is time-Frequency domain analysis. Due to the non-stationary nature of EEG signal, time-frequency features are very useful and efficient in extracting information. One such kind of very simple time-frequency analysis is Short Time Fourier transform (STFT). In STFT, signal is uniformly windowed and then FT is applied to each window to form the time frequency representation of the signal. Wavelet Transform (WT) is also a popular time-frequency transform in the field which utilize a bank of filters to decompose a signal different into frequency scales[23].

As shown in Fig.2.1, feature selection is the next part. in the feature selection section a minimum number of features is chosen to reduce redundancy. Statistical techniques such as sequential forward selection (SFS) and sequential backward selection (SBS) are popular in the field as simple feature selection methods[24][25]. The final section is classification, which separates the feature vector into classes. Some of the popular classifiers used in sleep stage scoring are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA) & Nearest Centers (NC), Neural Network (ANN), and Hidden Markov Model (HMM).

2.1.2 Support Vector Machine

Support Vector Machine (SVM) is a widely used supervised classification method in many classification problems. Each data item can be represented as a point in n-dimensional space (where n is number of features). The value of each feature is then be a particular coordinate. In SVM algorithm, the classification is performed by finding the best hyper-plane that separates the two classes with the maximum the margin width around the separating hyper-plane while minimizing the training error. SVM is trained using Lagrange method, by considering the problem as constrained optimization problem. SVM is recognized as a good classifier because of

its appealing features such as clearer separating margin, effectiveness in high dimensional margins where samples are lesser than the number of dimensions, and memory efficiency. And also, SVMs are criticized because of its higher training time for large data sets, inability of performing well for noisy data. However, SVM has been successfully adopted for sleep stage classification in previous studies [26][27][28][29].

2.1.3 K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is one of the simplest algorithms among all machine learning algorithms used in many classification problems. This algorithm assumes that similar things exist in close proximity. In other words, KNN exploits the fact that similar things are near to each other in the feature space. KNN assigns a label to input patterns depending on the majority vote of k-nearest samples [30]. A typical Euclidean distance measure is used to calculate how far each sample is to the target class.

2.1.4 Random Forest

Random Forest (RF) consists of an ensemble of tree-structures [31], where each individual tree plays the role of a single classifier. In RF, the training samples are fed to the trees as random as possible through a random selection followed by different bootstrap selections. This process continues several times to blend the samples in the training phase. The output is determined using a voting method of each trees' outputs. In order to train a RF classifier, an arbitrary number (usually large number of trees are used) of decision tree are randomly generated. Each tree in the tree forest are trained independent to each other so that they have a partial observation of the train samples. The performance of a random forest depends on its trees' individual performance [32].

2.1.5 Linear Discriminant Analysis & Nearest Centers

Linear Discriminant Analysis (LDA) is developed in 1936 by Fisher. In binary class cases, LDA can be considered as a classifier and, is used as a feature-extraction method in other cases.

LDA provides separable features for the next classifier. In implementing LDA, input samples are projected onto a few number of hyper-planes (depends on the number of classes) such that the separability is maximized in the projected space. LDA is optimized by a Fisher criterion. The final decision or the classification is made by utilizing a distance-based classifier to the LDA outputs. LDA works as a feature extractor and the projected features are then put into the relevant classes based on the minimum distance of the input to the center of each class [33].

2.1.6 Artificial Neural Network (ANN)

ANN can be identified as an artificial information processing system associated with interconnected nodes called neurones. An artificial neural network works similarly to the human brain's neural network. These networks are designed following the mechanism of real biological neural cells. A "neuron" can be recognized as a mathematical function that accumulates and classifies information according to a specific rule or function. The individual neurons employ activation functions such as sigmoid, hyperbolic and linear functions. Generally a neural network contains multiple layers of interconnected nodes. Once the number of hidden layers gets large, the network become deep, in which the memory of the network is also improved to conserve the information of input samples. Recently, ANNs are heavily utilized in automatic sleep stage scoring [34][35][36][37]. The schematic diagram of a feed-forward multi-layer neural network is shown in Fig.2.2.

ANNs are trained by back propagation algorithms. Even though training of an ANN is relatively slower, ANNs are comparatively fast in the test phase. One of the major drawback of ANNs is the over-fitting to the noisy samples placed along the margin space between classes. However, the hyper parameter such as network size , layer size, and activation functions should be selected carefully since such parameters can affect the classifier's performance.

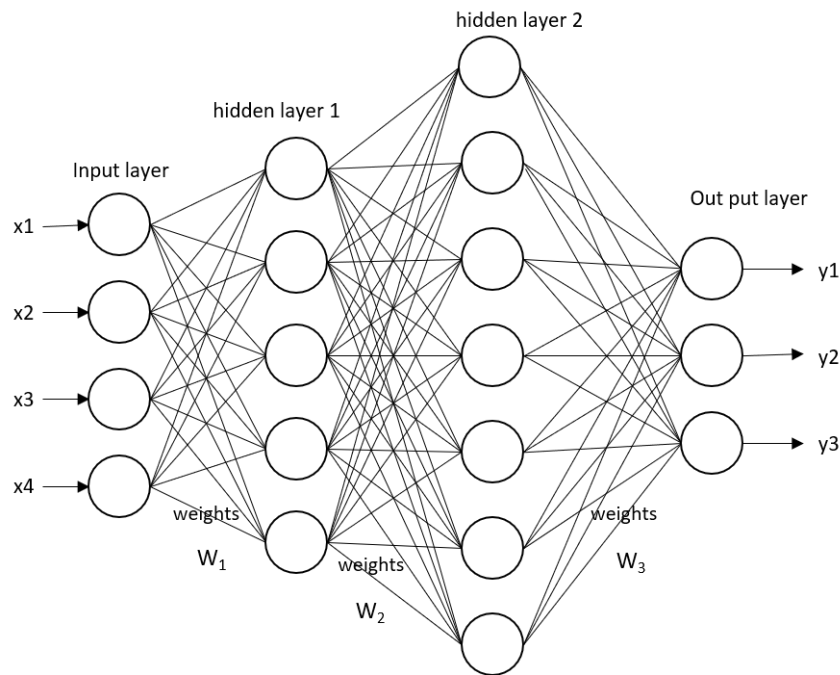


Figure 2.2: ANN architecture, feed-forward network, 'W' represent the weight matrices , and 'x' and 'y' represent inputs and outputs respectively

2.2 Automatic feature extraction methods involved with sleep stage classification

Mostly, automatic feature extraction based methods use pretrained Networks. Chambon *et al.* in has proposed end-to-end deep learning approaches to perform automatic sleep stage classification based on multi-channel and multi-model PSG signals [38]. Their method can correctly classify the sleep stages with a confidence of 91 %. In summary, a deep architecture can extract information from EEG, EOG, and EMG channels and perform classification using learned representations along with an end classifier. Furthermore, Akara *et al.* [39] proposed a deep learning model, named DeepSleepNet, based on raw single-EEG channel without utilizing any hand-engineered features. The approach uses CNN based feature extraction method and bidirectional long-short term memories (LSTM) to perform the classification of sleep stages from raw EEG epochs. CNN combined with fine-grained segments is successfully used to classify sleep stages with different combinations of multiple PSG channels including EEG, EOG,

and EMG [40]. The experiments performed with the proposed CNN architecture for multiple channels have been obtained the highest accuracy of 90.11 % for 11-channel configuration [40]. In addition to the 11-channel configuration, they proposed 5-, 7-, and 9-channel configuration with classification confidence of 83 %, 88 %, and 89 % respectively. The present study emphasizes on the capability to handle different kinds of physiological signals including EEG and EOG.

2.2.1 A brief summary of automatic sleep stage approaches

As explained above, the EEG is the best signal which represent the electrical activity of the brain. EEG patterns show various behaviors during sleep stages. These kind of behaviors have successfully been used to develop automatic sleep stage scoring systems. Table.2.1 shows research approaches that use only EEG signals to classify the sleep stages.

Electrical activity of the human heart is monitored with ECG recordings. Some studies show that sleep stage classification with ECG is less complex compared to complete PSG analysis [47–49]. Table.2.2 summaries such kind of methods which use ECG signals to score the sleep stages.

Human scorers use a combination of multiple physiological signals to determine the sleep stages. Even though multiple signals introduce redundant in formations, it is important for sleep technicians to confirm an interested event from different channels. The standard method to diagnose sleep disorders rely on multiple PSG signals. EEG signals in combination with other PSG signals, such as ECG, EOG and EMG have also been utilized to implement computer based sleep stage scoring systems as shown in Table.2.3.

2.3 Obstructive sleep apnea syndrome

2.3.1 Sleep apnea

Sleep apnea (aka. sleep apnoea, SA) is one of the most common chronic diseases and is caused by the complete or partial discontinuation of airflow that accompanies an obstruction

Table 2.1: Research works that used EEG signals to perform sleep stage classification

Author	Data	Feature extraction method	Classification	Results (Accuracy %)
Mousavi <i>et al.</i> , 2019 [41]	The benchmark Sleep-European Data Format (EDF) dataset	time and frequency-domain, sequence to sequence features	Deep learning	84.26% (two class)
Michielli <i>et al.</i> , 2019 [42]	The benchmark Sleep-EDF dataset	time and frequency-domain features	Deep learning	83.6% (two class)
Bajaj and Pachori, 2013 [43]	Sleep-EDF database	time-frequency image based on the Wigner-Ville distribution (WVD)	Multiclass least squares SVM.	88.47
Shi <i>et al.</i> , 2015 [44]	25 adult subjects, Sleep Apnea Dataset provided by St. Vincent's University Hospital and University College Dublin.	A two-stage multi-view learning algorithm based on a joint collaborative representation	K-means clustering	81.10
Koley and Dey, 2012 [45]	28 subjects aged between 35 and 56 suspected to have sleep apnea	SVM based feature elimination technique	Binary SVMs, one against-all strategy.	85
Hsu <i>et al.</i> , 2013 [46]	Sleep-EDF (Fpz Cz channel)	Energy feature extraction using FIR band pass filters	Recurrent neural classifier	87.2

Table 2.2: Research works that used ECG signals to score sleep stages

Author	Data	Feature method	extraction	Classification	Results (Accuracy %)
Yu celba, s <i>et al.</i> , 2018 [47]	Sleep laboratory of Necmettin Erbakan University database and PhysioNet	Morphological methods	Random Forest, Wake, Non-REM, REM (WNR)		Up to 87.11
Yoon et al., 2017 [50]	21 healthy subjects and 30 subjects with Obstructive Sleep Apnea (OSA) recorded at Seoul National University Hospital	HR Statistical parameters, Spectral power, variability measurements	Threshold, REM duration		87.54
Kesper <i>et al.</i> , 2012 [49]	Apnea-ECG and SIESTA Database	HR, Spectral power evaluated by ANOVA	threshold		57.8
Xiao et al., 2013 [48]	Public database Sleep and Stroke Volume Data ear Bank	HR, linear spectral power, nonlinear ear	WNR, random forest		88.67
Redmond <i>et al.</i> , 2007 [51]	31 male subjects	ECG derived respiration and HR statistics.	WNR, Discriminant Analysis (LDA) and a quadratic LDA.	Linear	up to 76.1
Mendez <i>et al.</i> , 2010 [52]	24 subjects	HR statistics power	Spectral REM-NREM, HMM.		79.3

Table 2.3: Research works that used combination of PSG signals to score sleep stage

Author	Data	Feature extraction method	Classification	Results
Fonseca <i>et al.</i> , 2015 [53]	Data from 48 subjects	ECG: Spectral variability, measurements, and network analysis. PSG: Time/frequency, and network analysis	LDA, NREM and REM	Wake, 80
Holland <i>et al.</i> , 2015 [54]	EEG, ECG and respiratory signals from the SIESTA database	HR: statistics. PSG: Time/frequency, and network analysis	LDA, Wake, REM and REM	80
Willemen <i>et al.</i> , 2014 [55]	36 healthy subjects	HR statistics and spectral power, Breathing Rate (BR) statistics, and movement statistics	SVM WNR	81
Kushida <i>et al.</i> , 2001 [56]	Full PSG 100 patients with sleep disorders	Visual scoring of wake and sleep states	Threshold	Accuracy-77

of the upper airway for a short time [57, 58]. A complete pause of at least 10 s in the airflow through the upper airway during sleep is usually considered as an apnea-episode as shown in Fig.2.3.

Mainly three types of sleep apneas can be recognized based on the breathing style, namely central SA (CSA), obstructive SA (OSA), and mixed apnea [58]. CSA is mainly caused by the instability in the central nervous system. In a CSA episode, a blockage of the airway at the back of the throat causes to create an apneic event. If the airway block is disturbed partially, then the pathology is termed hypopnea [59]. Such a hypopnea event involves at least 10 s of shallow breathing. This kind of shallow breathing lowers the air volume entering the lungs to low levels (below normal levels) and causes blood-oxygen desaturation of at least 4%. An occurrence of OSA is recognized when a person has a complete airflow pause in the upper airway for at least 10 s. During an OSA event, the airway is clogged while there are still respiratory efforts happens against the obstruction [60]. Furthermore, mixed apnea is identified when the apnea starts as a CSA and terminates as an OSA. Therefore, mixed apnea has both features from CSA and OSA.

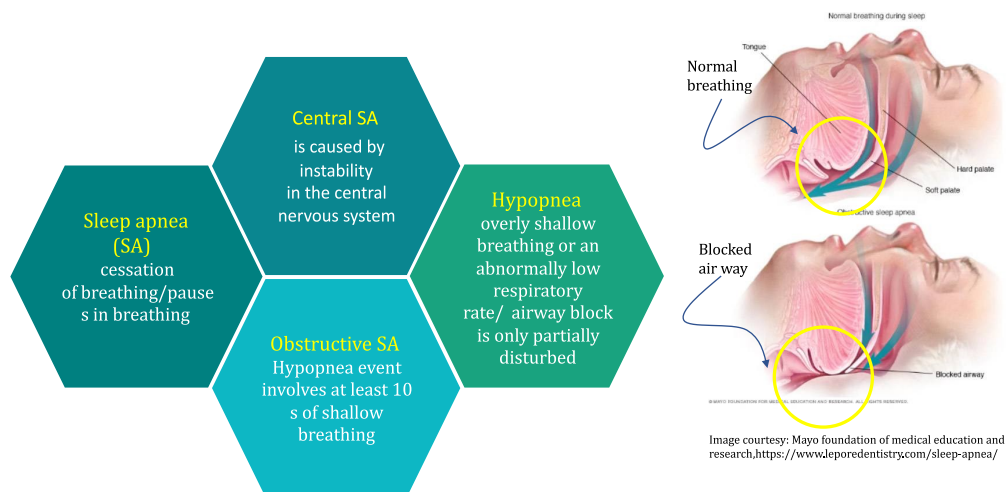
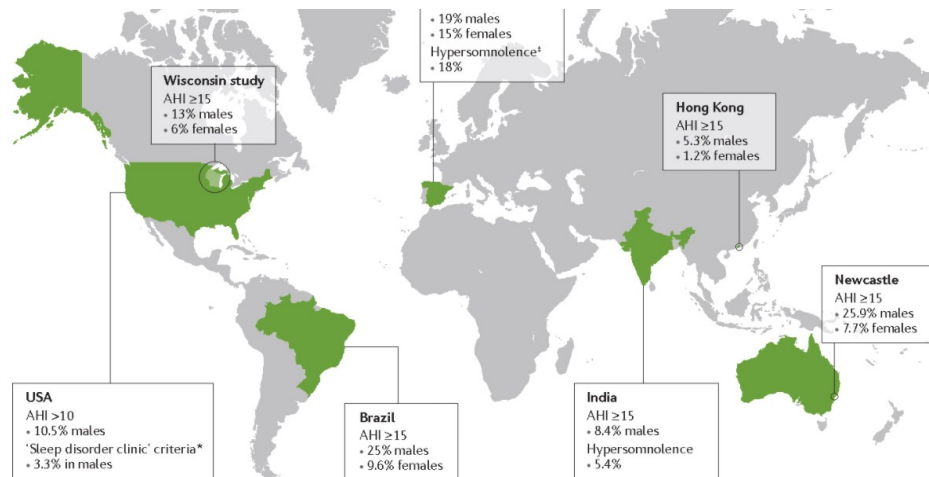


Figure 2.3: Sleep apnea syndrome

Undiagnosed and untreated repetitive apneic episodes can cause a variety of health complications, including excessive sleepiness in daytime, cardiovascular and neurological issues such as memory impairment, high blood pressure, acute coronary syndrome, and congestive heart

failure [60, 61]. According to previous studies [58, 62, 63], around 3–7 % of adult men and 2–5 % of adult women worldwide suffer severely from SA. Specifically, 14 % of men and 5 % of women in the United States suffer from OSA syndrome, and the prevalence of the OSA is continuously growing in various populations worldwide[64].



Nature Reviews Disease Primers
Obstructive sleep apnoea syndrome Patrick Lévy *et. al*

Image curtesy: <https://www.semanticscholar.org/paper/Obstructivesleep-apnoea-syndrome-L%C3%A9vy-Kohler/011dd0f204957e2805d09bafd77b3da171c1baab>

Figure 2.4: Prevalence of obstructive sleep apnea

In clinical work, the seriousness of apnea and hypopnea events is qualitatively measured using the apnea–hypopnea index (AHI) (*see Fig.2.5*). The AHI value is defined as the average number of apnea/hypopnea episodes occurring within one hour. In general, a subject showing an AHI value greater than five is considered as a person with SA [59, 65]. A mild OSA case is identified if the AHI value lies between 5 and 15. The moderate OSA patients show AHI values between 15 and 30. Severe cases have AHI values above 30 [66].

The most common diagnostic method for OSA suspects is polysomnography (PSG). In PSG, various physiological signals are acquired from sleeping patients including, airflow, respiratory effort, electroencephalogram (EEG), electrocardiogram (ECG), and oxygen saturation (SpO_2). The specific patterns in these physiological signals are then scrutinized by sleep experts to detect sleep-related disorders such as OSA. Therefore, PSG is known as gold standard for OSA detection for long time and the the study is also used to perform comprehensive evalua-

tion of the cardio-respiratory system. More detail about PSG can be found in subsection 1.1.1. Even though OSA detection based on PSG test, the some cases are still not recognized [67].

A typical PSG test requires dedicated nursing staff and expensive medical equipment specifically designed for polysomnography related bio data acquisition. The PSG diagnosis method therefore needs dedicated supervision of expert human personals. Besides that, a normal PSG test is time-consuming, expensive, and uncomfortable for patients since many sensors are attached to the body when the subject is sleeping. One of a main objectives of this study is to minimize these technical and economic complications of sleep studies based on conventional PSG studies. There are number of automatic SA detection methods proposed during the past two decades. These methodologies are usually relies on the analysis of the cardiopulmonary (CP) bivariate signal (a combination of heart rate (HR) and respiratory rate (RR) signals), or ECG-derived respiration (EDR).

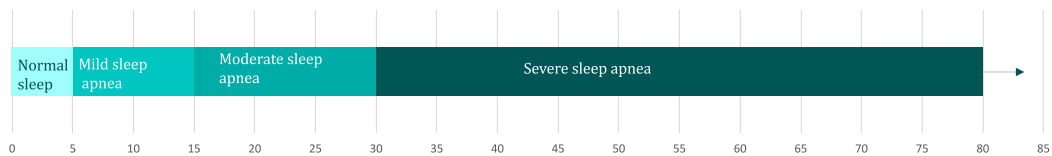


Figure 2.5: Apnea–hypopnea index (AHI)

Recording respiratory activity via sensors positioned around the nose is uncomfortable for the subject. Because, respiratory activity is completely or partially clogged during an apnea/hypopnea episode, variations in the RR signal can be observed. Therefore the RR signal is sometimes obtained indirectly via EDR signals or inductance plethysmography [68, 69]. The EDR signal is widely used to detect sleep associated disease or issues since it accurately reflects respiratory activity during the the sleep. Besides that, ECG electrodes can be easily attached to the body without disturbing sleep, with comparison to direct respiratory sensors placed close to the nose. Therefore, ECG signal can be identified as an appealing option for apnea detection .

Some studies have shown that SA event is strongly associated with variations of the signal, including Heart Rate Variability (HRV), morphological variations in the ECG signal [70–72], and variations in the ECG signal’s QRS duration [73, 74]. Therefore, many studies have been carried out based on these observations, and algorithms based on morphological-variation fea-

tures have tended to show improved performance [71, 75].

This experiment also focuses on ECG signal variability, HRV, and morphological variations during an apneic event. This study aims to capture such features using time-frequency representations. A fused combination of time–frequency representations (TFRs) is utilized to intensify the HRV-based and QRS-based variations in ECG signals. The presence or absence of apneic episode is then detected by means of a deep convolutional neural network (CNN) using a fused spectral images generated from two TFR (scalograms and spectrograms). Compared with other existing methods, this experiments showed improved performance in detecting apneic events because it not only combines a variety of TFRs but also exploits recent advances in CNN-based classifiers.

2.3.2 OSAS and apnea detection)

During the previous two decades, many works have been proposed to separate sleep apnea events from the normal event, using a number of physiological signals including ECG, EDR, and respiratory signals [76–78]. According to Guilleminault *et al.* [79], the presence of an apneic episode is related to the concomitant variation in the RR intervals in the ECG signal. A considerable number of research attempts have been made to implement automatic OSA-detection methods using a single ECG lead. Khandoker *et al.* have proposed an sleep apneas detection method using features extracted from successive wavelet-coefficient of the RR intervals and the EDR signal from the R waves in the QRS complex [80]. They have successfully adapted a support vector machine (SVM) classifier to perform the classification. In their work, more than 90 % of test subjects were identified so that the apnea cases can be distinguished from the normal cases. Song *et al.* developed a apnea detection method using a discriminative hidden Markov model (HMM) based on the ECG signals for ECG segments [81]. In that study, frequency-domain and time-domain features were extracted from the EDR and ECG signals were used to distinguish the apneic events. The per-segment detection accuracy of *Songs* model was 86.2 % for PhysioNet Apnea-ECG database.

Kunyang *et al.* have also proposed a neural-network (NN)-based model that used an HMM

for identifying sleep apnea episodes using ECG signal [82]. In their work, a combination of sparse auto encoders, NNs, and HMMs was used to develop the framework. A classification accuracy of 84.7 % was achieved for per-segment apnea detection. Hayano *et al.* proposed a screening method for OSA using cyclic variations in heart rate (CVHR) [83]. The agreement between the SA and the presence or absence of CVHR in each one-minute period was found to be 83 %. Sharma *et al.* achieved 84.4 % accuracy for one-minute ECG signals in detecting apnea using a least-squares (LS) SVM classifier with a Gaussian radial-basis-function (RBF) kernel for features derived from Hermite expansion coefficients [75]. Later, Viswabhargav *et al.* proposed an apnea detection method whereby EDR and RR signals were utilized to extract sparse residual entropy (SRE) features, using an SVM classifier [65]. In their study, an RBF-kernel-based SVM classifier achieved 85.43 % sensitivity and 92.60 % specificity for the SRE features.

Tripathy *et al.* introduced a novel method that analyzed the CP signal using fast and adaptive bivariate EMD coupled with cross time-frequency [84]. The CP signal was formulated using both the HR and RR signals derived from the ECG signal. Their method achieved average sensitivity and specificity values of 82.27 % and 78.67 %, respectively, using an SVM classifier and “random forest” classifiers in a 10-fold cross-validation method.

Singh *et al.* proposed a method based on the heartbeat interval and EDR, where sliding-mode singular spectrum analysis was used to extract features, with sensitivity and specificity values being 82.45 % and 79.72 %, respectively [85].

In these methods, many of the features of the ECG signal used in the classification are derived manually. These include waveform parameters such as instantaneous amplitude (IA) and instantaneous frequency (IF), residual entropy features, statistical features, and other specifically derived features.

Table 2.4: Comparison of automated systems proposed for OSA detection using CinC-2000 PhysioNet database [3]

	Author	Methodology	Performance
1	Chazal <i>et al.</i> [86],2000	<p>No. of Features: 128</p> <p>Features: Statistical features from RR duration, PSD of RR interval, R-peak Spectrum of R-wave</p> <p>Classifiers: Linear discriminants (LD)</p>	<p>Avg.acc= 89.8%</p> <p>Avg.sen= 86.5%</p> <p>Avg.spec= 91.9% (using 35-fold cv)</p>
2	Chazal <i>et al.</i> [87],2003	<p>No. of Features: 88</p> <p>Features: Statistical features of EDRP and RR duration Power spectral density of RR interval Power spectral density of EDR signal</p> <p>Classifiers: LD, quadratic discriminants (QD)</p>	<p>Avg.acc = 90%</p> <p>Avg.sen = 86.4%</p> <p>Avg.spec =92.3% (using 35-fold cv)</p>
3	Chazal <i>et al.</i> [88],2004	<p>No. of Features: 88</p> <p>Features: Statistical features of EDRP and RR duration Power spectral density of RR-interval Power spectral density of EDR signal</p> <p>Classifiers: LD</p>	<p>Avg.acc = 89.5%</p> <p>K-value =0.8 (Using 35-fold CV)</p>

4	Babaeizadeh <i>et al.</i> [89],2010	No. of Features: 12 Features: Spectral Power in frequency ranges different Classifiers: Quadratic Classifier	Avg.acc= 84.70% Avg.sen= 76.70% Avg.spec= 89.60% Training:testing Data = 50:50
5	Bsoul <i>et al.</i> [90],2011	No. of Features: 111 Features: Daubechies DWT based Statistical features From EDRP and HRV in time domain: 20 From EDRP and HRV in spectral domain: 91 Classifiers: Gaussian SVM	Avg.acc= 89.08% Acse=96.05% F1-score = .90 (using 30-fold cv)
6	Xie and Minn [91], 2012	No. of Features: 111 Features Statistical features in time domain 20 Statistical features in spectral domain 91 Classifiers: Bagging with REPTree (Reduced Error Pruning Tree)	Avg.acc= 77.74% Avg.sen= 69.82% Avg.spec= 80.29% (using 10-fold cv)
7	Liu <i>et al.</i> [92],2012	Features: Hilbert Huang transform	Avg.acc = 79.10% Avg.sen = 73.10% Avg.spec = 71.2% Without k-fold CV
8	Kesper <i>et al.</i> [93], 2012	Features: Hilbert Huang transform Statistical features in spectral domain	Avg.acc= 80.50% Data = 50:50 Training:testing

9	Sadr and Chazal [94], 2014	No. of Features: 68 Features: Statistical features in time frequency domain 4 and Power spectral density of interval: 32 RR Power spectral density of signal: 32 EDR Classifiers: Extreme learning machine classifier (ELM)	Avg.acc= 87.70% Avg.sen= 81.30% Avg.spec= 91.70% (using 35-fold cv)
10	Nguyen <i>et al.</i> [95], 2014	No. of Features: 32 Features: Recurrence Quantification Analysis (RQA) features of Heart Rate Variability (HRV) Classifiers: SVM and Neural Networks (NN)	Avg.acc = 85.26% Avg.sen = 86.37% Avg.spec = 83.47% Using three-fold CV
11	Hassan <i>et al.</i> [96], 2015	No. of Features: 25 Features: EMD based Statistical features of intrinsic mode functions (IMFs) Classifiers: Extreme learning machine (ELM)	Avg.acc= 83.77% Without cv

12	Varon <i>et al.</i> [97],2015	No. of Features: 28 Features: Standard deviation of RR (S1) serial correlation coefficients (r): 5 Standard deviation of EDR (ECGDerived respiration signals (S2): 3 Principal Component's relative power (PC) orthogonal subspace projections(F): 18 Classifiers: Least Squares-Support Vector Machine (LS-SVM)	Avg.acc= 84.74% Avg.sen= 84.71% Avg.spec= 84.69% Using 10-fold cv
13	Hassan <i>et al.</i> [98],2016	No. of Features: 36 Features: TQWT based -scale factor of normal inverse gaussian probability distribution function (NIG pdf): 18 tail Heaviness of NIG pdf: 18 Classifiers: Adaptive Boosting	Avg.acc= 87.33% Avg.sen= 81.99% Avg.spec= 90.72% Training:testing Data = 50:50
14	Surrel and Murali [99], 2018	Features: Power spectrum analysis	Avg.acc= 83.2% Without k-fold cv
15	Li <i>et al.</i> [100],2018	Features: Markov model based features with decision fusion, SVM and ANN classifiers	Avg.acc= 84.7% Avg.sen= 88.9% Avg.spec= 82.1% Training:testing Data = 50:50

Some features are derived from the QRS complex and selected manually. In some cases, much manual preprocessing is required when performing specific derivations, including EDR

signal extraction and QRS localization prior to the extraction of specific features.

Moreover, most existing approaches use frequency-domain and time-domain representations and nonlinear features derived from physiological signals, where substantial knowledge and relevant experience is required. To address this issue, Wang *et al.* proposed a method based on a modified LeNet-5 CNN, where feature extraction is automated with an accuracy of 87.6 % in the classification of OSA [101].

Recently, deep learning has become widely implemented in medical imaging and signal analysis because of its advances in pattern recognition and image-based studies. Researchers have also used deep-learning techniques to address ECG-related research issues such as arrhythmia detection [102–107] and other research applications [108, 109]. In these studies, deep neural networks (DNNs) were introduced successfully to extract descriptive and distinguishable features automatically from the input data, which were then used to perform the classification.

McNames *et al.* employed spectrogram signatures calculated from ECGs via Fast Fourier Transform (FFT) to classify OSA [110]. They obtained a case-based detection accuracy of 92.6 %. Singh *et al.* proposed a method based on ECG scalograms that were created via wavelet transforms to detect OSA using a DNN [111]. Their method achieved an accuracy of 86.22 % and a sensitivity of 90 % in per-minute OSA classification.

3

Automatic Detection of Sleep Stage Using PSG Signals

This chapter describes three methods to use artificial neural networks and deep neural networks for sleep stage classification. Mainly the proposed methods are trained with fairly large data sets from healthy and non healthy subjects. The architecture of the proposed methods and the training procedure is discussed in-detail in later sections of this chapter.

3.1 Dataset and pre-processing

Two PSG datasets were used in this study. The data were recorded with two devices in two different places. The dataset(D1) was performed at the Biomedical Information Engineer-

ing Lab, The University of Aizu, Japan with healthy volunteer subjects. The second data set (D2) was recorded at the Fukushima Otsuki Clinic, Fukushima, Japan between 2014.02.03 – 2016.02.17, and all recording were recorded from sleep patients. All the recordings have been recorded by Premium Alice 6 LDxS PSG Sleep Systems (Philips Corp. USA). Mainly these recordings consist of 4 EEG, 2 EOG channels, one EMG channels. Besides that, recordings of ‘RR’, ‘ECG II’, ‘Leg 1’, ‘Leg 2’, ‘Flow Patient’, ‘Flow Patient’, ‘Snore’, ‘Effort THO’, ‘Effort ABD’, ‘Body’, ‘Pleth’, ‘SpO2’, ‘Technical’ were included in those recordings. Basically, all the data processing and analysis were performed with Python programing language.

To convert the recorded signal into a representative signal which can be further worked with, the original signal should be processed. In other words, biomedical signals are often corrupted or deformed by other waves during the process of data collecting. Mostly these signals are likely to be affected by biologically or other technical issues. The recorded data set is originally sampled with 200 Hz. Artifacts can be identified as breathing movements artifacts, pulse artifacts caused by wrong placement of electrodes, power line hum etc. Typically, ECG artifacts such as spike occurrence in the stage of QRS complex, and EOG artifacts caused by blinking and eye moments can be identified as common biological artifacts.

However, some hardware pre-filters have already been applied for EEG and EOG signals as defined below. Table3.1 shows parameters for a EEG signal corresponds to an one hour time frame.

Table 3.1: Parameters of raw EEG signal

parameter	value
samples in file	720000 samples
maximum	300 μV
minimum	-300 μV
digital maximum	32767
digital minimum	-32768
prefilter	HP:0.32 Hz LP:93.6 Hz N:50 60Hz
sample frequency	200 Hz

Each recording was found as one-hour long data chunk (.edf file), and the sleep stages correspond to the PSG recordings have been annotated by professional sleep technicians. Basically

6 channels consisting of EEG (C3-M2, C4-M1, O1-M1); EOG and (LOC-M2, EOG ROC-M1) were taken into consideration in the analysis. The sleep stages of each recording have been manually scored by professional sleep technologists as wake, REM, S1, S2, S3, S4, MOMENT, and NOT-SCORED epochs. The PSG data has been collected as one hour long data chunk .edf file. As the first step, PSG data were extracted from the raw EDF (European Data Format) files using a python library called PyEDFlib. After the annotated labels were extracted from the .rml files provide by expert sleep technicians. The annotated data were then rearranged by merging S3 and S4 into one sleep stage, so that the analysis can be done using five stages (Wake, N1, N2,N3,REM). Epochs labeled as NOT-SCORED and MOMENT were excluded as they were not included in five stage sleep scoring criterion. All the PSG channels taken into analysis were sampled at 200 Hz. For the sake of simplicity, each PSG signal was then decimated by 2. After then each signal was feature-scaled using min-max normalization method before it was segmented into segments of 30 s. All the signals have been filtered with hardware filters. However, in some patients, some parts of the PSG signals have been heavily affected by the movements of the body. Fig. 3.1 shows such kind of effected epoch due to the movements of the electrode placement, and Fig. 3.2 shows a normal epoch.

Finally 30 s of time segments extracted from each PSG signals were concatenated into $(6*3000,1)$ array and the corresponding label was attached to each epoch. Python Data Analysis Library called pandas – was used to perform the data rearrangements discussed above. After performing all the rearrangements and artifact removing, labels were converted into one hot encoding and saved as Python numpy arrays for further analysis. Each numpy array contains shuffled epochs from 10 subjects. The preprocessing block diagram is shown in Fig. 3.3

3.2 Evaluation metrics for multi-class scenario

To evaluate the proposed models, we use overall accuracy (acc), per-class recall (RE), per-class precision (PR), and per-class F1-score (F1) as defined in (3.1), (3.2),(3.3), and (3.5), respectively [112]. Furthermore, we computed Cohen’s kappa coefficients (κ).

$$\text{acc} = \sum_{c=1}^C \frac{TP_c}{N} \quad (3.1)$$

Recall (also termed as true positive rate, sensitivity, probability of detection) reflects the correctly predicted proportion of all positive samples.

$$\text{recall} = \frac{TP}{TP + FN} \quad (3.2)$$

Precision (also termed as positive predictive value) reflects the proportion of positive predictions that is actually correct.

$$\text{precision} = \frac{TP}{TP + FP} \quad (3.3)$$

Specificity (also known as True Negative Rate) reflects the proportion of negatives that are correctly detected.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3.4)$$

The F1 score denotes the harmonic mean of precision and recall, and thus considers both metrics into an optimal blend for analyzing model performance.

$$F_1 = 2 \cdot \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (3.5)$$

where TP, TN, FP, FN, and N denote true positives, true negatives, false positives, false negatives, and total samples, respectively. TP_c is the true positives of class c.

For per-class metrics, the interested class is considered as a positive and the rest is considered as negative class. Furthermore, we computed area under curve-receiver operating characteristic curve (AUC - ROC) to evaluate the performance of our model. The area under the curve (AUC) of a receiver operating characteristics (ROC) curve is used to check or visualize the performance of the multi-class classification problem [113].

3.3 Experiment I (CNN model)

In the experiment I, only patient data set was used. The sleep stages correspond to the PSG recordings have been annotated by professional sleep technicians. In a nutshell, 176,864 epochs were used to train the proposed in experiment 1 and 3,857 epochs were used to evaluate the network as shown in Fig. 3.4

The architecture of the proposed method I is consisting of two CNN branches as shown in Fig. 3.5. Mainly, the model starts with 2D input array (6,3000) as shown in Fig. 3.8. This is formed by rearranging the epoch discussed above. As shown in Fig. 3.5 three preliminary convolutional layers are used to extract sleep related the appropriate temporal and frequency domain information. Secondly, the network is divided into two branches formed with different filter sizes and different stride values. The first set of convolution layers is formed by rectangular shaped kernels ((1,2),(1,4) and (1,4)) followed by rectangular shaped max pooling layers, which affects to one channel when convolving. The two branches are mostly formed with square shaped kernels, which can extract details from multiple channels at a time. As illustrated in Fig. 3.5, the final section consists of several fully connected layers with Sigmoid and ReLU activations. SoftMax layer with 5 units corresponding to each sleep stage was utilized as the classification layer. The proposed CNN was designed and trained from starch using Keras with TensorFlow as back-end.

3.3.1 Training the network

Adam optimizer was used to minimize categorical cross entropy between predicted classes and actual labels on mini batches size of 100 epochs (training samples). The learning rate was set to 10^{-4} and, Tensor Board was used to visualize learning curves to further fine tune (adjusting hyper-parameters) the network using test data. Validation accuracy and loss were used as an early stopping criterion to avoid over-fitting. Since the data set is fairly large, it is not possible to load the training data set at once. Therefore, Keras “fit_generator” function was used to train the network.

Especially, Adam optimizer is a very popular optimizing algorithm due to its appealing characteristics such as adaptive learning rate and fast convergence over mini batch updates [114][115]. The update rules of Adam optimizer is illustrated in equation (3.6).

Specifically, Adam optimizer algorithm concerns an exponentially decaying average of the gradient \hat{m}_t and exponentially decaying average of past squared gradients \hat{v}_t to control the adaptive learning rate with constant values β_1 , β_2 , and ϵ . We used the default values of 0.001 for η , 0.9 for β_1 , 0.999 for β_2 , and 10^{-8} for ϵ , as suggested in the previous study [115].

$$\vartheta_{t+1} = \vartheta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (3.6)$$

where, \hat{v}_t denotes the exponentially decaying average of past squared gradients, \hat{m}_t denotes the exponentially decaying average of past gradients, ϑ denotes model parameters, η denotes learning rate, and ϵ denotes a very small number to prevent any division by zero.

3.3.2 Results

Fig. 3.6(a) shows the normalized confusion matrix and Fig. 3.6(b) shows per-class evaluation matrices for the proposed method I. The overall accuracy of the proposed model is evaluated as 84.13 %. The Precision, Recall and F₁-Score also calculated approximately as 84 %. There are 38,569 trainable parameters in the proposed method,

3.3.3 Model visualization

Fig. 3.8 shows an input epoch reshaped to (6,3000) corresponding to a deep sleep stage. The height of 2D input corresponds to 6 channels of 4 EEG channels and 2 EOG channels and the length corresponds to time samples. As described earlier each channel has 3000 data points per epoch. Several experiment were conducted to see the output of convolutional filter activations. Fig. 3.9 shows the output of the max pooling layer corresponds to layer 6, the feature maps highlight two things. Similarly, Fig. 3.9 the corresponding max-pool output from the right side which carries 32 feature maps. Fig. 3.11 shows the output of a layer 7 convolutional layer. As can be seen in Fig. 3.11 the feature map extract more complex features. The darker segments

are closer to zero and lighter segments show values close to 1. The values of each feature map is scaled into 0 - 1 range.

3.4 Experiment II

In Experiment I, 5-stage sleep classification was considered. However, temporal inter connections of the features were not effectively utilized to classify the five sleep stages. The purpose of this approach was to implement a four-stage classification. Mainly, the model is constructed so that it can extract features without utilizing hand engineered features. Subsequently, a Recurrent Neural Network (RNN) is used to learn temporal sequence information of sleep epochs. In this experiment one base model was developed using six PSG channels. The base model was then adopted for single channel configurations yielding multiple models. The architecture of experiment II is embedded with two main parts inspired by the work done by A. Supartak *et al.*[116]. As mentioned above, the first segment is designed so that the model can extract discriminative features from the raw PSG input. As illustrated in Fig. 3.12 a set of convolutional branches are dedicated to pull out selective time-invariant features from the raw PSG signals. The second block is attached to the end of the feature extracting block to learn the sequential trends of the PSG signals in an epoch. The whole CNN architecture is designed for 30 s PSG epoch similar to the experiment I.

3.4.1 Feature Learning

As mentioned above, the first block of the CNN model is designed for extracting different wave patterns of PSG signals using convolutional operations. In this model, each single branch was designed so that it can perform convolutional operation with different sizes of 1D convolutional kernels. Because we were interested in capturing maximum possible the obligatory features correlated with sleep stage classification, we considered to utilize few CNN branches. After a series of experiments, we decided to use five CNN blocks such a way that the kernel sizes are varied gradually from smaller to larger as shown in Fig. 3.12. Fundamentally, filters with smaller kernel sizes are likely to be working well in isolating highly localized features in the

PSG waves. On the other hand, the filters with larger kernel sizes are sufficiently performs well in segregating highly globalized features. In other words, smaller kernels can pull out the features associated with abrupt changes, and filters with larger kernel sizes are good at producing features associated with the basic patterns of the signal. As shown in the Fig. 3.12, the input is formed separately from 6 PSG channels including 4 EEG channels and 2 EOG channels like the previous experiments. Each 1D convolutional layer is followed by a rectified linear (ReLU) activation layer. The parameters of each layer can be found in Fig. 3.12. Since we are interested in 6 PSG channels in this study, the input can be identified as a combination of six 1D arrays. Formally the raw input can be represented as $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$, where N is the number of training epochs. All inputs are then concatenated as illustrated in Fig. 3.14. Each time step in the concatenated input carries six features representing six PSG channels explained above.

$$\tilde{x}_i = x_i^{EEG^{C3}} || x_i^{EEG^{C4}} || x_i^{EEG^{O1}} || x_i^{EEG^{O2}} || x_i^{EOG^L} || x_i^{EOG^R} \quad (3.7)$$

$$b_i^k = CNN_{\mathcal{D}_k}(\tilde{x}_i), \{k = 1, 2, \dots, 5\} \quad (3.8)$$

$$G_i = b_i^1 || b_i^2 || b_i^3 || b_i^4 || b_i^5 \quad (3.9)$$

Where, $||$ is the concatenation operation; \tilde{x}_i is the i^{th} input (after concatenation operation); $CNN_{\mathcal{D}_k}$ is the CNN branch parameterized by \mathcal{D}_k ; b_i^k is the set of features extracted from k^{th} CNN branch for i^{th} sample; and G_i is the combined feature sequence. After conducting several experiments (pre experiments), the kernel sizes were picked up as fractions of sampling rate f_s . Combined features G_i are then passed to the next layer.

3.4.2 Sequence Learning

In the sequence learning section, we utilized Gated Recurrent Units (GRU) to learn sequential trends of extracted features from the first section of the model. According to the AASM manual, sequence trends of the PSG wave patterns are considered in determining the corresponding sleep stage. The sleep scoring is made depending on how long wave pattern appeared in the epoch. Besides that, the transition sequence of the wave patterns such as K-complexes and sleep spindle are also thoroughly examined before labeling the epoch. For instance, one or more trains of sleep spindles can be observed in a light sleep epoch with low amplitude and mixed frequency activities. Such kind of sequential trends and temporal information are expected to be learned with stacked GRU units. Basically, two stacked GRU units were employed as the sequence learning part. Considering the feature set as (G_i) , the sequence learning part can be described as follows. Considering the G_i as a sequence of features with M time steps, G_i can be redefined as G_i^T so that,

$$G_i^T = \{g_i^{t_1}, g_i^{t_2}, \dots, g_i^{t_m}, \dots, g_i^{t_M}\} \quad (3.10)$$

$$a_{t_m}^{1,i} = GRU_{\alpha_1}(a_{t_{m-1}}^1, g_i^{t_m}) \quad (3.11)$$

$$a_{t_m}^{2,i} = GRU_{\alpha_2}(a_{t_{m-1}}^2, a_{t_m}^{1,i}) \quad (3.12)$$

Where, $g_i^{t_m}$ is an extracted feature vector of t_m^{th} time step for i^{th} training example; $a_{t_m}^{1,i}$ and $a_{t_m}^{2,i}$ are the recurrent outputs of the RNN layers respectively. GRU_{α_1} and GRU_{α_2} layers are parameterized by α_1 and α_2 .

3.4.3 Complete Model

After series of experiments we fixed the Kernel sizes of the first layers of the CNN branches, starting from $\frac{1}{10} \times f_s$. Kernel sizes of the convolutional layers are $\frac{10f_s}{100}$, $\frac{8f_s}{100}$, $\frac{4f_s}{100}$, $\frac{3f_s}{100}$, and $\frac{2f_s}{100}$

respectively.

The stride sizes of the first two convolutional branches, were set to $\frac{1}{25}^{th}$ of sampling rate, while the other three were set to $\frac{1}{50}^{th}$ of sampling rate. The kernel sizes and strides of subsequent convolutional layers were kept at smaller values. The feature extracting section and the sequence learning section is combined via two convolutional layers. Finally, fully connected layers followed by a SoftMax layer is employed to classify four sleep stages.

3.4.4 Multi-step training

Multi-channel models

The training was done in a few steps. Firstly, the feature extraction CNN branches were trained. In the first training session only the feature extraction section was trained with a SoftMax layer attached directly to the concatenation layer. Categorical cross-entropy loss was used as the loss function to evaluate the error between model predictions and the actual sleep stages. Adam optimizer was used to perform the supervised pre-training session, with a learning rate of ($lr_1 = 0.001$). Since our data set is large, we set the mini-batch size as 50 for each training iteration. Subsequently two fine-tuning training sessions were performed to train the rest of the model.

After determining parameter sets $\mathcal{D}_k, \{k = 1, 2, \dots, 5\}$ for all branches of $CNN_{\mathcal{D}_k}$, SoftMax layer was removed from the pre-trained section. The sequence learning part was attached to the pre-trained part. The sequence learning part consists of two convolutional layers and a max pooling layer prior to the stacked GRU units. The output of the sequence learning portion was attached to a fully connected neural network followed by a SoftMax layer as illustrated in Fig. 3.13. In the second training session, the pre-trained part was disabled for training, and only the newly attached part was enabled for training.

Since the prior layers have been already trained for extracting features, a lower learning rate was employed ($lr_2 = 0.0007$) during this training session. The best model was continuously saved during training if the validation accuracy is increased against the training iteration. In

order to save the best model, we used early stopping criteria in Keras library. In the final training session, all the layers were frozen except the fully connected layers and the model was retrained with a learning rate of $lr_3 = 0.0007$.

Single-channel models

We performed some additional experiments to verify the feasibility of using this model with a single EEG channel. In this experiment, we assigned only one EEG channel for all inputs at a time and retained the model with the same data set. In all single channel experiments, training was disabled for all the feature extraction block and the whole model was then retrained using Adam optimizer with a learning rate of 0.0009. After performing the same experiment for the four EEG channels, we yielded 4 single channel configurations for each data set.

3.4.5 Implementation

Similar to the Experiment I, the second CNN was also implemented and trained from ground up using an Python environment using Keras deep learning API library with TensorFlow as back-end. Furthermore, this model was trained with GPU support (NVIDIA GEFORCE GTX 1070). As mention earlier we used two data sets. But in in all experiment scenarios, we used same configuration to retrain the model for healthy data set (*mod_D1*) and patient data set (*mod_D2*). And, the same configuration was used to train and test the single-channel models. Since the patient data set is comparatively large, we used $\sim 6\%$ of data from 172,512 total epochs to test the trained model for very experiments. On the other hand $\sim 12\%$ of 59,480 total epochs from the healthy data set was used to test the models. The validation split ratio was kept same for all experiments during all experiments. During each pre-training session, the best model was saved based on the validation accuracy.

3.4.6 Results

Multi-channel models

Fig. 3.14 shows a raw PSG input epoch corresponding to a deep sleep epoch. As described above, each channel has 3000 data samples. The visualizations of few convolutional filter activations corresponding to the epoch illustrated in Fig. 3.14 is shown in Fig. 3.15. As can be seen in Fig. 3.15, the feature maps highlight two things. Firstly, the most influential features have been extracted from the raw PSG signal. Secondly, trivial information has been cut off from the feature map as a result of the ReLU activation function. The visualized feature map in Fig. 3.15(a), which corresponds to the first convolutional layer of the 4th branch, illustrates how the convolutional filters have been adapted to capture the generic features from the raw signal. As shown in Fig. 3.15(d) in the final layers, most of the unnecessary features have been removed from the feature map, while the pertinent information for deep sleep stage have been emerged.

In this phase, we evaluated trained models with an unseen data set from both data sets. As shown in Fig. 3.16, (*mod_D1*) performed well in classifying light sleep stage with a confidence of 96 % . Besides that, wake, deep, and REM showed 85 %, 80 %, and 76 % respectively. The overall accuracy is recorded as 89.30 % for healthy data set. On the other hand, (*mod_D2*) also having an accuracy of 95 % for light sleep stage, while the REM, wake, and deep stages showed confidence levels of 88 %, 82 %, and 74 % respectively. Receiver operating characteristic (ROC) curves for the multi-channels models are shown in Fig. 3.17. The trained CNN models showed macro-average Area Under Curve (AUC) of 0.99 for both data sets. The macro-average AUC for healthy subjects and patients was calculated as 0.98 and 0.99 respectively. ROC AUCs for wake, light, deep, and REM sleep stages were calculated as 0.99, 0.96, 0.99, and 0.98 respectively for healthy data set. The similar measures for patient's data were 0.99, 0.98, 0.99, and 0.99 for wake, light, deep, and REM stages respectively. Table 3.2 tabulates the per-class evaluation metrics for both multi models. W, L, D, and R represents the wake, light, deep, and REM sleep respectively. In terms of precision, recall, and F1 measures, *mod_D1* has shown a weighted average of 89 %, while *mod_D2* is observed with 92 %. It is interesting to note that the micro

averages of above measures are observed unchanged even though the per-class values lie a range between 74 % ~ 95 %.

Table 3.2: per-class evaluation metrics for multi-channel models

	<i>precision</i>		<i>recall</i>		<i>f1-score</i>		<i>Actu. occurrences</i>	
	<i>mod-D1</i>	<i>mod-D2</i>	<i>mod-D1</i>	<i>mod-D2</i>	<i>mod-D1</i>	<i>mod-D2</i>	<i>mod-D1</i>	<i>mod-D2</i>
W	0.95	0.88	0.85	0.82	0.9	0.85	855	1214
L	0.89	0.95	0.96	0.95	0.92	0.95	4202	6698
D	0.88	0.74	0.8	0.74	0.84	0.74	758	189
R	0.88	0.84	0.76	0.88	0.81	0.86	1182	1500
μ . Avg	0.89	0.92	0.89	0.92	0.89	0.92	6997	9601

Single-channel models

Fig. 3.18 shows the confusion matrix for single-channel models of healthy subjects obtained by replacing the input with the same EEG channel. In a nutshell, we can observe that EEG-O1 model shows the maximum predicting confidence in predicting deep sleep (88 %) and wake stage (80 %) while EEG-O2 and EEG-C4 models are having the highest confidence level in predicting light sleep stage (94 %) and REM sleep stage (78 %) respectively. * Note- bold numbers indicate the maximum value of each column.

Similarly, Fig. 3.19 shows the corresponding confusion matrix for patients. Table 3.3 and Table 3.4 show the corresponding per-class evaluation metrics for both data sets. As noted in Table 3.3 micro averages of precision, recall and F1 measures for all single channel models lie between 82 % ~ 86 %. Correspondingly, Table 3.4 exhibits a range of (86 % ~ 88 %) for the models trained with patients for similar measures.

3.4.7 Overall Performances of Proposed Models

In Table 3.5, the overall performances of multi-channel models and single-channel models are shown. As can be seen, the best overall performance is observed with the multi-channel model for both data sets.

In contrast, the best overall performance (85.29 %) for single-channel is given by the model trained for EEG C4-A1 for the healthy data set while the EEG C3-A2 model shows the best

Table 3.3: per-class evaluation metrics for single-channel models trained for healthy data set

	<i>precision</i>				<i>recall</i>				<i>f1-score</i>			
	O1	O2	C3	C4	O1	O2	C3	C4	O1	O2	C3	C4
W	0.88	0.94	0.8	0.76	0.8	0.73	0.72	0.76	0.84	0.82	0.75	0.76
L	0.86	0.83	0.87	0.92	0.89	0.94	0.93	0.9	0.87	0.88	0.9	0.91
D	0.78	0.87	0.88	0.84	0.88	0.63	0.78	0.86	0.83	0.73	0.83	0.85
R	0.68	0.73	0.75	0.73	0.59	0.64	0.68	0.78	0.63	0.68	0.71	0.75
μ . Avg	0.82	0.83	0.84	0.86	0.83	0.83	0.85	0.86	0.82	0.82	0.84	0.86

Table 3.4: per-class evaluation metrics for single-channel models trained for patient data set

	<i>precision</i>				<i>recall</i>				<i>f1-score</i>			
	O1	O2	C3	C4	O1	O2	C3	C4	O1	O2	C3	C4
W	0.85	0.84	0.8	0.84	0.76	0.74	0.72	0.74	0.8	0.79	0.75	0.79
L	0.89	0.91	0.87	0.91	0.94	0.93	0.93	0.93	0.91	0.92	0.9	0.92
D	0.86	0.94	0.88	0.94	0.5	0.32	0.78	0.32	0.63	0.47	0.83	0.47
R	0.72	0.74	0.75	0.74	0.62	0.78	0.68	0.78	0.67	0.76	0.71	0.76
μ . Avg	0.86	0.86	0.88	0.87	0.86	0.86	0.88	0.87	0.86	0.86	0.88	0.87

overall test accuracy (87.7 %) for patient data set.

Table 3.5: Overall performances of proposed models

Channel	<i>Healthy data set</i>					<i>Patient data set</i>				
	all	O1	O2	C3	C4	all	O1	O2	C3	C4
Test Acc. %	89.3	82.5	82.1	84.6	85.9	91.9	86.1	85.9	87.7	87.4

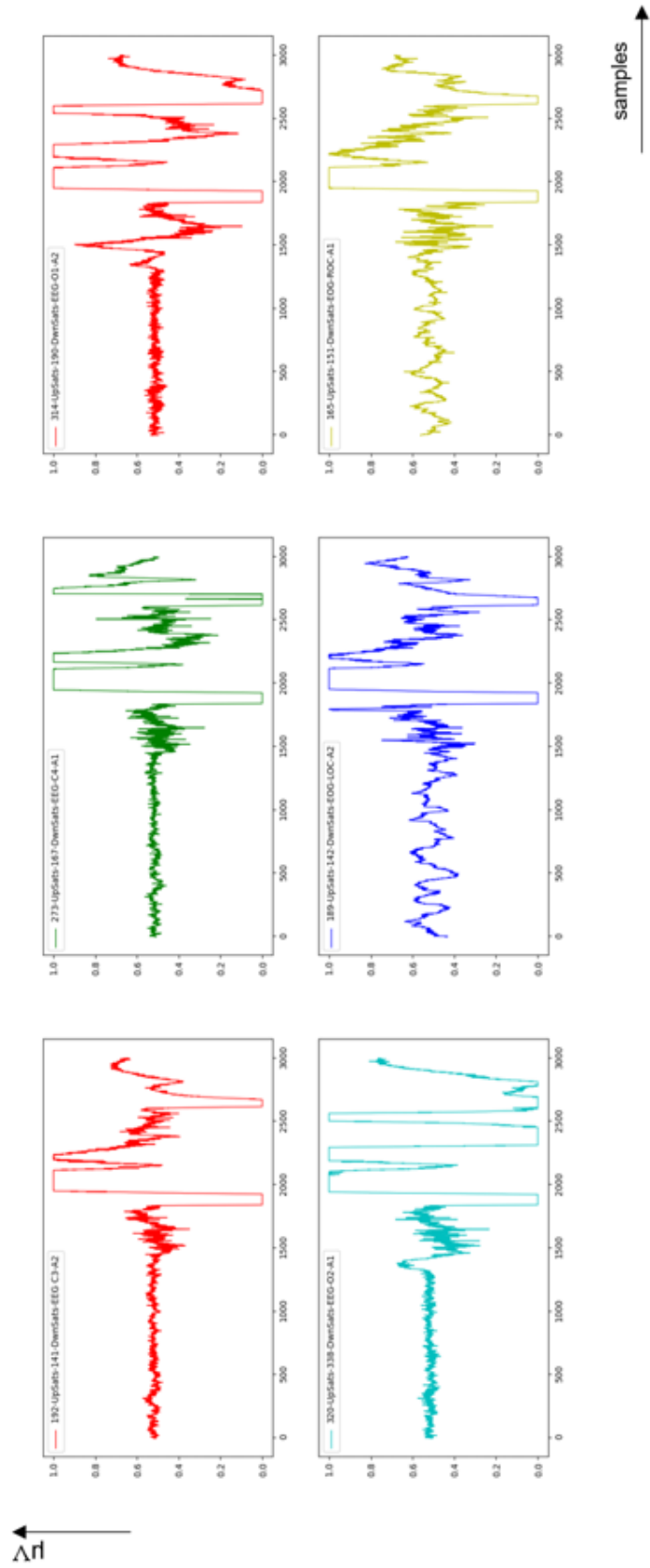


Figure 3.1: Epoch with artifacts due to the moment of the electrodes

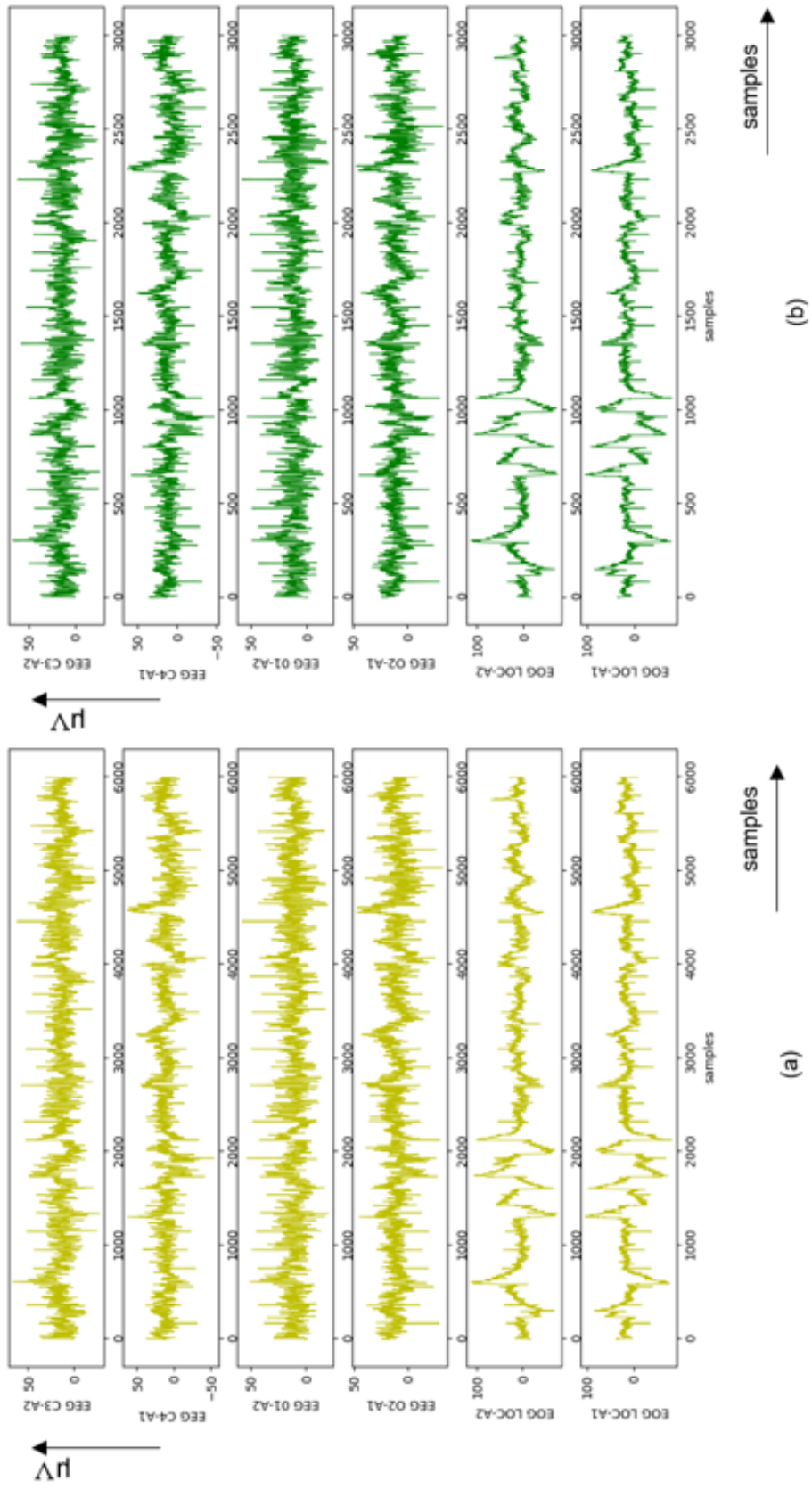


Figure 3.2: Normal epoch (a) original signal (b) decimated signal



Figure 3.3: Pre-processing and rearranging data

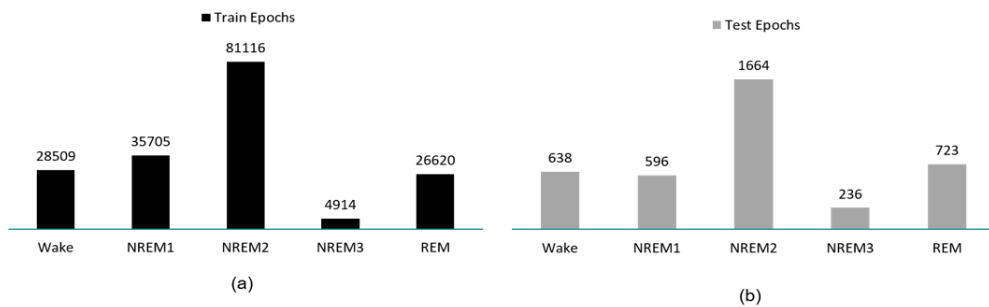


Figure 3.4: Training and test data

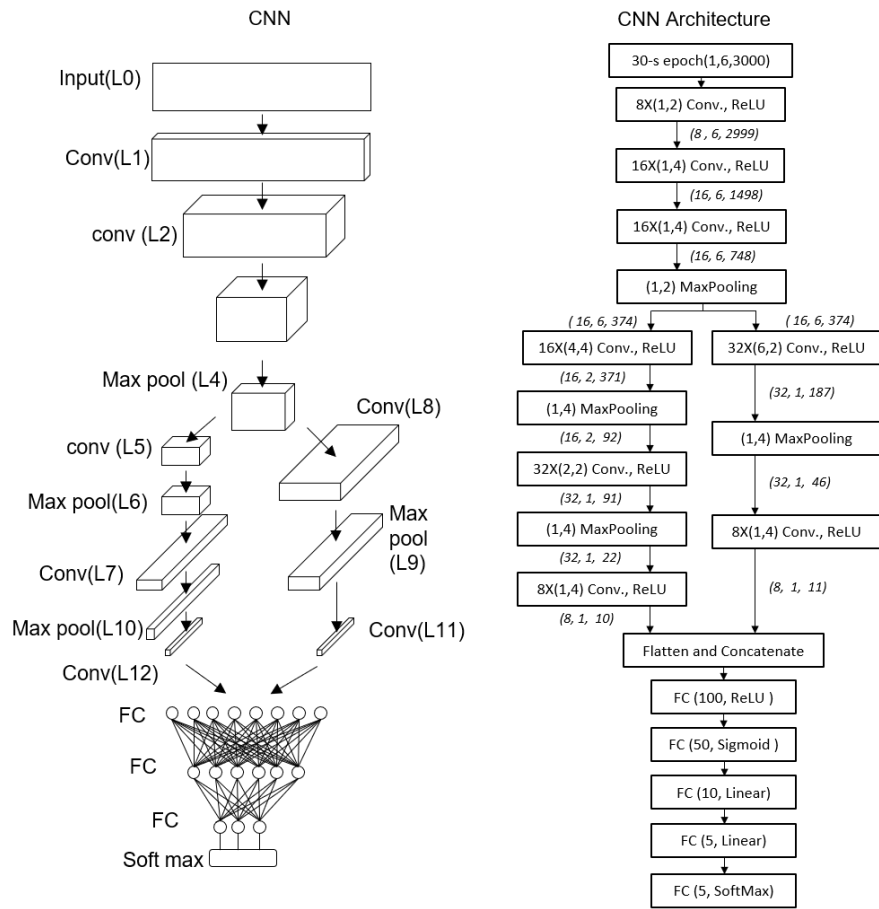


Figure 3.5: Architecture of proposed method I

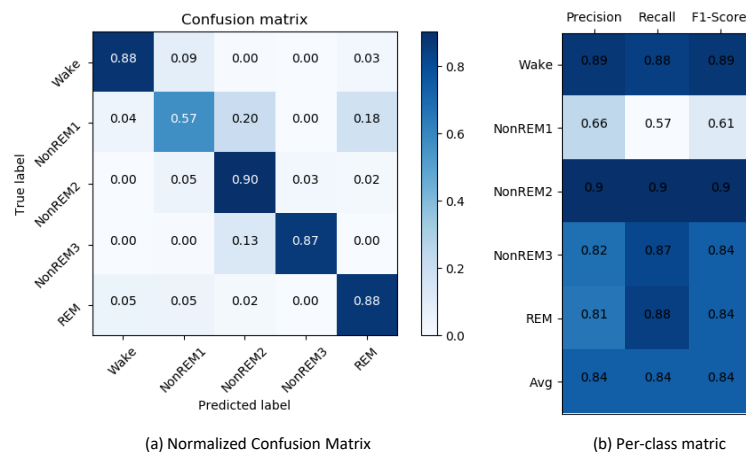


Figure 3.6: Evaluation matrices (a) Normalized Confutation Matrix (b) Per-Class Metric

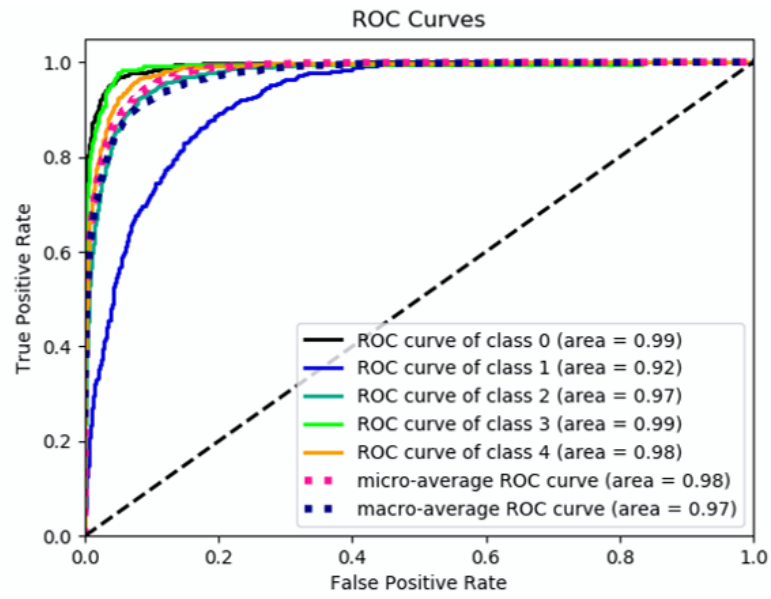


Figure 3.7: Receiver operating curve (ROC) for proposed Method I

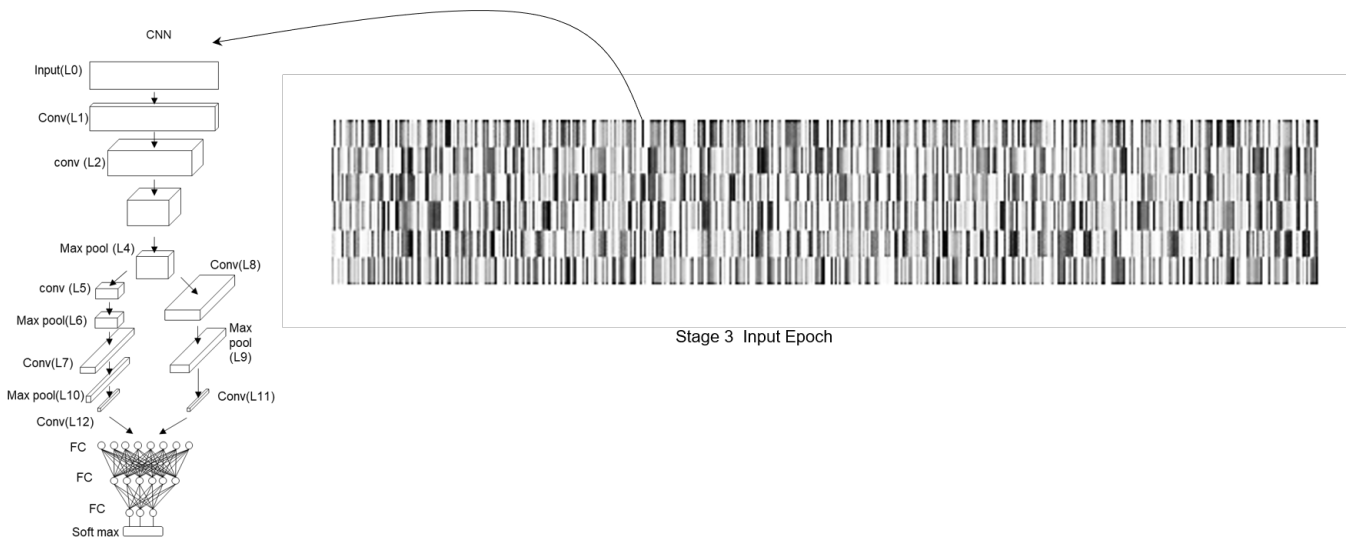


Figure 3.8: Reshaped input corresponds to a deep sleep epoch

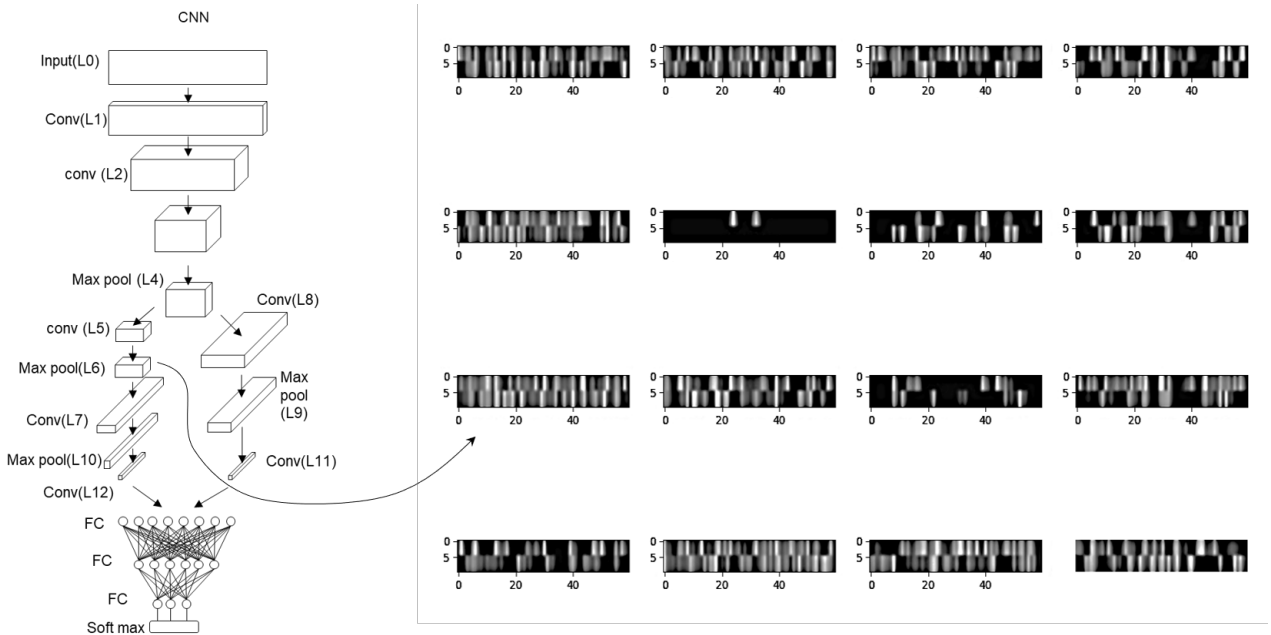


Figure 3.9: Feature map corresponds to the max pool layer (L6) of left branch

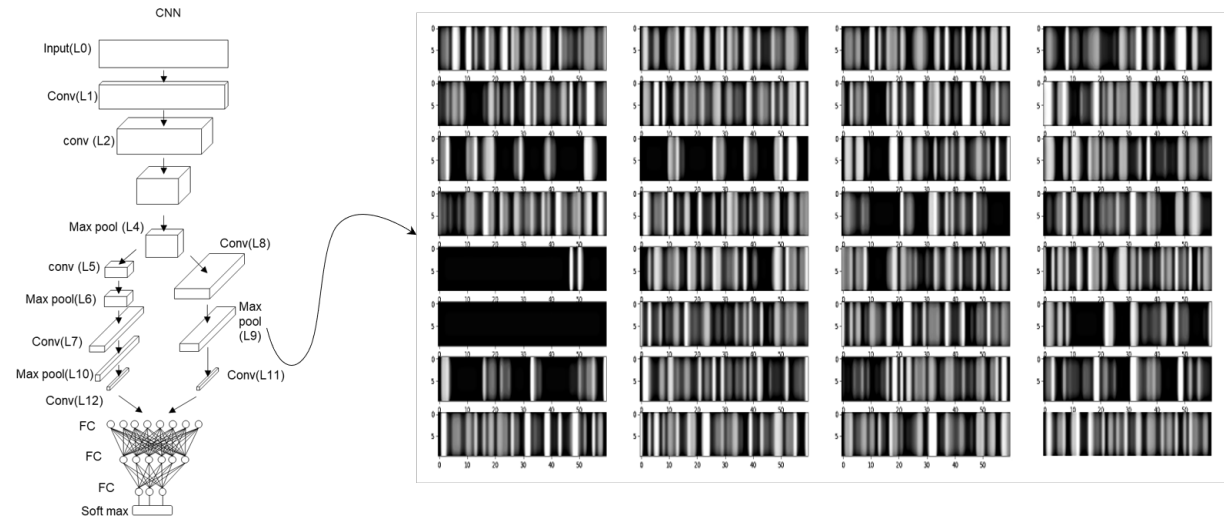


Figure 3.10: Feature map corresponds to the max pool layer (L9) of right branch

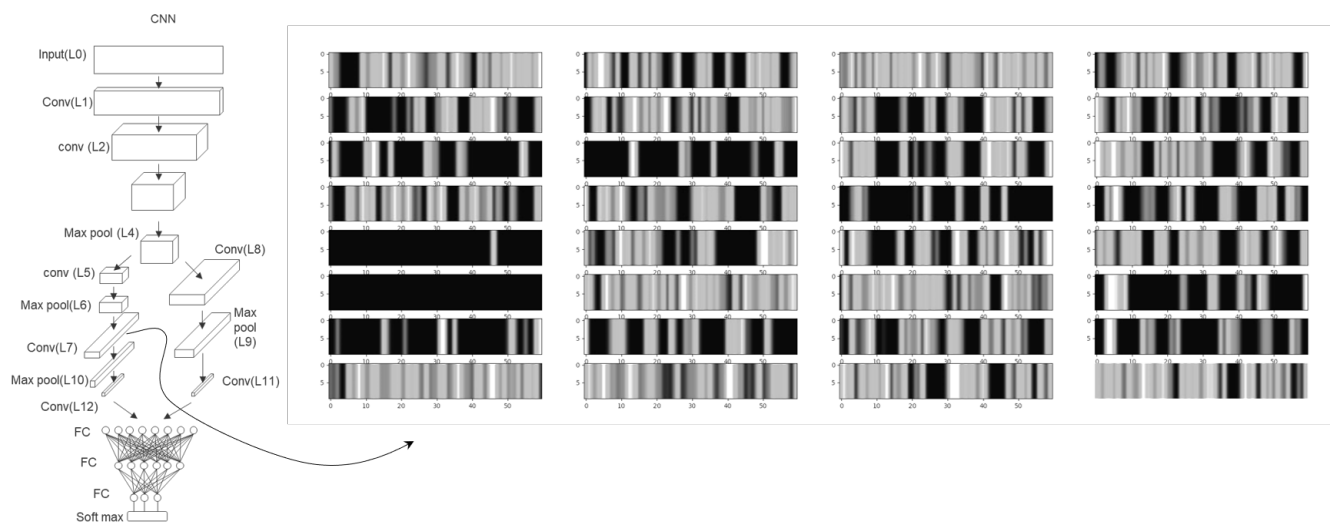


Figure 3.11: Feature map corresponds to the convolutional layer (L7) of left branch

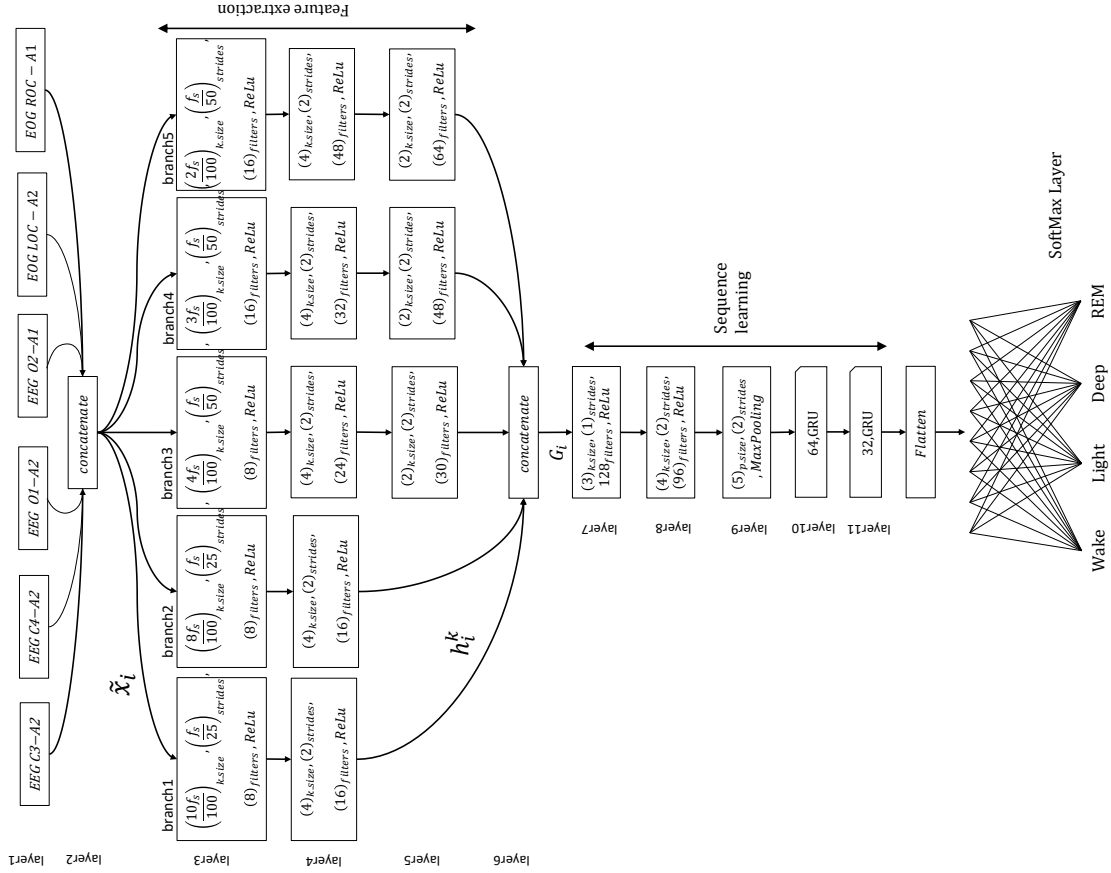


Figure 3.12: Proposed Multi-channel model, f_s is the frequency

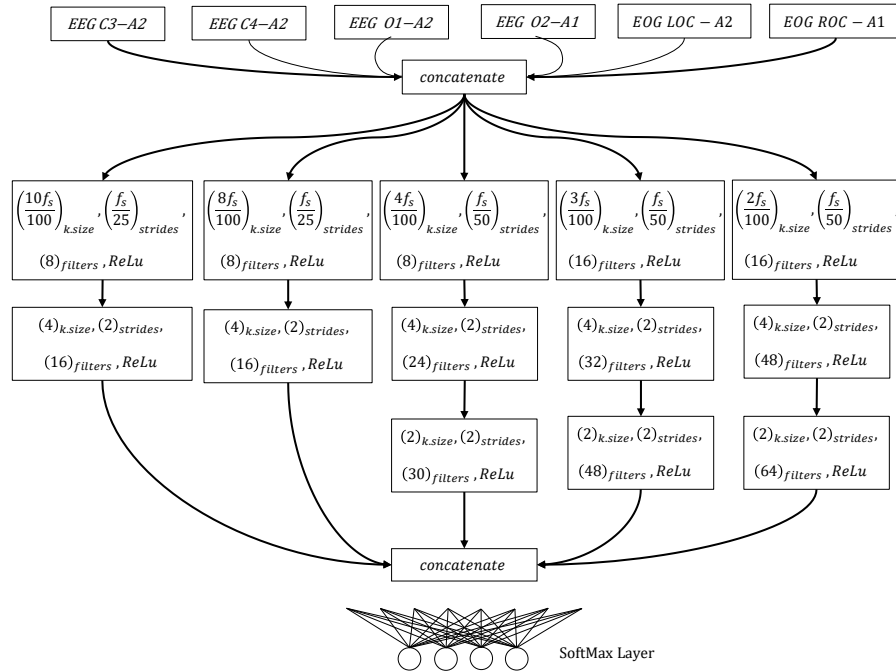


Figure 3.13: Feature extraction section

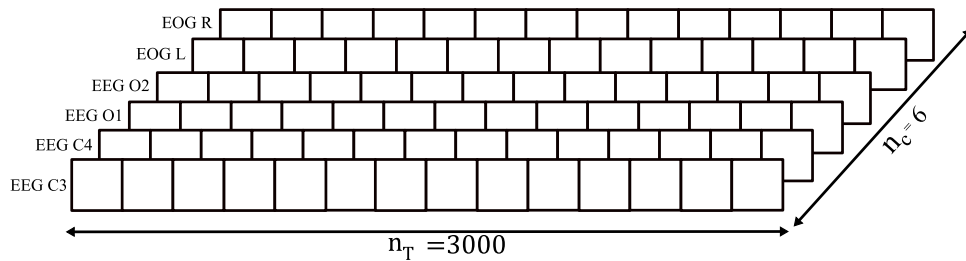


Figure 3.14: Input of 1-D time series array in (1).

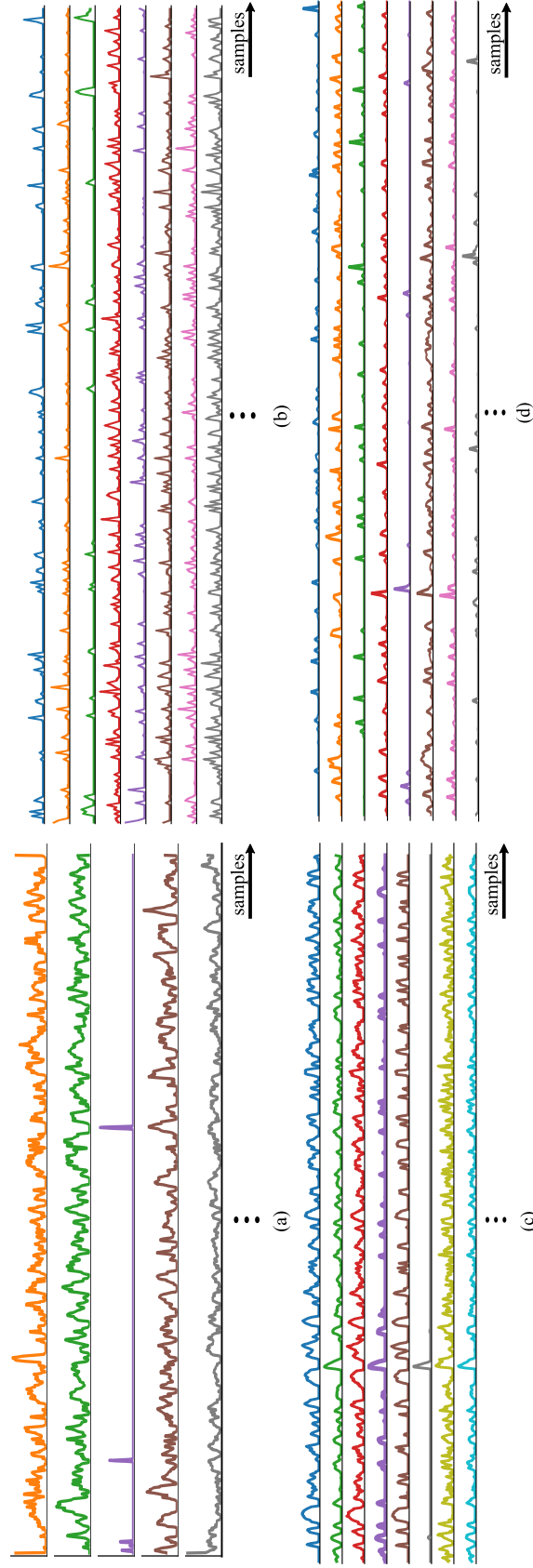


Figure 3.15: Visualizations of convolution layers of trained multi-channel model corresponding to a deep sleep epoch (a). activations of the first 5 filters (layer-3, branch-4) (b). activations of the first 8 filters (layer-4, branch-4) (c). activations of the first 8 filters (7th layer)

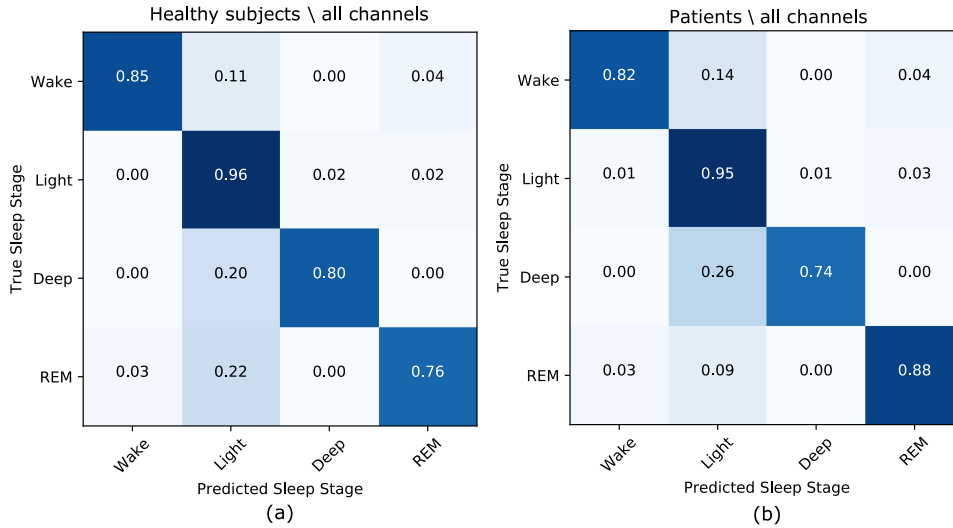


Figure 3.16: Normalized confusion matrix for multi-channel models. (a). *mod_D1* (model for healthy subjects); (b). *mod_D2* (model for patients).

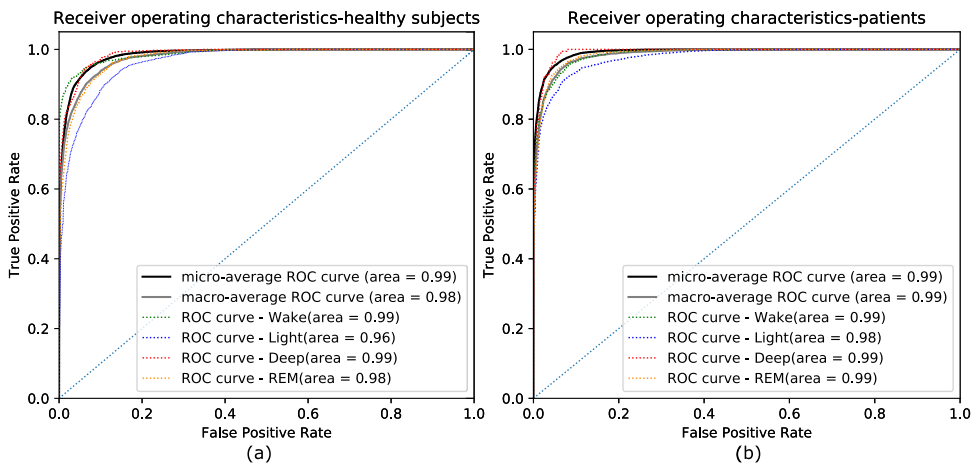


Figure 3.17: Receiver operating characteristic curves for multi-channel models (a). *mod_D1* (model for healthy subjects) (b). *mod_D2* (model for patients)

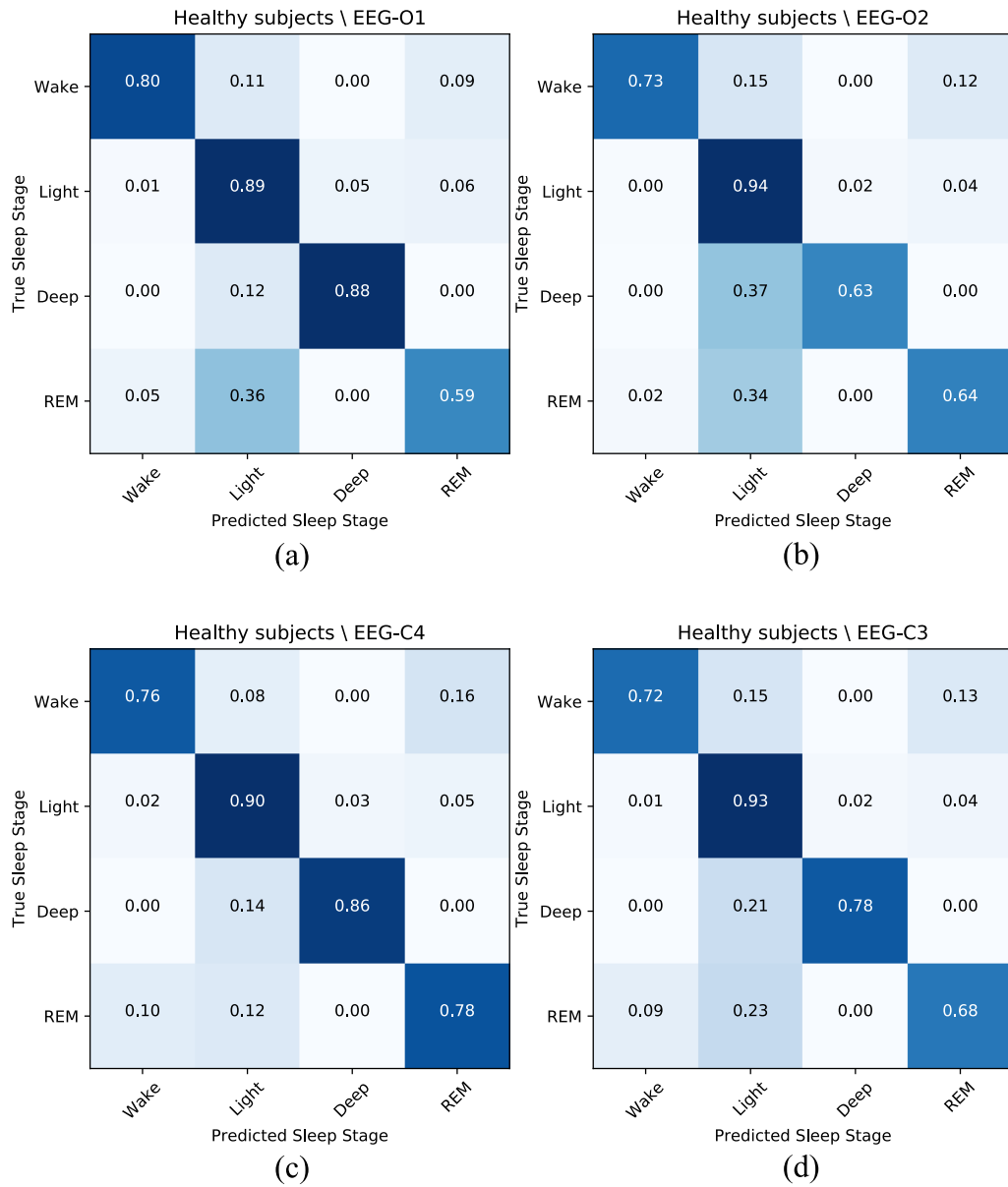


Figure 3.18: Normalized confusion matrix for single-channel models for healthy data set (a). EEG-O1 model (b). EEG-O2 model (c). EEG-C4 model (d). EEG-C3 model.

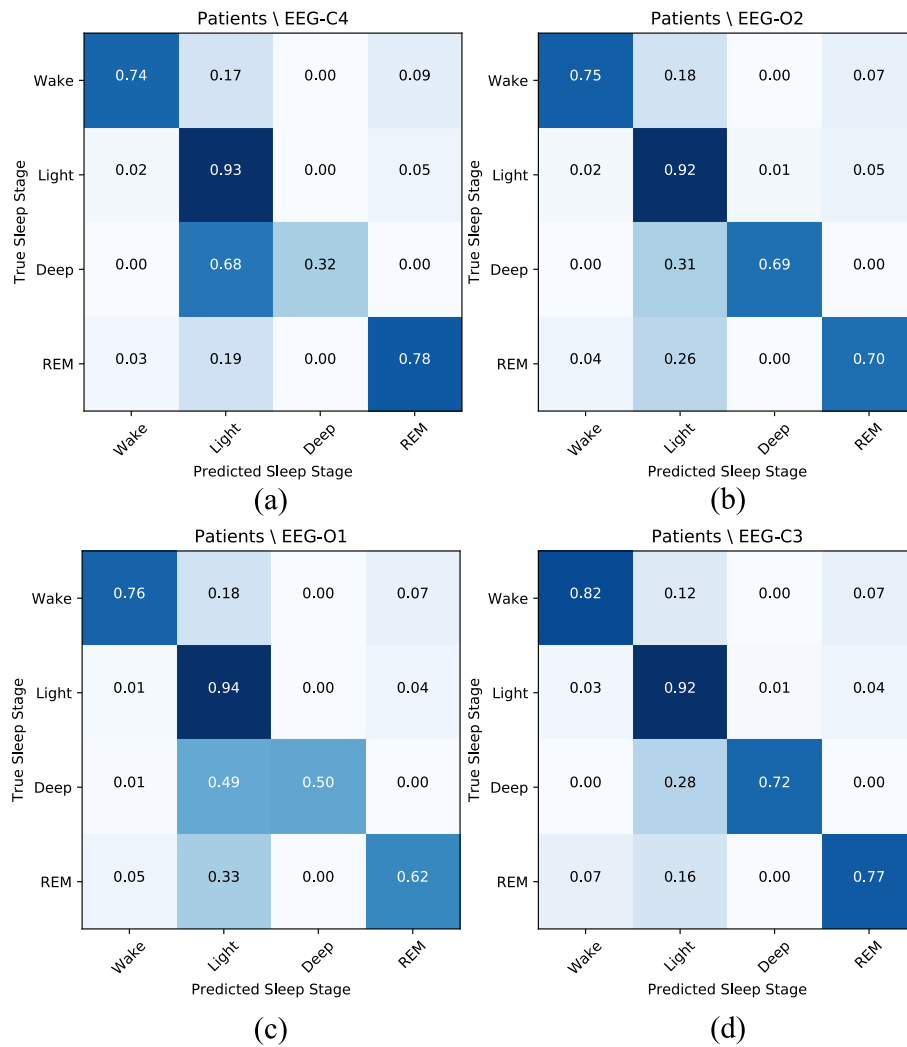


Figure 3.19: Normalized confusion matrix for single-channel models for Patient data set (a). EEG-C4 model (b). EEG-O2 model (c). EEG-O1 model (d). EEG-C3 model.

Table 3.6: Result comparison

Approach	Sleep stages	Automatic feature selection	Classifier/Method	Subject/data set/channels	Performance, Acc. %
1 Gudmundsson <i>et al.</i> 2005[117]	W, S1 + S2, SWS, REM	No	SVM, KNN	4 recordings, EEG (C3-A2)	81
2 Herrera <i>et al.</i> 2011[118]	W, S1/S2, S3/S4, REM	No	SVM	10 recordings, EEG (C3-M2)	70
3 Li <i>et al.</i> 2009[119]	W, S1 + REM, S2, S3/4	No	KNN	8 recordings, Sleep-EDF dataset	81.7
4 Huanga, <i>et al.</i> 2013[37]	W, S1/S2, S1/S4, REM	No	SVM	EEG(Fpz-Cz, Oz) 10 recording, Sleep laboratory of NCTU dataset EEG (Fp1, FP2)	77.1
5 Hassan <i>et al.</i> 2015[120]	W, S1-S2, SWS, REM	No	Bootstrap Aggregating (Bagging) ANN	8 recording, Sleep-EDF database EEG (Fz-Oz)	87.49
6 Ebrahimi <i>et al.</i> 2008[121]	W, S1 + REM, S2, SWS	No	ANN	8 recordings, Sleep-EDF dataset, EEG (Pz-Oz)	93.0
7 Phan <i>et al.</i> 2013[122]	W, S1 + REM, S2, SWS	No	KNN	4 recordings, Sleep-EDF EEG (Fpz-Cz)	94.49
8 Prop. Method	W, N1+N2, N3, REM	Yes	CNN, RNN	184 Patient subjects, 70 Healthy subjects	data set1: Multi-channel 89.3, fl-89 Single-channel 85.9, fl-86 data set2: Multi-channel 91.9, fl-92 Single channel Acc 87.7, fl-88

3.5 Experiment III

In Experiment II, we consider 4 stage sleep classification. In order to improve the performances of the models proposed in Experiment I and Experiment II, we designed series of experiments in Experiment III. In this chapter, we present the architecture of another automatic sleep stage classification model with improved performances. This model is capable of performing 5 stage and 4 stage classification at once. This experiment is an extension of Experiment II, and most of the experimental setups are identical to the procedures followed in Experiment II.

We detail the whole process of multi-channel CNN-GRU architecture, in this section. The architecture is inspired from an existing study [39] and an extension of the Experiment II model. Even though the Experiment II model consisted of five branches, we noticed that some branches does not contribute adequately for the classification process. Therefore the proposed deep learning architecture in Experiment III consists of only three main sections as shown in Fig. 3.20. As similar to the Experiment II the first block is dedicated to learning sleep related features from raw PSG inputs. The second part consists of stacked bidirectional and unidirectional gated recurrent units (GRU) to learn the interconnections among extracted features within an epoch (sequential patterns of extracted features in an epoch). In addition to the main blocks described above, bypass CNN blocks with global pooling operations and convolutional layers with small kernels were employed to extract fine details. As shown in the Fig. 3.20 the final section is divided into two sections of fully connected layer to simultaneously perform 5-stage and 4-stage classifications.

Similar to the previous experiments, we utilized 6 PSG channels including 4 EEG (C3-A2, C4-A1, O1-A2, O2-A1) and 2 EOG (LOC-A2, ROC-A1) to train the newly proposed network. The raw input can be represented as $X = \{x_1, x_2, \dots, x_i, \dots, x_M\}$, where M denotes the number of training samples, and each x_i consists of time series data from different channels mentioned above. We denote 30 s of PSG signal segment by $x_i \in R_{n_c * n_T}$ with its label $y_i \in Y$, where n_c denotes the number of channels, n_T denotes time sampling points, and $Y = \{W, N1, N2, N3, REM\}$. Subsequently, the proposed CNN-GRU classification

model is defined by a function $\hat{f} : x_i \rightarrow y_i$.

3.5.1 Feature learning

Each PSG channel consists of 3,000 time samples $n_T = 3000$. x_i , ($i = 1$ to M) corresponds to raw time series data from 6 PSG channels shown in Fig. 3.14. The first block (feature learning) consists of three 1-D CNN branches attached to the concatenation as illustrated in Fig. 3.20. The kernel size of the first convolutional layer in each CNN branch is selected as defined in Table 3.7. The parameters were determined experimentally via trial and error.

The model starts with 6 inputs (each corresponds to a PSG channel), followed by a concatenation layer. The concatenation layer concatenates all input arrays into a \tilde{x}_i time series array as described in equation (3.13). Subsequently, the reshaped 1-D time series array (3000, 6) is passed to CNN branches termed as $\text{CNN}_{\mathcal{G}_1}$, $\text{CNN}_{\mathcal{G}_2}$, and $\text{CNN}_{\mathcal{G}_3}$, where $\text{CNN}_{\mathcal{G}}$ denotes a function that transforms raw PSG segment into a feature sequence parameterized by \mathcal{G} .

$\text{CNN}_{\mathcal{G}_1}$ starts with fairly larger convolutional kernels and $\text{CNN}_{\mathcal{G}_2}$ starts with medium sized convolution filters to perform different convolutional operations. $\text{CNN}_{\mathcal{G}_3}$ exhibits the smallest starting kernel size. The outputs of each CNN layers from each branch are then further reduced via sub subsequent convolutional layers and pooling layers. Three feature maps are created at the end of each convolutional branch. The concept behind the use of multiple CNN branches is to construct a sequence of features, which can discriminate sleep stages from each other. In other words, assorted kernels helps to handle the trade-off between temporal and frequency information.

$$\tilde{x}_i = x_i^{EEGC^3} || x_i^{EEGC^4} || x_i^{EEGO^1} || x_i^{EEGO^2} || x_i^{EOGL} || x_i^{EOGR} \quad (3.13)$$

$$b_i^k = \text{CNN}_{\mathcal{G}_k}(\tilde{x}_i), \{k = 1, 2, 3\} \quad (3.14)$$

$$G_i = b_i^1 || b_i^2 || b_i^3 \quad (3.15)$$

where, $||$ denotes the concatenation operation; \tilde{x}_i denotes the i^{th} input (after concatenation operation); $\text{CNN}_{\mathcal{D}_k}$ denotes the CNN branch parameterized by \mathcal{D}_k ; b_i^k denotes the set of features extracted from k^{th} CNN branch for i^{th} sample; and G_i denotes the combined feature sequence.

As explained in previous experiments, 1-D CNN filters with smaller kernels likely to be distinguishing signal patterns with abrupt changes while convolutional filters with larger kernels are good at identifying frequency information [39]. A combination of the three branches can obtain a detailed time–frequency representation at the second concatenation layer. All branches are designed such that the output dimension of each branch is identical. In all the CNN blocks, the convolution operation starts by convolving the input signal with a pre-defined number of filters with stride as shown in Fig. 3.20. From the perspective of the kernel size, the amount of information extracted from larger convolutional filters are lower when compared that from the smaller filters. Furthermore, the smaller filter exhibits better weight sharing and slow reduction in input dimension. Based on those factors, a few number of filters and long strides are employed for larger kernels. Conversely, more filters were dedicated smaller kernels and the stride sizes were set to smaller values since the small convolutional filters collects very fine details in the interested receptive field. All the subsequent convolutional layers of the CNN branches were set as nearly identical. The number of filters is increased gradually in each consecutive layer of each branch, as shown in Fig. 3.20. In this configuration, we expected that most of the sleep related information, including k-complexes, sleep spindles, saw-tooth waves, and low amplitude mixed frequency waves are fully extracted. The feature maps formed at the end of each convolutional layer can be represented as in equation (3.15), where b_i^k denotes the reduced feature map from each CNN block.

Table 3.7: Model specifications

	Layer type	Kernal size	Number of filters	Strides	Output dimension (samples)
CNN _{g₁}	Convolutional	16	8	4	750,8
	Convolutional	3	12	2	374,12
	Convolutional	3	16	2	186,16
	Max-pooling	4		2	92,16
CNN _{g₂}	Convolutional	8	12	2	1500,12
	Convolutional	3	16	2	749,16
	Max-pooling	4		2	373,16
	Convolutional	3	24	2	186,24
	Max-pooling	4		2	92,24
CNN _{g₃}	Convolutional	4	12	2	1500,12
	Convolutional	3	16	2	749,16
	Max-pooling	4		2	373,16
	Convolutional	3	24	2	186,24
	Max-pooling	4		2	92,24

3.5.2 Sequence learning

The temporal interconnections of extracted features are expected to learn in the sequence learning section. We employed one bidirectional GRU layer and two unidirectional GRU along with a fully connected layer to reformulate the information extracted from the feature extractor for better classification performances. Based on the AASM manual, sequence trends of wave shapes are considered to determine the sleep stage. For example, a major body movement followed by a slow eye movement and low amplitude mixed frequency without non arousal associated k-complex defines the end of an N2 sleep. The Recurrent Neural Network (RNN) is used to learn the types of sequential trends of the extracted spatial information. Some studies [123][124] indicated that deep RNN architectures such as LSTM and GRU build-up progressively higher level representations of sequence data. The output of an LSTM hidden layer can be fed as the input to subsequent LSTM hidden layer to enhance the performance of the network[125].

We assume that feature sequence G_i contains N number of feature vectors $g_i^{t_n}$ after concatenation operation, and thus the feature sequence G_i can be redefined as G_i^T as follows:

$$G_i^T = \{g_i^{t_1}, g_i^{t_2}, \dots, g_i^{t_n}, \dots, g_i^{t_N}\} \quad (3.16)$$

where, $g_i^{t^n}$ denotes a concatenated feature vector that corresponds to the t^n time step at for the i^{th} example. It is important to note that the t^n denotes time step at the second concatenation layer. The concatenated feature vector sequence G_i^T is then passed to the GRU block, as shown in Fig. 3.20. The output of the feature sequence learning section \mathcal{A}_{gru}^i is defined as follows:

$$\mathcal{A}_{gru}^i = \text{GRU}_{\delta}(G_i^T) \quad (3.17)$$

where, GRU_{δ} is a function that converts the feature vector sequence to a vector using stacked GRU blocks parameterized by δ . We set the return sequence as “False” for the final GRU layer to obtain a single vector \mathcal{A}_{gru}^i for the input sequence G_i^T , which later passed to a fully-connected layer FC_{α} parameterized by α to form a vector $O_{gru_block}^i$.

$$O_{gru_block}^i = \text{FC}_{\alpha}(\mathcal{A}_{gru}^i) \quad (3.18)$$

3.5.3 Complete model

As shown in Fig. 3.20 we used extra stacked CNN block CNN_{β} (parameterized by β) containing small kernels followed by a global max pooling layer to form a feature vector directly extracted from the input. Additionally, we employed a shortcut CNN block CNN_{γ} parameterized by γ followed by an average pooling layer to map the extracted features into a vector $O_{avg_block}^i$. Finally, all sections were concatenated to form O^i shown in (3.21) and passed to the fully connected layers to perform classification. In the model designing we employed aforementioned global pooling layers to vigorously summarize the extracted features from raw input [126]. The global average pooling layer makes the model more robust to spatial translations in the data in extracted features and works as a regularizer [127].

$$O_{outer_block}^i = \text{CNN}_{\beta}(\tilde{x}_i) \quad (3.19)$$

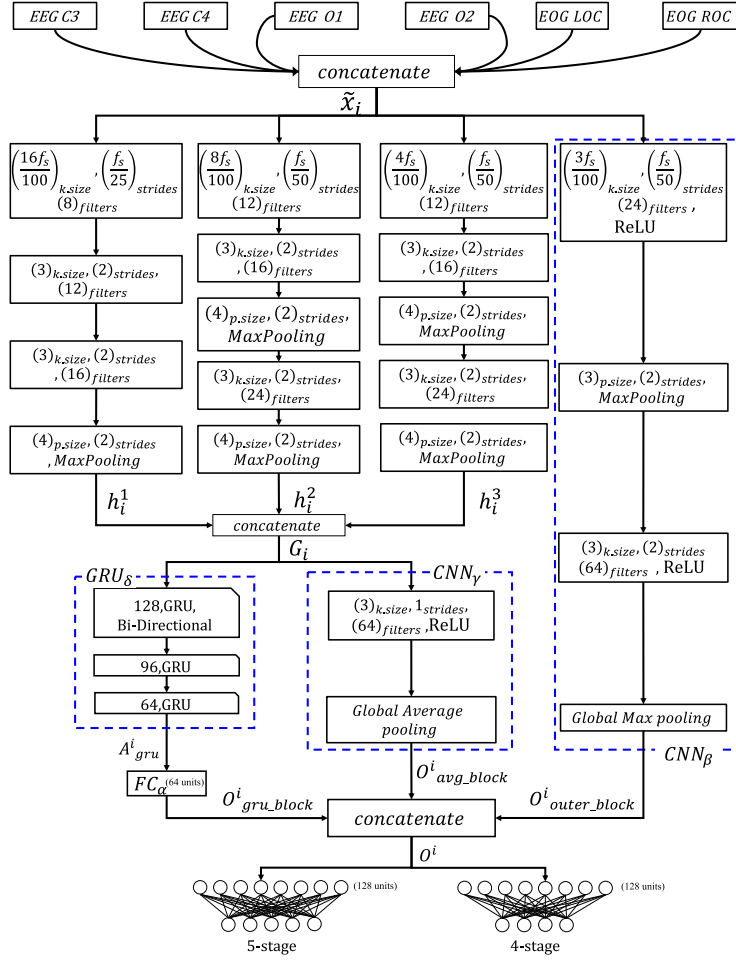


Figure 3.20: Architecture of CNN-GRU model.

$$O^i_{avg_block} = \text{CNN}_\gamma(G_i^T) \quad (3.20)$$

$$O^i = O^i_{gru_block} || O^i_{outer_block} || O^i_{avg_block} \quad (3.21)$$

In the classification section, a pair of fully connected layers are employed separately for the 4-stage case and 5-stage case. For the 4-stage and 5-stage cases, soft-max layers are utilized to obtain probability vectors corresponding to each sleep stage. For all convolutional layers in the feature extraction section, we employed linear activations, and a rectified linear unit (ReLU) activation is applied subsequently to each convolutional layer in both CNN_γ and CNN_β . In the study, raw PSG data was used to implement the model without any further preprocessing.

We randomly selected 14 subjects that were set as the test data set, and 170 subjects were chosen to train the network, as shown in Fig. 3.21.

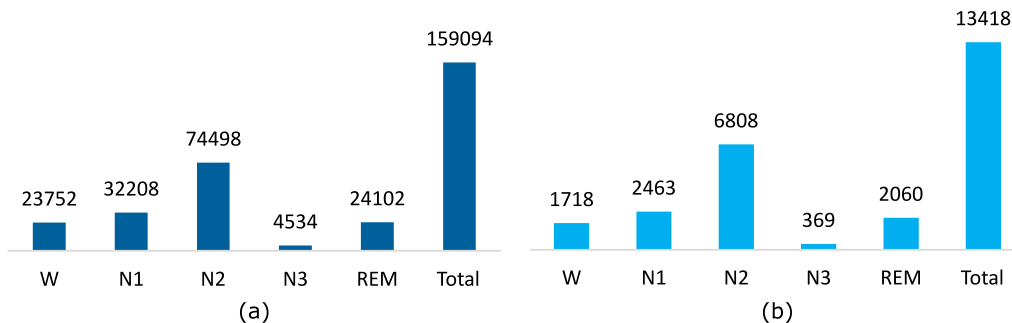


Figure 3.21: Number of epochs in each sleep stage.(a). training epochs; (b)test epochs

3.5.4 Multi-phase training

Similar to the previous experiments, we employed several training sessions to effectively train our network using back propagation. Essentially, we trained our network section by section using an Adam optimizer to minimize the loss using mini-batch gradient descent.

Firstly, we trained the feature extractors $\text{CNN}_{\mathcal{D}_k}$, $\{k = 1, 2, 3\}$. The model is trained after setting all parameters for $\text{CNN}_{\mathcal{D}_1}$ branch followed by fully connected layer with a soft-max activation. The soft-max layer consists of five units to calculate the probability for 5-stage sleep. The min-batch size is set to 128, and the training data is shuffled in each batch iteration for better convergence and to prevent learning of insignificant features native to individual subjects. After obtaining the learned parameters \mathcal{D}_1 , the soft-max layer is discarded and $\text{CNN}_{\mathcal{D}_2}$ is attached to the network. The outputs of each branch were concatenated and passed to a new fully connected network followed by a soft-max layer as in the previous training phase. The first branch is then frozen for training, and the network is retrained for the new block. The same training procedure is followed for $\text{CNN}_{\mathcal{D}_3}$. At the end of three training steps, we obtained all the parameters for a feature learning section \mathcal{D}_k , $\{k = 1, 2, 3\}$. In all training phases, model check points are set for validation loss, and the model weights and architecture are saved after each successful mini-batch update. The training is stopped when the validation loss begins to stop improving over mini-batch updates. We conducted a series of experiments with different

CNN architectures and several CNN branches for the best performance.

Secondly, the soft-max layer is detached from the model, and the feature sequence learning part (GRU_δ and FC_α) and the outer CNN block CNN_β are attached as shown in Fig. 3.20. Subsequently, the outputs of FC_α , CNN_γ , and CNN_β were concatenated. The pre-trained network blocks $\text{CNN}_{\mathfrak{g}_k}$, $\{k = 1, 2, 3\}$ were then frozen for learning and the network was retrained with Adam optimizer using the same configuration following the same procedure in previous training session.

In the final training, two fully connected networks are attached (each consists of 128 units with ReLU activation) blocks as shown in Fig. 3.20. The model was then trained as a multi-input multi-output convolution neural network. In this phase, all pre-trained parts of the network were set to training disable mode before continuing training with the same optimizer. After obtaining the optimal model, we only trained the fully connoted blocks with a low learning rate to tune the model further. In the final tune up, we used l_2 regularization that adds “squared magnitude” of the coefficients as penalty term to the loss function. Specifically, l_2 regularization is applied before ReLU activation layers in fully connected layer. The results indicated that without using regularization, the model tends to over-fit for unnecessary information presented in the data as noise and high frequency components. Slight performance improvements in the final model is observed with this regularization. After experiments, we observed that ($l_2 = 0.0001$) works optimal for our model.

After we obtained our main CNN-GRU model ($\mathcal{M}_{\text{base}}$) with all channels, we performed some additional experiments with different combinations of channels. Essentially, six additional configurations are tested with the same architecture. A transfer learning technique is used to train models for different combinations of input channels. First, we test the feasibility of adapting the model with a single channel configuration. To achieve four single channel models (\mathcal{M}_{c_3} , \mathcal{M}_{c_4} , \mathcal{M}_{o_1} , \mathcal{M}_{o_2}), the same EEG channel signal was assigned for each input in the main model. Subsequently, the model was retrained with a lower learning rate with Adam optimizer (0.0008). Secondly, the model is tested for the other three channel configurations. In this experiment, two models was trained using only three electrodes chosen from left and right

regions of the head (\mathcal{M}_{left} , \mathcal{M}_{right}). Each electrode is replaced with its counter side channel to obtain a 3-channel configuration. For example, EOG-left, C3, and O1 were replaced by their counter side electrodes, EOG-right, C4, and O2, to obtain \mathcal{M}_{right} . In all the experiments, training is disabled for feature extraction section. Weights of the other parts of the network are then updated in the retraining process for different channel configurations. When adapting the base model for different configurations, it is expected that the global pooling operations (global max and global average) in CNN_β and CNN_γ were re-adjusted for each channel configuration.

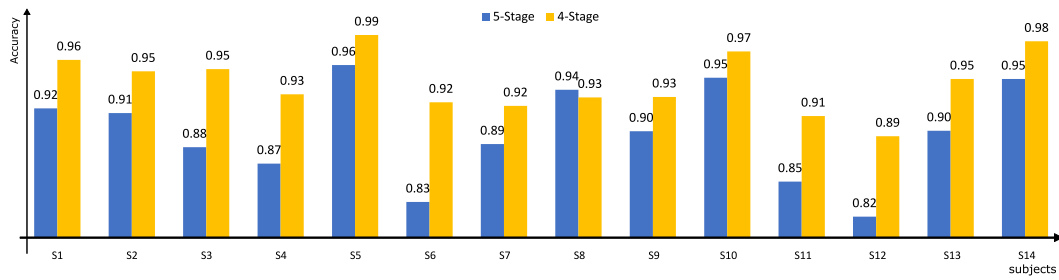


Figure 3.22: Accuracy comparison of 4-stage and 5-stage cases for each subject.

3.5.5 Experimental setup

Prior to finalizing the model architecture, we perform a series of experiments to verify the optimal configuration in CNN blocks and RNN blocks by varying possible hyper parameters. Specifically, for the feature extraction section we commence with 5 CNN branches with different configurations e.g., varying filter parameters, pooling layers, and activation functions, and test several architectures by varying the number of branches until the optimal performance is observed. In the subsequent training phases, we change the configuration of the RNN block, and several bypass blocks are also used based on the learning curves and overall performance. Additionally, we use several values for the learning rate and regularization factor ranging from $10^{-2} - 10^{-4}$ and $10^{-2} - 10^{-5}$, respectively.

3.5.6 Results

In the following sub-sections, we clarify the outcomes of the proposed approach III and discuss the significance of the result using the performance measurements described above. In

order to evaluate the proposed CNN-GRU model, we divided our experiments into three sections as described below.

1. Sleep stage classification (5-stage, 4-stage) using multi-channel model \mathcal{M}_{base} and generate performance metrics
2. Sleep stage classification (5-stage, 4-stage) using single channel models ($\mathcal{M}_{c_3}, \mathcal{M}_{c_4}, \mathcal{M}_{o_1}, \mathcal{M}_{o_2}$) and generate performance metrics for each channel to compare the performance channel wise
3. Sleep stage classification (5-stage, 4-stage) using right and left channels and generate performance metrics to compare the effect of left and right regions on the head

Base model: Experiment I

In this section, we illustrate the detailed results of the base model \mathcal{M}_{base} . Fig. 3.22 shows the overall accuracy of the base model for individual subjects for both classification scenarios. Evidently, in majority of the cases, the 4-stage classification exhibits higher accuracy when compared to the 5-stage case. The average accuracy for the 5-stage case is computed as 89.66 % with a 95 % confidence interval (CI) of (87.17 – 92.15 %). In 4-stage case, the accuracy is reported as 94.21 % with a 95 % CI of (92.65 – 95.76 %). The maximum accuracy for the 5-stage case is 95.81 % and that for the worst case is 81.92 %. Conversely, the maximum accuracy for the 4-stage case was 98.57 % while that for the worst case was 89.29 % of confidence.

Fig. 3.23 shows the IQR plots for per-class performance achieved by the base model with multiple channels. These plots show how well the classifier generalizes across patients. Narrower ranges indicate good performance. It is important to note that subjects with no N3 (deep sleep) presented in sleep are omitted when plotting the box plot for N3. Based on the box plot in Fig. 3.23, it is observed that all the measures for N3 (deep sleep) exhibit larger variations when compared to other sleep stages. The lowest observation for any measurement is also observed for N3 stage except for precision. Furthermore, N2 stage exhibits the second large variation for

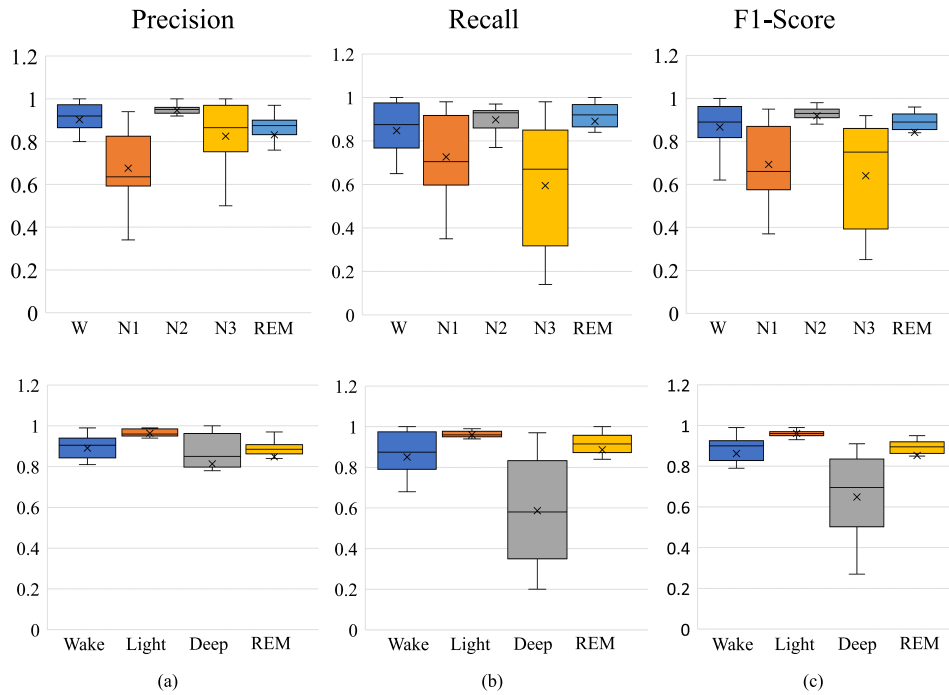


Figure 3.23: per-class interquartile range (IQR) plots for precision, recall, and F1 score for 5-stage and 4-stage classifications obtained via the proposed method for all subjects, (center line: median; box limits: upper and lower quartiles; whiskers: $1.5 \times \text{IQR}$; \times : mean) (a). Precision values (b). Recall (c). F1-score.

any measure in the 5-stage case. The two observations box plots are fairly condensed for other stages.

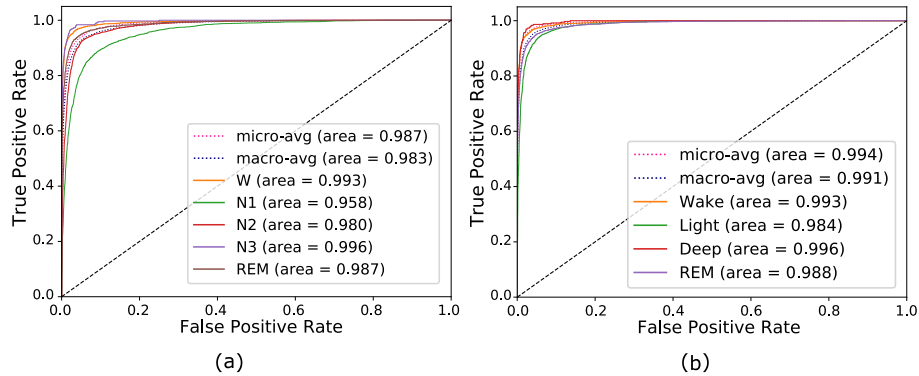


Figure 3.24: Receiver operating characteristic (ROC curve) for \mathcal{M}_{base} (a) 5-stage classification (b). 4-stage classification.

To evaluate the usefulness of the proposed approach, ROC curves are plotted for both cases as shown in Fig. 3.24. In each case, per-class ROC is plotted by considering the rest as negative class. In Fig. 3.24, per-class area under curve (ROC-AUC) and averaged (macro averaged and micro averaged) ROC are also indicated for both cases. the predicted classes contribute equally

for the overall macro-avg ROC-AUC case. In the micro-avg ROC-AUC case, the contribution from each class is aggregated to compute the final average value. Sleep stage classification corresponds to an imbalanced classification problem, therefore micro ROC-AUC is preferable. As shown in Fig. 3.24 (a), all the per-class ROC-AUC are > 0.96 with the exception of the N1 sleep stage, while ROC-AUCs are > 0.98 for both micro and macro averaged values calculated for 5-stage classification. With respect to the 4-stage case in Fig. 3.24 (b), ROC-AUCs are always > 0.98 for all per-class and averaged values. The observation confirms that the classifier is a good choice for both staging cases although N1 exhibits a comparatively moderate performance for the 5-stage case.

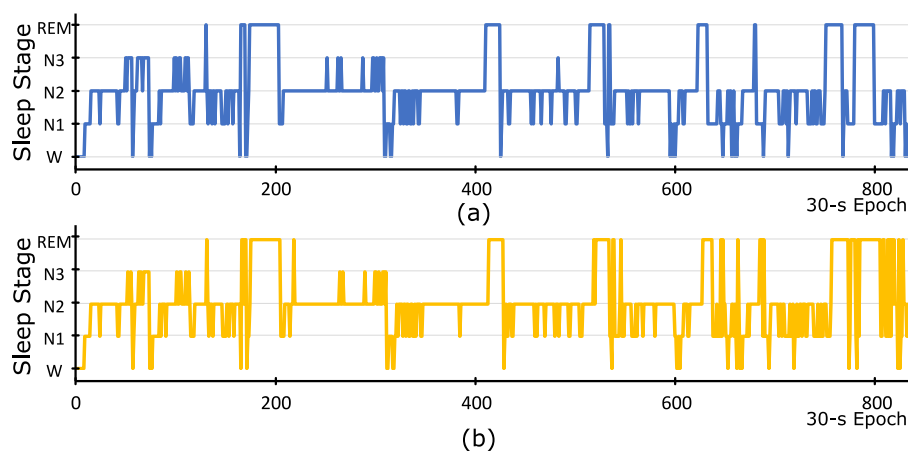


Figure 3.25: Hypnogram for the 5-stage case (a). manually scored by a sleep expert (b). automatically scored by the proposed multi-channel model.

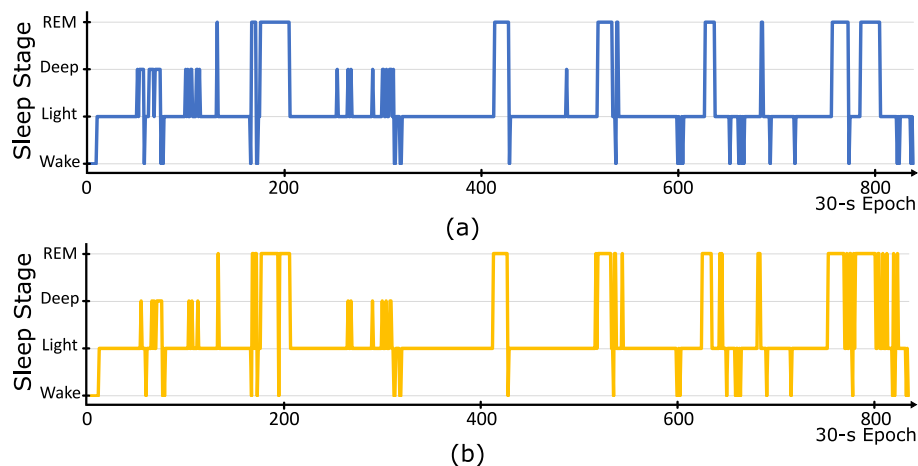


Figure 3.26: Hypnogram for the 4-stage case (a). manually scored by a sleep expert (b). automatically scored by the proposed multi-channel model.

Fig. 3.25(a) and Fig. 3.25(b) demonstrate the hypnograms constructed for the 5-stage case for both true scored sleep stages and predicted sleep stages for subject ‘S2’. Fig. 3.26 illustrates the hypnograms graphed for the 4-stage case for the same subject.

Fig. 3.27 shows the normalized confusion matrix for the multi-channel model for both classification scenarios considering all subjects. The numbers in dark blue denote the normalized values of correctly classified sleep epochs. Based on the Fig. 3.27(a), the lowest classification confidence is observed for N1 and N3 stages. Conversely, deep sleep (N3) exhibits the lowest performance for the 4-stage scenario.

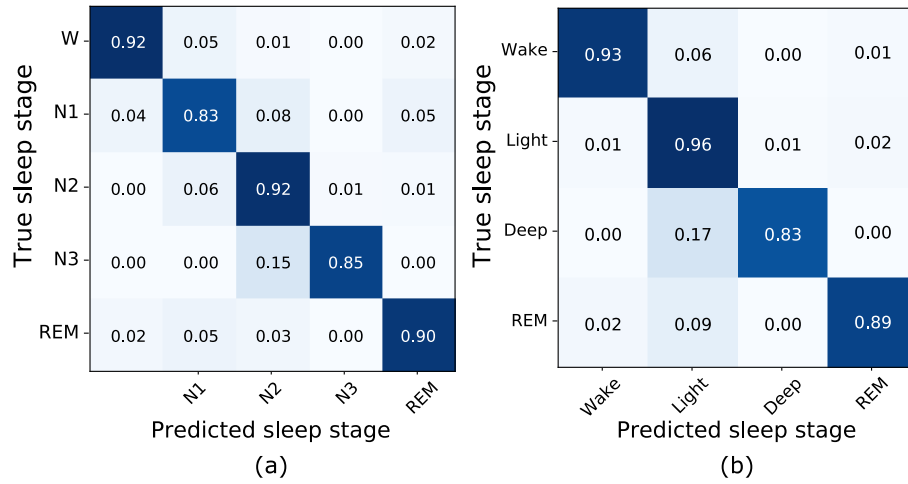


Figure 3.27: Confusion matrix obtained for test subjects (a). 5-stage classification (b). 4-stage classification.

Experiment II and III

In this section, we illustrate the detailed results of experiments II and III.

Fig. 3.28 shows a comparison of overall accuracies and Cohen’s kappa coefficient (κ) for all the models. The minimum κ value and overall accuracy are observed from the models corresponding to occipital area channels. Specifically, M_{left} and M_{right} (three-channel models) exhibit the highest κ and overall values besides the base model.

Fig. 3.29 shows the radar charts of performance measures tabulated in Table 3.8.

Fig. 3.29(a) shows the evaluation matrix for the multi-channel models including the base model as radar charts. As shown in Fig. 3.29(a), the base model yields a more balanced perfor-

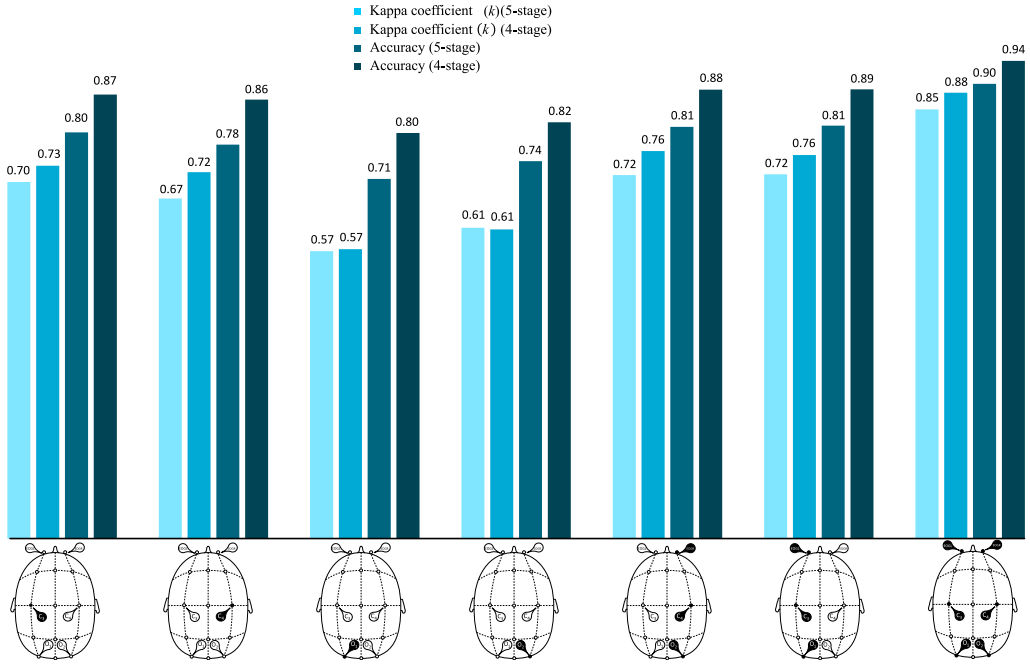


Figure 3.28: Overall accuracy and Cohen's kappa coefficient comparison for all experiments.

Table 3.8: per-class evaluation metrics for proposed classification models

		C3			C4			O1			O2			EOGR, C4, O2			EOGL, C3, O1			all		
		PR	RE	F1	PR	RE	F1	PR	RE	F1	PR	RE	F1	PR	RE	F1	PR	RE	F1	PR	RE	F1
5-stage	W	0.82	0.82	0.82	0.81	0.82	0.82	0.74	0.81	0.77	0.76	0.86	0.81	0.91	0.76	0.83	0.72	0.92	0.81	0.9	0.92	0.91
	N1	0.6	0.65	0.62	0.53	0.6	0.56	0.48	0.44	0.46	0.59	0.4	0.47	0.66	0.53	0.59	0.63	0.6	0.61	0.77	0.83	0.8
	N2	0.89	0.87	0.88	0.89	0.84	0.86	0.85	0.83	0.84	0.84	0.88	0.86	0.88	0.91	0.89	0.88	0.93	0.9	0.95	0.92	0.93
	N3	0.68	0.77	0.72	0.76	0.7	0.73	0.87	0.46	0.61	0.68	0.68	0.68	0.65	0.82	0.72	0.88	0.64	0.74	0.83	0.85	0.84
	REM	0.78	0.75	0.76	0.72	0.76	0.74	0.49	0.6	0.54	0.56	0.62	0.59	0.7	0.88	0.78	0.9	0.65	0.75	0.89	0.9	0.89
4-stage	Wake	0.83	0.81	0.82	0.79	0.84	0.82	0.76	0.78	0.77	0.75	0.87	0.8	0.93	0.75	0.83	0.72	0.91	0.81	0.91	0.93	0.92
	Light	0.91	0.93	0.92	0.91	0.9	0.91	0.85	0.88	0.87	0.86	0.92	0.89	0.93	0.91	0.92	0.92	0.95	0.93	0.96	0.96	0.96
	Deep	0.68	0.76	0.72	0.7	0.76	0.72	0.87	0.49	0.63	0.68	0.68	0.68	0.68	0.79	0.73	0.85	0.66	0.75	0.82	0.83	0.83
	REM	0.79	0.7	0.74	0.75	0.73	0.74	0.55	0.5	0.52	0.64	0.38	0.47	0.72	0.87	0.79	0.93	0.63	0.75	0.9	0.89	0.89

mance when compared to three-channel models while the left and right electrodes exhibit nearly identical performance. However, the right side electrodes exhibit more balanced performance in terms of the aforementioned performance measures.

In Fig. 3.29(b), the highest F1 measure for both classification scenarios is observed for C4 channel. Although C4 exhibits comparatively better performances, both central lobe electrodes exhibit nearly identical performances in terms of precision, recall, and F1-score. As shown in Fig. 3.29(b) O2 and O1 polygons are the lowest in majority of the radar charts, and this indicates poor performance in-terms of all the measures discussed above. Furthermore, central electrodes exhibit more balanced performance over each sleep stage.

In this chapter, we compare the results of all phases. The aim of this study was to imple-

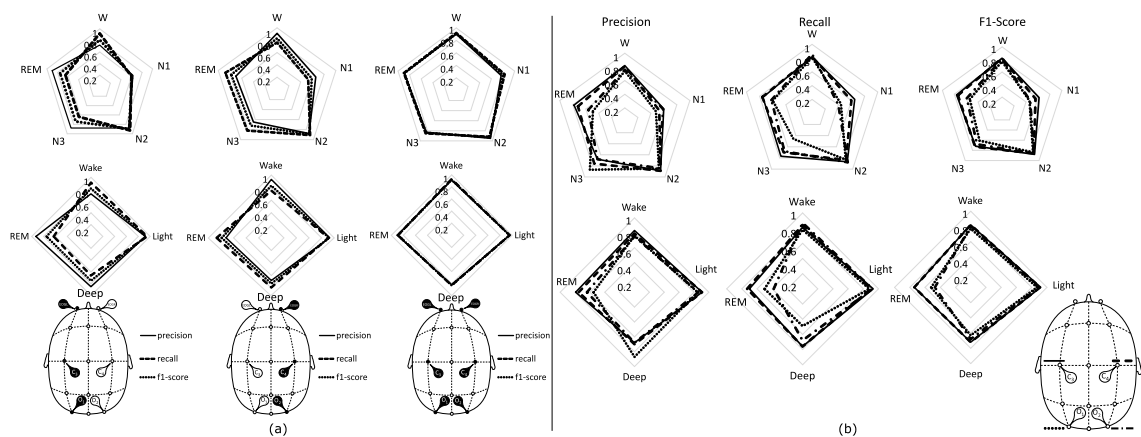


Figure 3.29: Comparison of evaluation metrics for all models for two classification scenarios (a). Multi-channel models (b). Single-channel models.

ment automatic sleep stage scoring for 5-stage and 4-stage classification. In Experiment I we implemented 5 stage classification system by using a deep neural network. Since Experiment I and Experiment III are focused on 5-stage classification we compare the basic performances metrics. Fig. 3.30 shows the improvement of performance metrics. On the accuracy point of view there is a 6% of improvement compared to Experiment I. Other performance metrics also show significant improvement compared to Experiment I. Especially, in Experiment III, the architecture consists of CNN and RNN layers which improve the classification performance.

3.6 Results comparison of Experiment I, Experiment II, and Experiment III

In Experiment II and III we implemented 4 stage classification system. Fig. 3.31 shows the improvement of performance metrics. There is a 2% accuracy improvement in between Experiment II and Experiment III . Furthermore, there is a improvement in other metrics compared to Experiment II. In the AUC-ROC point of view, there is significant improvement in light sleep stage due to the arrangements in Experiment III.

However, the accuracy of single channel models are slightly better in Experiment II models. Fig. 3.32

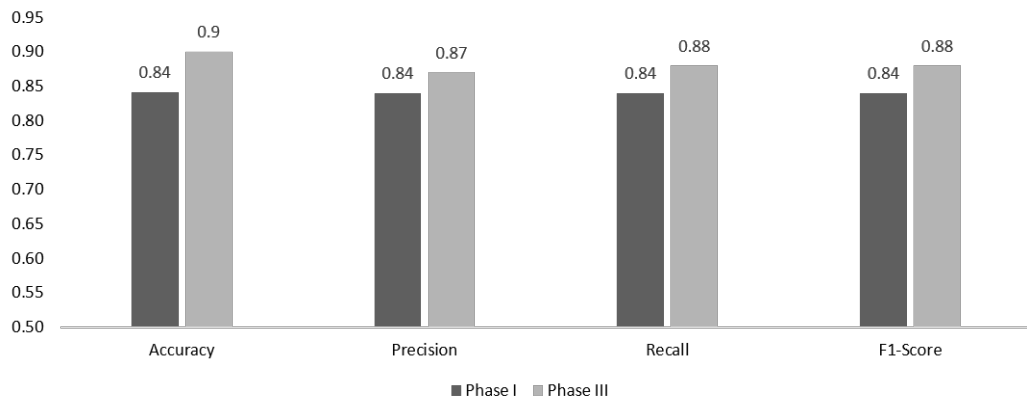


Figure 3.30: Result comparison Experiment I and Experiment III (5-stage case)

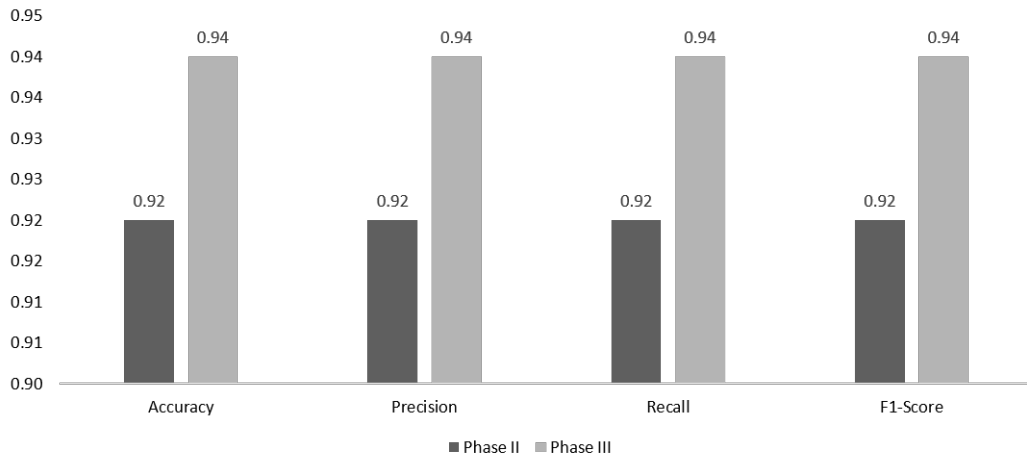


Figure 3.31: Result comparison Experiment II and Experiment III (4-stage case)

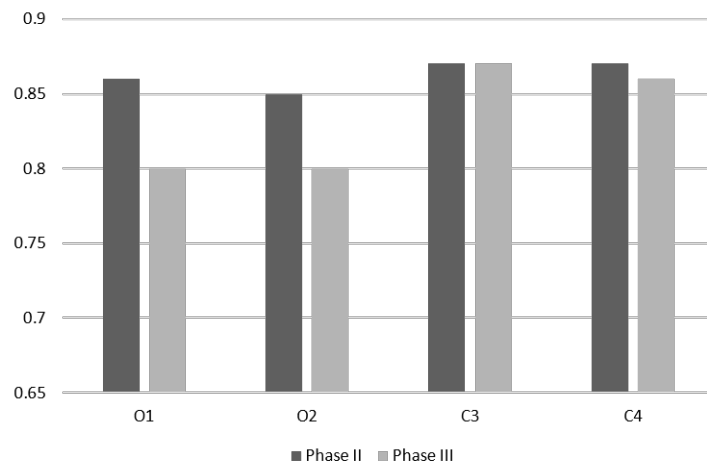


Figure 3.32: Accuracy comparison Experiment II and Experiment III (4-stage case, single)

4

Obstructive Sleep Apnea Syndrome Detection Based on Fused Time-Frequency Spectral Images

4.1 Materials and Methods

As described in Chapter 2, Obstructive Sleep Apnea (OSA) is a common chronic sleep disorder that disrupts breathing during sleep. On the other hand, apnea is associated with many other medical complications, such as hypertension, coronary heart disease, and depression. The golden standard for diagnosing and treating OSA involves nocturnal polysomnography (PSG)

study. As the detailed PSG study requires special equipment, specialized human intervention, dedicated analyzing skills, the availability of OSA diagnosis in public health sectors are not satisfactory. As the traditional PSG uses lot of body sensors, it not comfortable have a good good sleep with all the sensors attached to the body. Therefore, a simpler and low cost OSA detection method is required to automate the polysomnography procedure and reduce the discomfort. ECG signals have been studied in many researches to replace the PSG based OSA detection. Most of such proposed approaches rely on feature engineering, which calls for advanced expert knowledge and experience. In this study, a novel fused-image-based technique is proposed using only a single-lead ECG signal. A CNN is used to extract features automatically from the spectral images created with one-minute ECG segments. The proposed network comprises 37 layers Fig.(4.3), which comprise with residual blocks, dense layer, dropout layer, and a soft-max layer.

In this phase of the study, novel methodology for OSA detection is proposed and implemented using fused spectral images created by combining Short-Time Fourier Transform (STFT) and continuous wavelet transform (CWT) representations (*see subsection 4.1.3*). A deep CNN model is employed to perform feature extraction and classification of apneic and non-apneic ECG segments using those fused image representation as inputs. further more the proposed method does not utilize any QRS-based features or other features manually derived manually in performing the classification. Instead of using a limited number of features extracted from the QRS complex or other EDR signals, more detailed and complex features derived from a combination of spectral images are used to detect the presence or absence of apneic events.

4.1.1 Dataset

To evaluate the proposed methodology, a popular and widely used dataset provided by PhysioNet is used (Apnea-ECG database provided by Dr. Thomas Penzel at Philips University [128, 129]). The dataset comprises single-lead ECG signal from 70 subjects. The recordings are in two groups (a released set and a withheld set), each with 35 subjects (see Fig. 4.1). The

training dataset has further divided into three groups based (a01 through a20, b01 through b05, and c01 through c10). Test set (withheld set) is numbered from x01 to x35. Apnea annotations are made by human experts based on simultaneously recorded respiration and related signals. Eight recordings (a01 through a04, b01, and c01 through c03) have additional signals (Resp C and Resp A, chest and abdominal respiratory effort signals obtained using inductance plethysmography; Resp N, oronasal airflow measured using nasal thermistors; and SpO_2 , oxygen saturation). However, only ECG signal is considered in this study as the main purpose of this study is to use ECG signals to classify the OSA events.

The ECG signals were recorded at a sampling rate of 100 Hz and with 16-bit resolution. Each ECG signal lasted 420–600 min with a mean of 492 ± 32 min. Non-overlapping one-minute ECG segments were annotated as either ‘OSA’ or ‘Normal’, but no distinction was made between cases of hypopnea or apnea. The PhysioNet Apnea-ECG database includes both male and female subjects aged from 27 to 63 years with a mean of 43.8 ± 10.8 years. The body weights of the subjects range from 53 to 135 kg with a mean of 86.3 ± 22.2 kg (see Fig. 4.2). The sleep recordings were obtained from 25 male and 7 female volunteers, including both healthy and OSA subjects [130, 131] (see Fig. 4.2).

4.1.2 Method

The Time Frequency Representation (TFR) of a signal is utilized to analyze the information contain in various types of signals, including physiological, acoustic signal, and geophysical signals. specially, TFRs are employed to identify complex and high-dimensional non stationary properties of a signal. STFT and CWT are two of the most widely utilized visual representations which play an important role in analyzing non stationary signals. In particular, the fluctuation of the frequencies, amplitudes over time, and morphological variations in ECG signals can be better represented using STFT and CWT instead of FT (see Eqs 4.1 and 4.3) as it focuses on both time domain and frequency domain attributes. Normally the TFR of a signal is illustrated as a colored image (heat map) in a spectrogram or a scalogram (see Fig. 4.6). Usually, the visual representation of the STFT is called spectrogram, and the scalogram is the visual representation

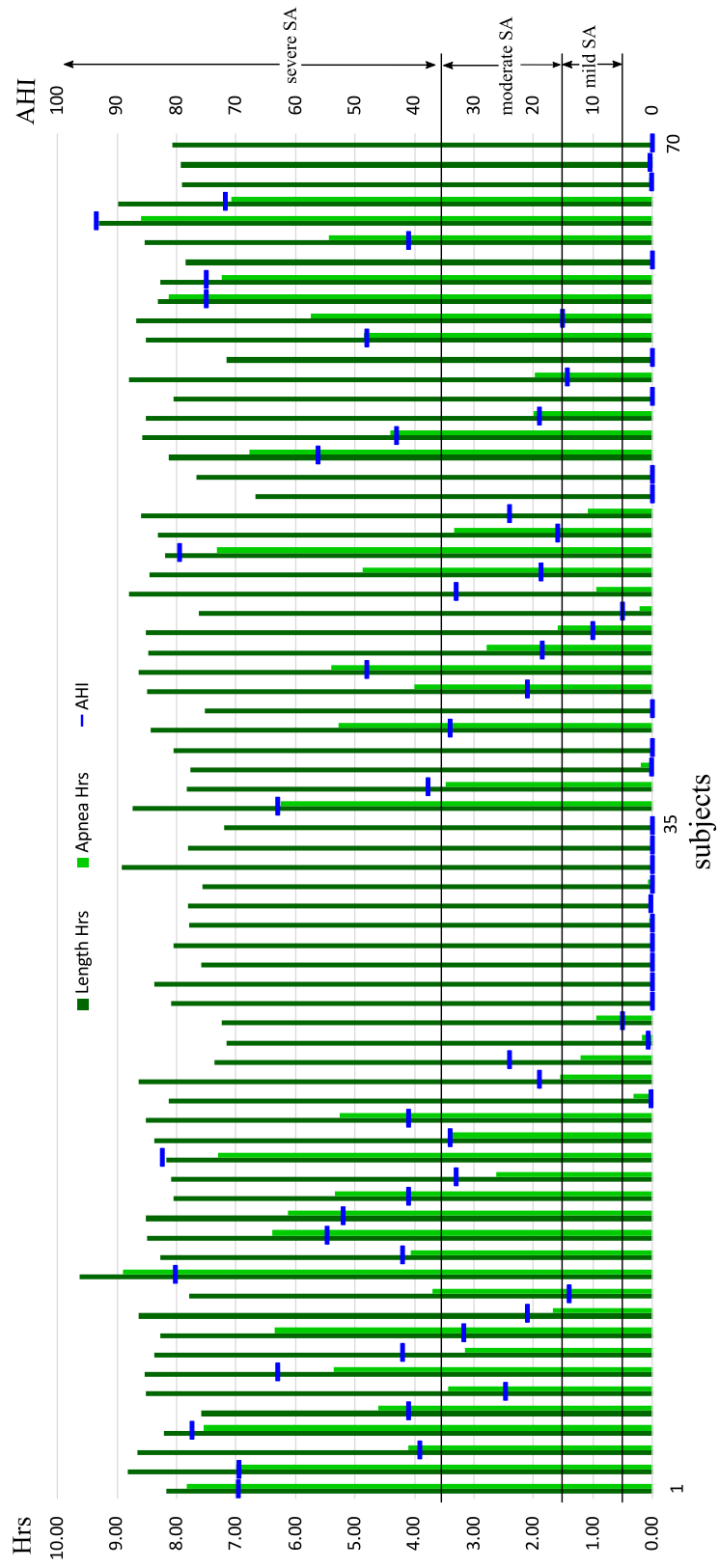


Figure 4.1: AHI and apnea episodes distribution in the dataset

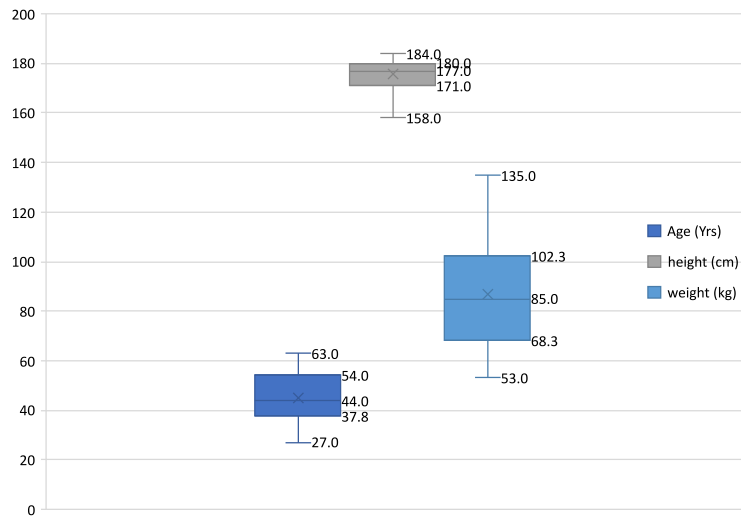


Figure 4.2: Additional information about the subjects in the PhysioNet Apnea dataset

of CWT [132].

STFT is used to construct the TFR of the physiological signals as a spectrogram with a constant time–frequency resolution (see Section 2, Eq 4.1). A constant sliding window along the time axis is employed to create a two-dimensional (2D) representation of the signal at this fixed resolution [130, 133, 134]. As a result of using a constant window, all the frequency information is analyzed at the same time–frequency resolution.

However the STFT based signal processing is simple, efficient, and robust method in solving many research problems associated with many kinds of signals. As mentioned earlier, the performance of STFT analysis is heavily dependent on the analysis window and it needs lot of book keeping to optimized the hyper parameters. Despite of the negative points aforementioned, STFT is a widely used primary tool for decades within speech and radar processing research communities. Customarily, the selection of window sizes and overlapping length is arbitrary and depend on the experience, and this is why the STFT based analysis is criticized as a heuristic method [135].

In contrast to STFT, the CWT’s wavelet window is scaled and shifted during the transformation. This provides long time windows for low-frequency regions and shorter time windows for high-frequency regions. Due to the scaled and shifted window, the wavelet transform is capable of comprehending the time and frequency information simultaneously providing a multi

resolution representation of the signal in both low- and high-frequency regions. The mathematical formula for calculating the wavelet coefficients is given in equation 4.3. As described in this equation, a basis function, i.e., the mother wavelet $\psi(t)$, and its scaled and dilated versions are used to decompose the time-domain signal. Due to the advancements emerged in CWT analysis, the abrupt changes transpired in the signal is grasped so that the TFR can be used for more complex analysis. However, the use of a window introduces a compromise between time localization and frequency localization in both TFR's.

In this study, we conducted experiments comparing four types of spectral images: scalogram images, spectrogram images, images based on smoothed pseudo Wigner–Ville distribution (WVD) [136, 137], and fused images (a hybrid version of CWT and STFT images, (see Fig.4.6 and Fig.4.7)), to identify apneic events. However, it should be noted that the Wigner–Ville distribution method has cross-term issues when used with non-stationary signals [138]. Similar to the CWT and STFT WVD can also provide quantitative information of the signal energy distribution in time-frequency domain.

The proposed apnea-detection method is based on deep learning, using a fusion (combination) of two spectral images (scalogram and spectrogram images) for one-minute ECG segments. Firstly, each one-dimensional ECG segment in the time domain is converted into more-detailed images format (scalogram, spectrogram, Wigner–Ville distribution, and fused image). the images are then passed to the CNN to perform automatic image feature extraction and classification.

As explained in the Introduction section, the idea behind combining the two TFRs is to increase and enrich the discriminative feature in newly formed images. As Apnea-ECG database provides one-minute-based annotations, the proposed methodology also utilizes one-minute ECG fragments to identify apneic and non-apneic episodes. Even though there are plenty of good reasons for employing TFR or fused TFR in solving various research problems, expertise knowledge is expressly required to extract and analyze specific features from such representations. In other words, it is not realistic to accomplish tasks such as selection, analysis, or identification of specific patterns or features in a TFR manually as TFRs contain very fine and

complex details. Therefore, the best option to analyze such a complex spectral image is employing the deep learning techniques since many deep learning techniques such as CNN can perform this task intelligently and automatically.

In this work, a residual learning approach is employed to perform OSA detection, which is schematically illustrated in Fig. 4.3. Generally, a plain CNN is obtained with a number of stacked layers of linear and nonlinear processing units. These layers allows the network to learn complex, detailed and fine representations at different levels of abstraction [139]. A typical residual network differs from a plain CNN due to its “skip” connections, as exemplified in Fig. 4.3. In a residual network, the activations from the earlier layer are reused until the posterior layer learns the weights. The skip connections are very important when mitigating the gradient vanishing and degradation, which are commonly seen complications in large plain networks. In general, a residual network can be easily trained to learn a residual mapping with fewer stacked layers than a plain network, with substantially good performance in image classification [140–142].

4.1.3 Preprocessing and image creation

Signal noise and baseline drift

Generally, it is common that a raw ECG signal is corrupted by various types of noises which implants unwanted information into the signal. Unwanted presence of noise can be generated from any external or internal source including power lines, conductors near to the device, RF transmitters, motors running near to the device which draws inrush currents. Electromagnetic interference (EMI) is the noise caused by other conductors or cables placed near to the device. Radio Frequency Interference (RFI) is also an external source of noise caused by radiating signals from wireless communication systems. On the other hand, the noise can be generated from the device itself. As an example, a presence of a noise can be observed due to a faulty components or loose connections. Further, a loose connection of leads also can generate an unwanted presence of noise. Another possible source of noise is environmental issues such as mechanical

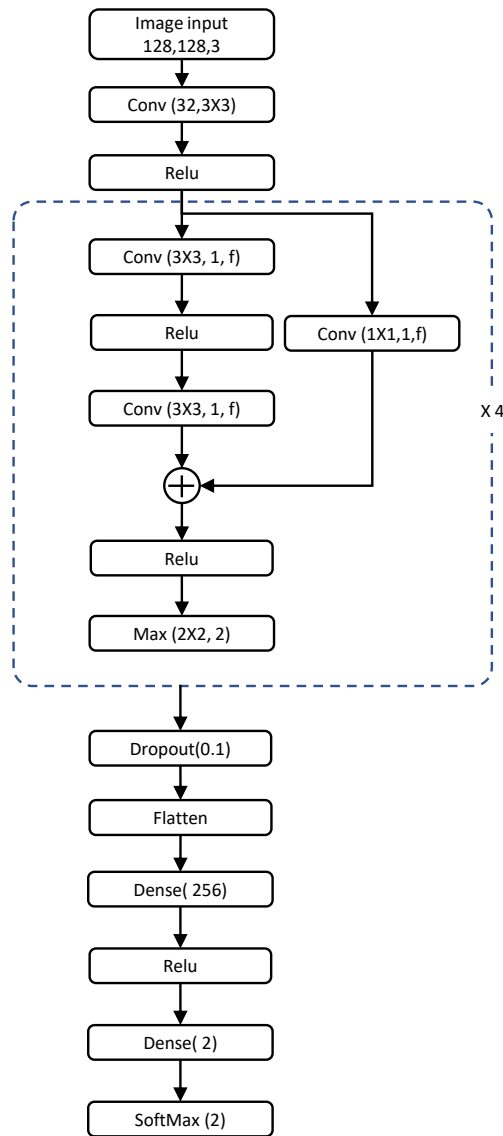


Figure 4.3: Proposed 2D-CNN network for OSA event detection. “ $Conv(k,s,f)$ ” denotes a convolutional layer where k , s , and f are the kernel size, stride size, and number of filters, respectively. “ $Max(p,s)$ ” denotes a max-pooling layer where p and s are the pool size and stride size, respectively. The values for the filter sizes “ f ” in the four residual blocks are 32, 64, 96, and 128.

vibrations and fluctuations in temperature. The internal noise of electronic components can be generated due to the changes of the temperature. The fundamental physical properties of the electronic components can fluctuate naturally on temperature variations. This noise is called thermal noise (Johnson noise). It is unrealistic to eliminate all the noise sources completely, even though noise removing hardware filters are used in the data acquisition. Therefore, in some cases, artifact removal is crucial to get the desired information from the signal. On the

other hand, presence of noise in the TFR image may cause imprecise estimation of characteristic points and features [143].

Besides the artifact aforementioned, temporal drifts may also presented in the ECG signal which are not related to the desired information. Similar to the signal noise case, many internal and external sources might induce time varying temporal drifts. The effect of such drifts can be reduced using baseline correction. Usually the baseline drift is generated due to the background fluctuations appears as slow but wide ranging ups and downs. Therefore it can be considered sort of low-frequency noise. There are many methods to perform baseline correction including using least squares method, computational geometry, Fourier analysis, Wavelet analysis, and neural networks. In many signal processing applications, wavelet methods are widely applied [144].

Therefore, signal denoising was performed using three of MATLAB's built in functions: "wavedec," "waverec," and "cmddenoise" [136]. First, the raw segments were transformed (one-minute fragments) into wavelet coefficients using "wavedec," with the "sym8" wavelet used to perform the baseline correction. After then, the "cmddenoise" function was employed to perform interval-dependent thresholding to the baseline-corrected signal. Fig. 4.4 shows part of a raw ECG segment and its prepossessed waveform before being transformed into its image format. The denoised dataset was then converted into four TFR image datasets as illustrated schematically in Fig. 4.5.

Image creation

A spectrogram dataset was created to evaluate the performance of the proposed model in Fig 4.3. MATLAB's builtin function "spectrogram" was employed with a "blackman" window to generate spectrogram images. While creating the spectrogram images, the window size was set to 64 (640 ms), and the overlap was set to 60 samples (600 ms) [137]. The definition of window function $\omega(n)$ is given in Eq 4.2. MATLAB's "cwt" was employed to create the scalogram images using the "Morse" analytic wavelet. The scalogram image was formed using the squared modulus of the CWT coefficients as a function of time and frequency, where the frequency is

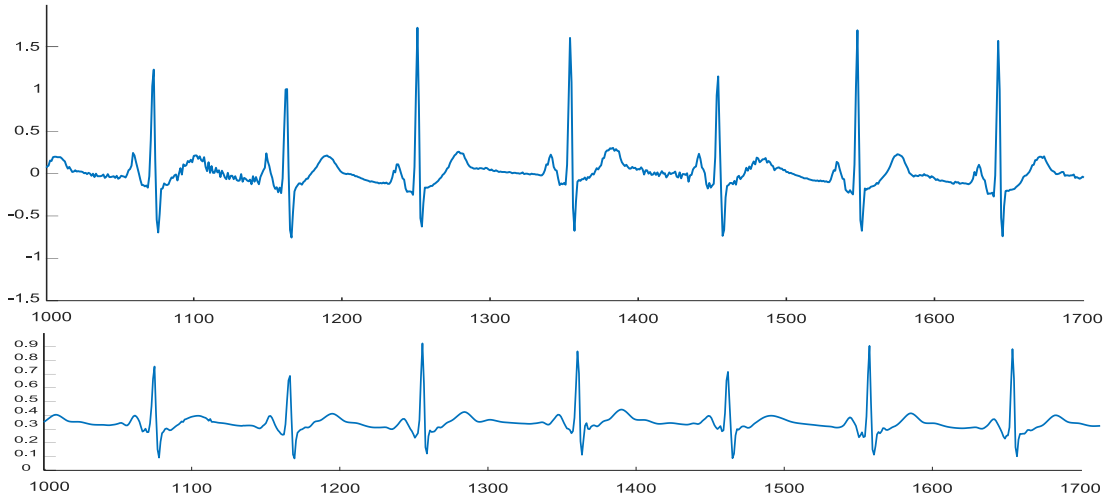


Figure 4.4: Preprocessing ECG segments. (a) Part of an original ECG segment. (b) The denoised and amplitude scaled version.

plotted on a logarithmic scale. The height and width of the created scalogram images represent the frequency and time, with the red/green/blue (RGB) colors representing the absolute values of the CWT mapped into a (three-dimensional) color map.

In this study, the “*Morse*” wavelet was used as the mother wavelet for the CWT of the ECG segments, given that it had already been used successfully in many research applications [109, 145, 146]. We saved both sets of images, generated with “*cwt*” and “*spectrogram*,” using the “*gcf*” command. Fig. 4.6 shows both TFR images created for normal and apneic ECG segments.

For comparison purposes, we also used MATLAB to prepare a TFR involving a smoothed pseudo Wigner–Ville distribution, with time and frequency windows used for the smoothing.

$$X_{STFT}[m, n] = \sum_{k=0}^{L-1} x[k] \omega[k-m] e^{-j2\pi nk/L}, \quad (4.1)$$

$$\omega(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{L-1}\right) + 0.08 \cos\left(\frac{4\pi n}{L-1}\right), \quad (4.2)$$

where L is the window length, $x[k]$ is the input ECG signal, m is the time (discrete), n is the frequency (discrete) and ω is the window function. The log values of $X_{STFT}[m, n]$ are used to create the RGB color image (the spectrogram image) [133].

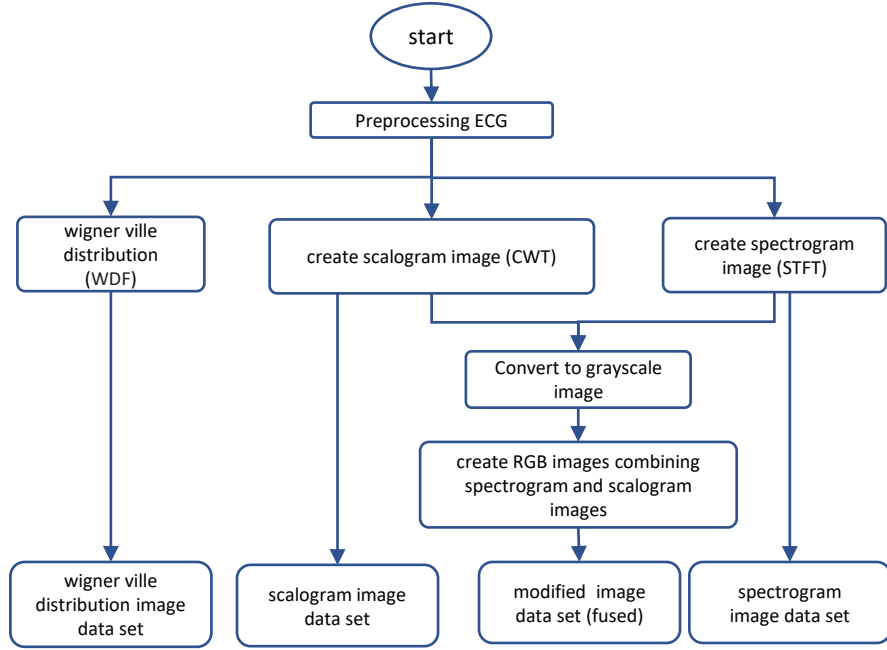


Figure 4.5: Image dataset creation

$$W_x(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \psi^* \left(\frac{t - \tau}{s} \right) dt, \quad (4.3)$$

where $W_x(s, \tau)$ is the wavelet coefficient, $x(t)$ is the ECG signal, $\psi(t)$ is the basis function (mother wavelet) conjugate, s is the scale, and τ is the time parameter.

Before creating the full dataset of spectral images, we experimented several window sizes, overlap lengths, window types, and other parameters, using a small amount of randomly selected data to make sure that appropriate and comprehensible images were generated. Here the images were inspected visually and tested using the proposed CNN (The CNN is trained using small part of the dataset).

After confirming the most appropriate parameters for image generation, the scalogram, spectrogram, and pseudo Wigner–Ville distribution image datasets were constructed and saved in the computer for further analysis. After constructing the different types of TFR datasets, the fused image dataset was prepared using the scalogram and spectrogram images as illustrated in Fig. 4.7. Fused were created by integrating gray-scale values of the scalogram and its matching spectrogram into three layers of an RGB image. To blend the CWT and STFT representations

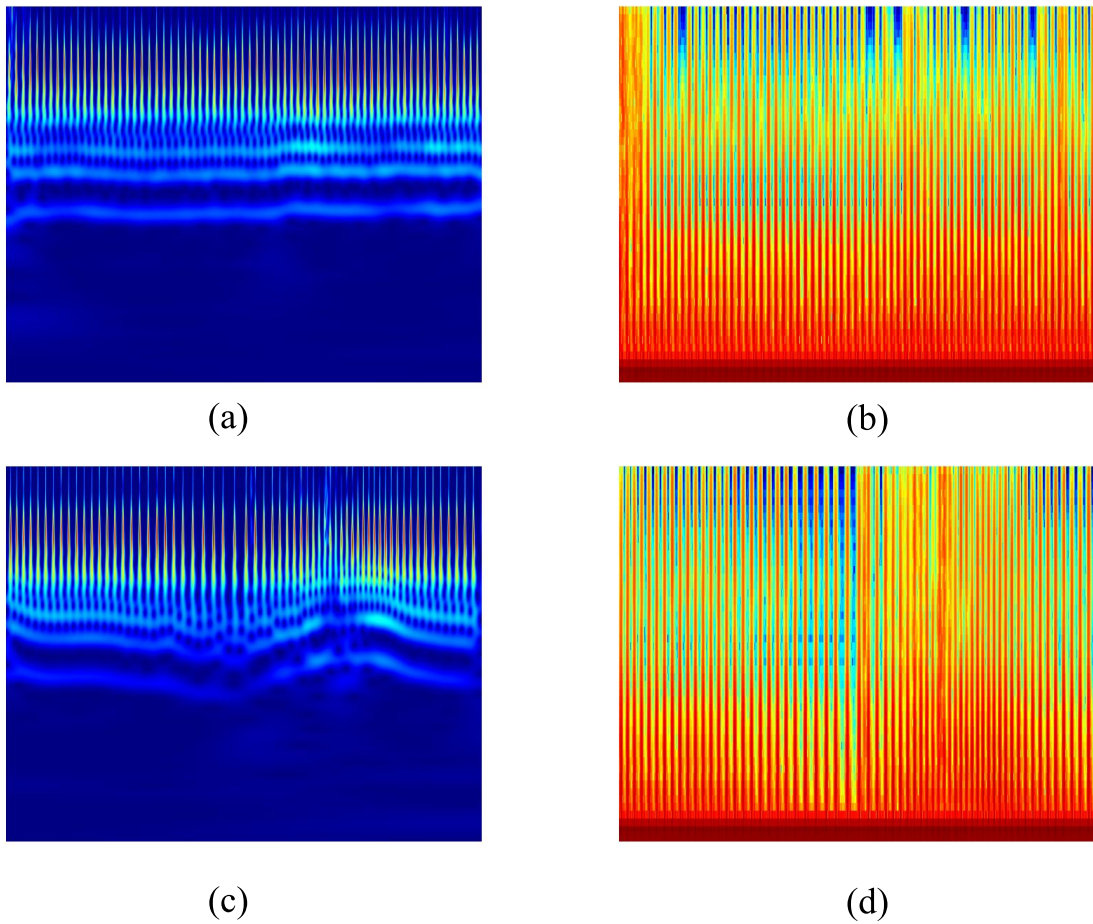


Figure 4.6: One-minute ECG segments transformed into (128, 128, 3) RGB images. (a) Scalogram image of a normal ECG segment. (b) Spectrogram image of the normal segment. (c) Scalogram image of an apnea ECG segment. (d) Spectrogram image of the apnea segment.

into one image, the gray-scale values of the scalogram image were employed as the “red” component of the new image (the red layer of the image), and the “green” component was formed using the corresponding gray-scale values of the spectrogram. The “blue” layer was created by the addition of the gray-scale values of the scalogram and spectrogram. In this way, the three red, green and blue layers of the fused image accommodated picture elements from both scalogram and spectrogram representations. As shown in Fig. 4.7, the modified image (fused image) is therefore a hybrid version of the CWT and STFT images which contains more discriminative information than the original forms of its TFRs. In other words, each pixel or point represents the spectral presence of the ECG wave derived from both Time-Frequency representations.

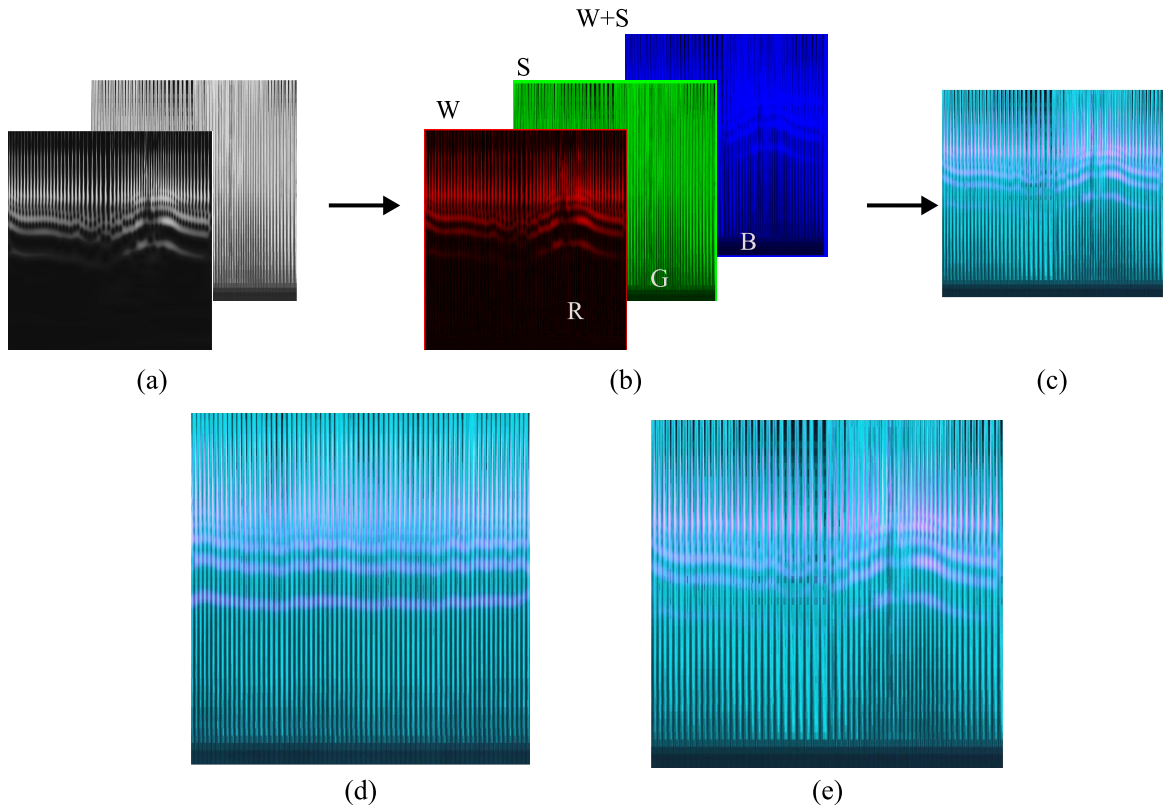


Figure 4.7: Fusing the scalogram and spectrogram for an apnea ECG segment. (a) Gray-scaled scalogram and spectrogram images. (b) RGB components of the modified image, where W is the gray scaled values of scalogram, and S is the gray scaled values of spectrogram (c) Fused image. (d) Fused image of a normal ECG segment. (e) Fused image of an apnea ECG segment.

4.1.4 Proposed model

The proposed architecture to which the images are fed, is shown in Fig. 4.3. The CNN comprises four residual blocks sharing the same architecture but with different hyper parameters. It has 37 layers, including the convolutional, max-pooling, dense, rectified linear unit (ReLU) layers and other layers. More specifically, there are 13 convolutional layers and four max-pooling layers. The model starts with 32 convolutional filters (3×3) followed by a (ReLU) activation layer. The output generate by the first convolutional layer is passed to a series of residual blocks, as shown in Fig. 4.3. Each residual block is formed with two consecutive convolutional layers and a skip connection through a 1×1 convolutional layer to make sure that the dimensionality is restored. Each residual block is followed by an addition layer, a ReLU activation layer and a 2×2 pooling layer with a stride size of 2 to summarize the feature map generated by each residual

block. The number of filters (f in Fig. 4.3) used in all convolutional layers in the same residual block is kept unchanged. The max-pooled output of each residual block is then passed to the next residual block consecutively. The number of filters (f) in a residual block is successively increased as (32, 64, 96, and 128). The max-pooled output of the last residual block is passed to a 0.1 dropout layer to prevent the CNN model being over-fitted. Adding a dropout layer is one of a recognized technique where some nodes are dropped out randomly during training. This is a very effective method in regularizing the model which limits the over-fitting, and also it reduce the generalization error in a DNN model.

Finally, the flattened output of the dropout layer is send to a Fully Connected (FC) layer with 256 units followed by “ReLU” activation layer. The classification layer is a soft-max layer, where the output of the network is normalized to a probability of y_k , as specified by Eq 4.4. The FC layer with 256 units followed by “ReLU” activation works as the classifier for the features derived from the deep stacked residual blocks as illustrated in the model architecture in Fig.4.3.

$$y_k = \frac{\exp \{a_k\}}{\sum_{j=1}^K \exp \{a_j\}}, \quad k = 1, 2, \quad (4.4)$$

where a_k is the activation (a linear weighted sum of the hidden nodes) of the k^{th} neuron in the soft-max layer, and y_k is the probability of the individual class.

4.1.5 Implementation of model training

The proposed CNN model was implemented and trained using the MATLAB R2020a deep-learning toolbox [147]. The training of the model is done with graphics processing unit support (NVIDIA GEFORCE GTX 1070) using 10-fold cross validation method [148]. 20,000 “normal” ECG segments and 13,062 “OSA” ECG segments were selected to train the proposed network. The images (D) were randomly split into 10 equal subsets $\{f_1, f_2, f_3, \dots, f_k, \dots, f_{10}\}$ of images. Then one image subset was chosen as the test dataset and the remainder of the dataset was used to train the CNN model, resulting 10 models (see Fig. 4.8).

After splitting the test and training datasets per each fold, random oversampling was em-

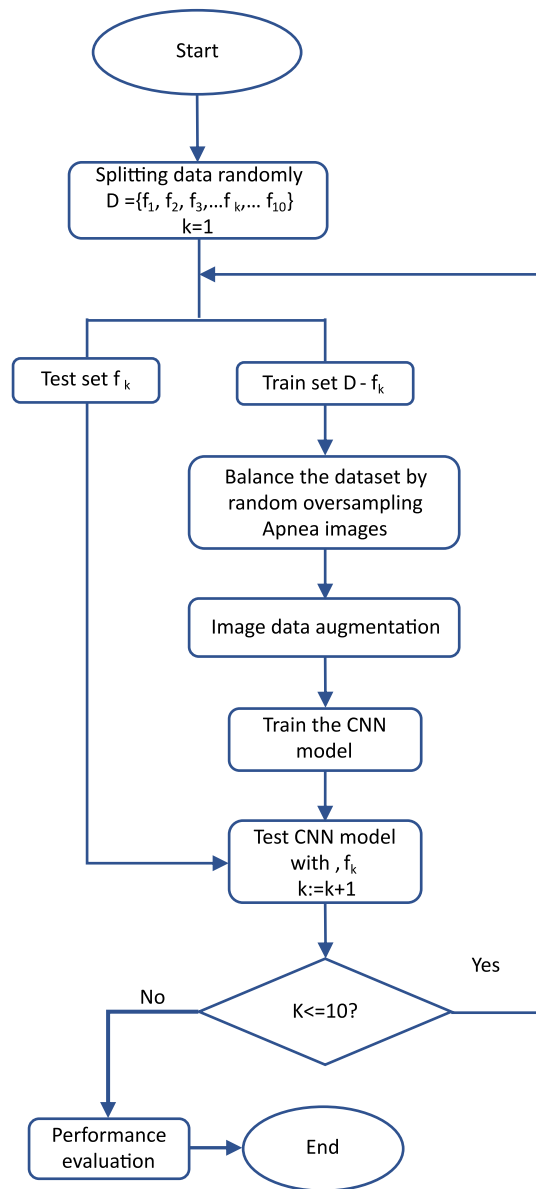


Figure 4.8: Schematic diagram of the training procedure for the proposed 2D-CNN model with 10-fold cross validation.

ployed to balance the dataset and prevent the model from being overfitted. More specifically, for each fold, the training set comprised with 18,000 normal one-minute ECG segments. To balance the dataset, apnea images were randomly copied (the minority class) so that the total number of OSA images was also 18,000. Then, all the training images, including the randomly oversampled images, were subjected to fold-wise image augmentation using

- random rotation (-8 to $+8$ degrees)

- random horizontal translation (−30 to +30 pixels)
- vertical translation (−10 to +10 pixels)
- random shearing (−5 to +5 degrees)
- random horizontal flipping

It should be noted that, small-scale augmentation was performed for the training images since the spectral images are usually consistent and steady compared with normal images. In other images taken by a camera, are attributed with large rotations, various scales, vivid colors, and special effects are presented compared to the spectral images created by computer. Therefore, the images were slightly to keep the consistency of the TFR images. The training procedure followed in this study is demonstrated in Fig. 4.8. When training each fold, the mini-batch size was set to 128, and the each CNN model was evaluated in every 256 iterations both to ensure that the model is converging and to visualize the training process (see Fig. 4.12) during training.

A back-propagation algorithm was used to train the whole model by optimizing the cross-entropy error E_{ce} between the predicted values and the actual ground truth values, as specified in Eq 4.5, using the “*Adam*” optimizer with an initial learning rate of 0.001, as suggested in [114][115]. Further, each fold was run for up to 48 epochs, until the training loss between consecutive batch updates ceased to improve. After training models for every fold (10 folds), the best model was chosen for evaluation, based on its validation accuracy.

$$E_{ce} = - \sum_n \sum_{k=1}^K t_{n,k} \log y_{n,k} \quad (4.5)$$

where $y_{n,k}$ is the actual output of node k , n is the number of examples in the mini-batch, and $t_{n,k} \in \{0, 1\}$ are the target outputs [149].

4.1.6 Evaluation metrics for binary class scenario

Overall accuracy, per-class recall (RE), per-class precision (PR), per-class specificity (SP), and per-class F_1 score (F1) as defined in Eqs (4.6), (4.7), (4.8), (4.9), and (4.10), respectively [150] were utilized to evaluate the proposed model. As illustrated in the Method section, the average result of the 10 folds for each performance metric was calculated in order to evaluate the final performance of the proposed CNN model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.6)$$

RE (also known as the probability of detection, true positive rate, or sensitivity) reflects the correctly predicted proportion of all positive samples.

$$\text{RE} = \frac{TP}{TP + FN} \quad (4.7)$$

PR (also known as the positive predictive value) reflects the proportion of positive predictions that are actually correct.

$$\text{PR} = \frac{TP}{TP + FP} \quad (4.8)$$

SP (also known as true negative rate) reflects the proportion of negatives that are correctly detected.

$$\text{SP} = \frac{TN}{TN + FP} \quad (4.9)$$

The F1 score denotes the harmonic mean of PR and RE, which considers both metrics to give an optimal measure for analyzing model performance.

$$\text{F1} = 2 \cdot \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (4.10)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false neg-

atives, respectively.

4.2 Results

In this work, an fused image-based method for OSA detection is presented using one minute ECG segments. To analyze the effectiveness of the proposed CNN model, other previous works were compared. However, all subjects were employed in training the network in the 10-fold cross-validation procedure. Therefore, only per-segment OSA-detection performance is compared. Further it should be noted that the test dataset is completely isolated from the training image dataset before performing random oversampling. The test image dataset is not augmented.

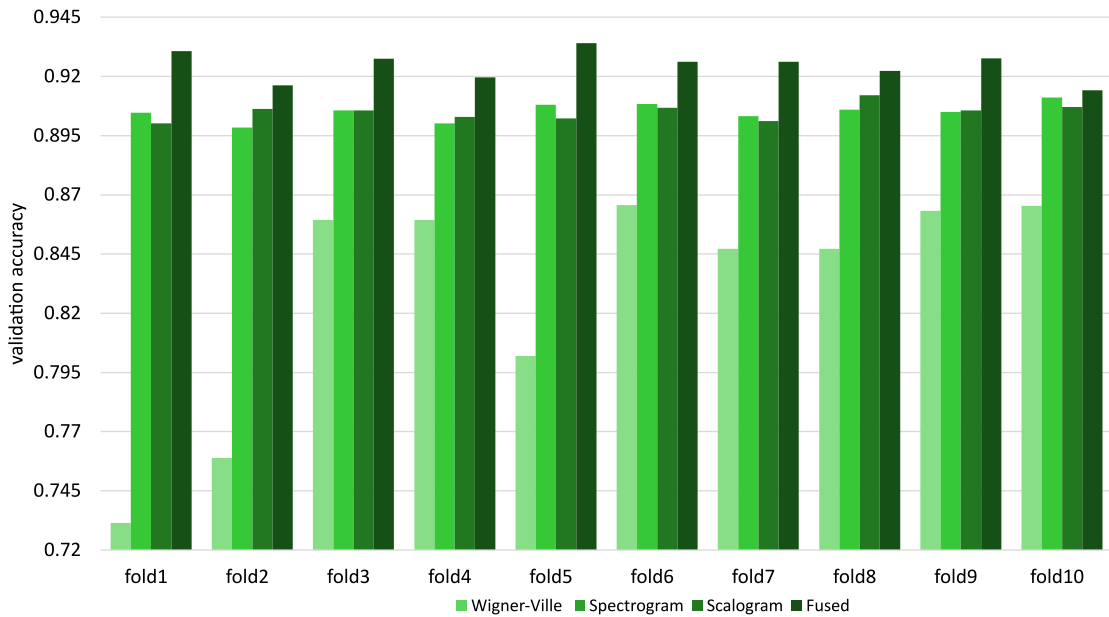


Figure 4.9: Distributions of validation accuracy for TFR images and fused images over 10 folds.

Figs 4.9 and 4.10 show the validation accuracies for each fold during the 10-fold cross validation and the confusion matrices, respectively. Fig. 4.11 shows interquartile range (IQR) plots for performance metrics calculated across all folds.

Table 4.1 demonstrates the overall macro average of few performance metrics for the proposed CNN model trained with 10-fold cross validation technique. The Table 4.1 clearly shows that the performance measures are very similar for all image types other than the Wigner–Ville

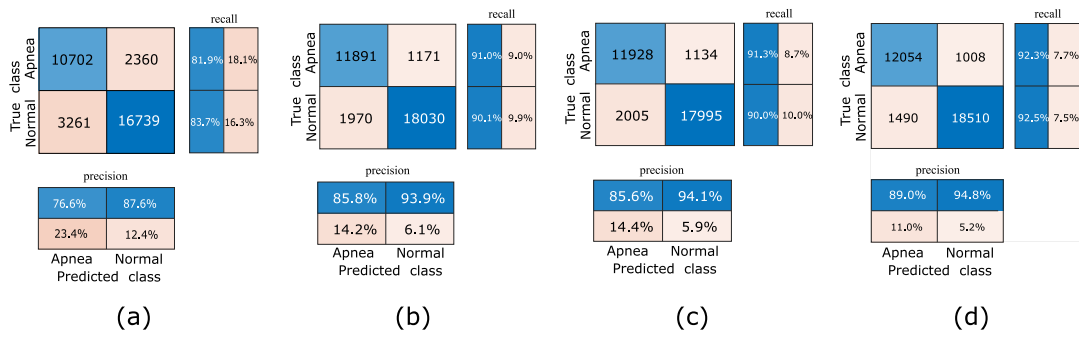


Figure 4.10: Confusion matrices for per-segment apnea detection, with classwise PR and RE shown in the bottom and right-hand boxes, respectively: (a) Wigner–Ville distribution images, (b) Scalogram images, (c) Spectrogram images, and (d) Fused images.

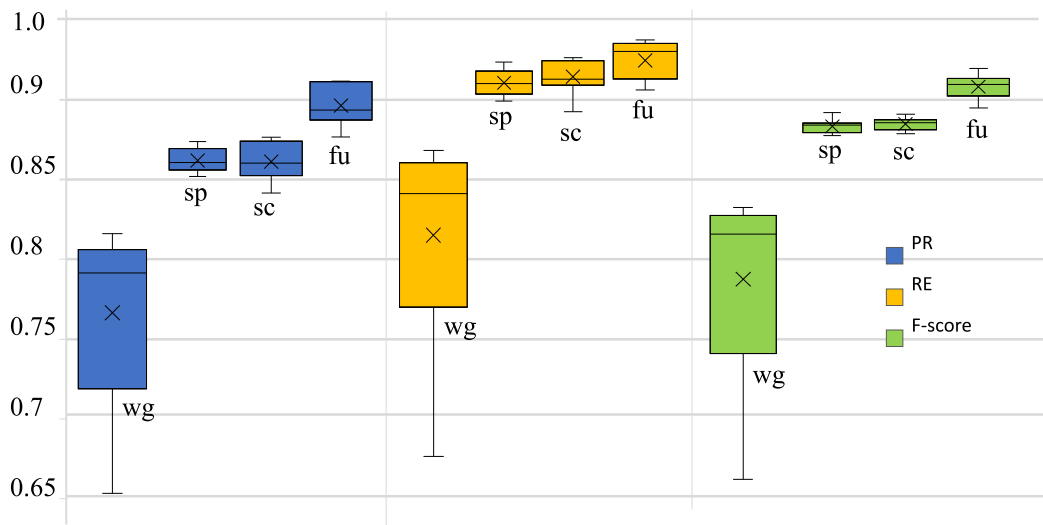


Figure 4.11: IQR plots of PR, RE, and F1 for apnea detection obtained across all folds. The center line indicates the median, the box limits indicate the upper and lower quartiles, the whiskers indicate $1.5 \times \text{IQR}$, and \times indicates the mean. The images are Wigner–Ville distribution images (wg), scalogram images (sc), spectrogram images (sp), or fused images (fu).

distribution images.

When considering the validation accuracy for all folds, as shown in Fig. 4.9, there is no great variation between the 10 folds. This observation indicates that the proposed CNN model can be generalized.

As shown in Fig. 4.11, the rest of the performance metrics, including the F_1 score, also show very small variation across the folds. Although the means of the performance metrics show nearly identical for scalogram and spectrogram images (cwt and STFT cases), the performance metrics show slightly higher variability across the folds for scalogram images. The overall ac-

Table 4.1: Overall performance in per-segment apnea detection TFR images and fused images

	accuracy %	PR %	RE %	SP %	F1 %
Wigner distribution	82.9	76.6	81.9	83.7	79.2
Scalogram images	90.5	85.8	91.0	90.2	88.3
Spectrogram images	90.5	85.6	91.3	90.0	88.4
Fused images	92.4	89.0	92.3	92.6	90.6

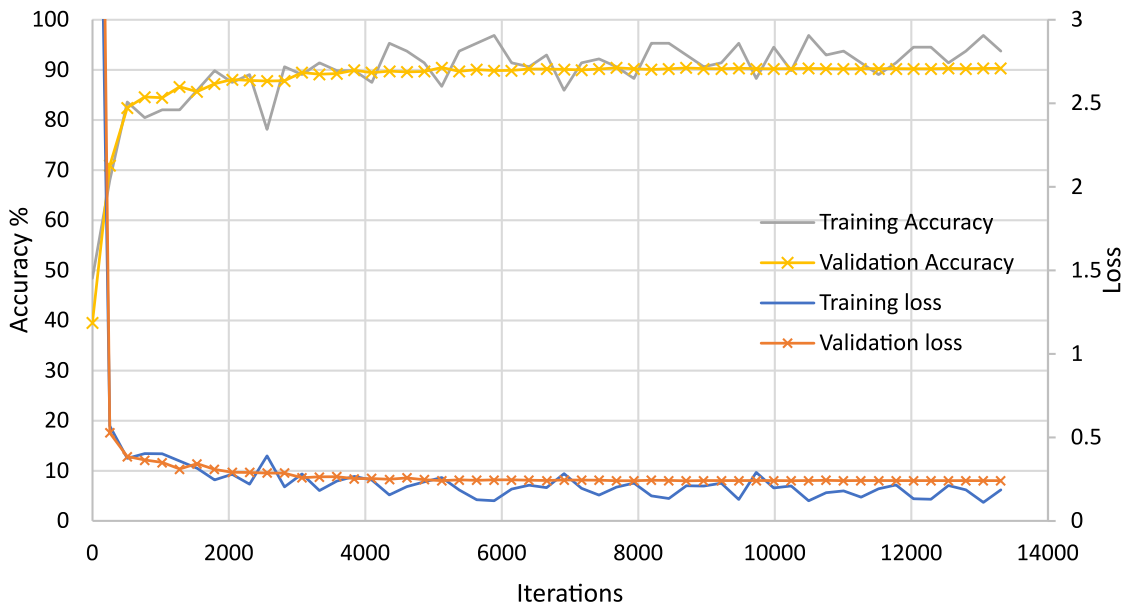


Figure 4.12: Accuracy-loss graph of the proposed CNN (for the lowest-performing model).

accuracy and F1 scores for per-segment OSA detection with scalogram images were calculated as 90.5 % and 88.3 %, respectively. The same measures for the spectrogram images are 90.5 % and 88.4 %. However, the proposed CNN model showed the highest performance for the fused images, achieving 92.4 % overall accuracy and a 90.6 % F1 score. The variability of all measures for the fused images is slightly high compared to corresponding scalogram and spectrogram images. The lowest performance is observed for the Wigner–Ville distribution images for all in all performance metrics , with the greatest variation in across the folds.

Fig. 4.12 shows an accuracy loss profile for the weakest classifier of fused images. The learning curves confirm that the parameters selected for image creation and the CNN are appropriate for discriminating between OSA and normal ECG segments.

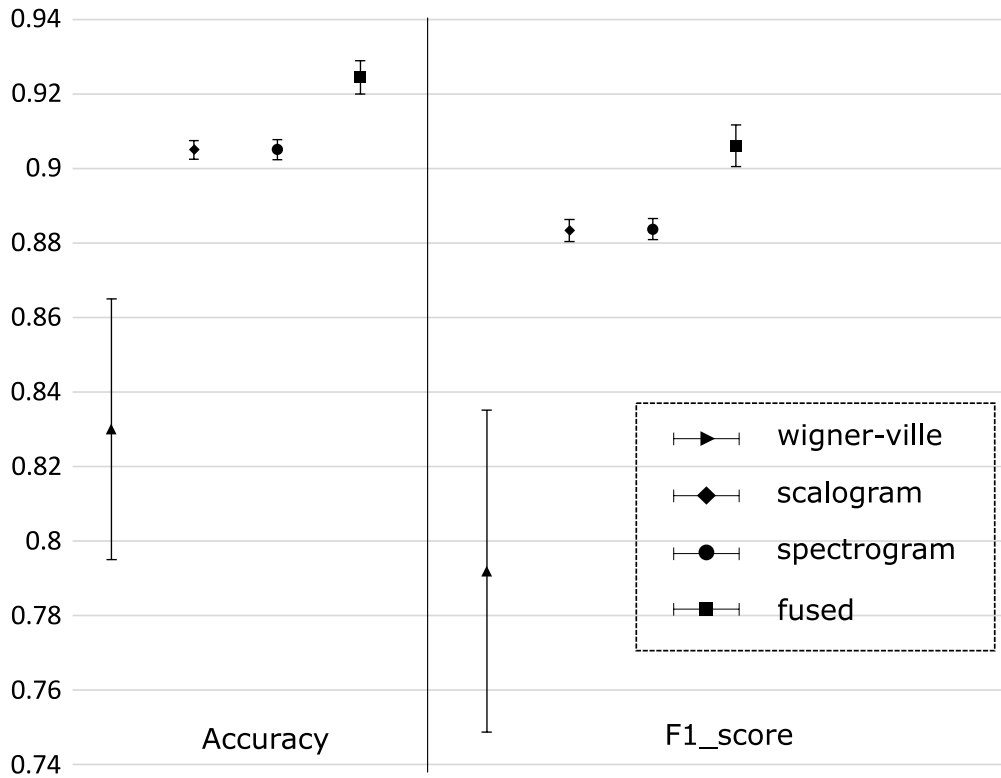


Figure 4.13: Overall 10-fold cross-validation results for per-segment apnea detection with Wigner–Ville distribution, scalogram, spectrogram, and fused images. Black lines indicate the corresponding 95 % confidence interval.

4.2.1 Robustness evaluation

The PhysioNet Apnea-ECG database is a relatively small dataset, with 70 subjects where withheld dataset and the training dataset containing 35 recordings each. Therefore, using a single withheld dataset for validation might be unfair, given that this work focuses on training a deep-learning model. To train such a CNN, more data is required compared to other machine-learning methods. As pointed out in the method section, 10-fold cross validation is employed to test the robustness of the proposed CNN with the entire dataset (70 recordings), which was randomly divided into 10 subsets, as shown in Fig. 4.8. Fig. 4.13 shows the average accuracy and F1 score for per-segment OSA detection, with their 95 % confidence intervals (CIs) calculated for 10 cross-validation steps. According to Fig. 4.13, the proposed CNN model shows quite consistent performance for all image types (excluding the Wigner–Ville distribution) in terms of validation accuracy and F1 score with a small 95 % CI. The model obtained Overall accu-

racies of $90.5 \pm 0.3 \%$, $90.5 \pm 0.3 \%$, and $92.4 \pm 0.5 \%$ for per-segment OSA detection using scalogram, spectrogram, and fused images, respectively. Similarly, the F1 scores were recorded as $88.3 \pm 0.3 \%$, $88.4 \pm 0.3 \%$, and $90.6 \pm 0.6 \%$, for the scalogram, spectrogram, and fused cases, respectively. However, the pseudo Wigner–Ville images, showed significantly low performance (accuracy = $82.99 \pm 3.49 \%$ and F1 score = $79.19 \pm 7.31 \%$) because the Wigner–Ville distribution method has cross-term issues.

4.2.2 Comparison with existing methods

Since the PhysioNet Apnea-ECG database has been available for some time, many automatic OSA-detection methods can be seen in the literature. Here, we compare our proposed method with those that also employed the PhysioNet Apnea-ECG database. However per-recording detection performance is not compared as this CNN is trained using 10-fold cross validation after aggregating the entire dataset.

Table 4.2 summarizes the overall performance of the proposed CNN method relative to other works, with respect to per-segment OSA detection. As shown in the table, the proposed CNN model has achieved the best performance in terms of overall accuracy, sensitivity, and specificity. In particular, our method can be compared with that of Tao Wangs *et al.* work [151]. Their model has showed an overall accuracy, sensitivity, and specificity of 87.3 %, 85.1 %, and 88.7 %, respectively for the withheld dataset, whereas the proposed method has shown an average accuracy, sensitivity, and specificity of 92.4 %, 92.3 %, and 92.6 %, respectively. This a significant performance improvement compared to the other works with the same dataset. On the other hand, in the sense of the robustness evaluation, this study has improved the overall accuracy by $\approx 6 \%$, with a smaller CI ($\approx \pm 0.45 \%$) (for 10-fold cross validation) than their $\pm 1.5 \%$ evaluated using 7-fold cross validation for the full dataset.

Additionally, the proposed CNN model has the best detection confidence, as demonstrated in Table 4.2 for all performance metrics comparing the works [60, 75, 81, 82] whereas the same dataset was employed. Similar to this study Singh *et al.* [111] has proposed an image-based OSA-detection method that used a CNN model based on AlexNet. This model has shown a

validating accuracy of 86.2 % , a sensitivity of 90 % , and a specificity of 83.8 % using scalogram images (227, 227, 3). Although their model has good sensitivity compared to this study, the proposed model has obtained much better performance in the senses of validation accuracy and specificity. Their classification model is a plain DNN model with 5 CNN layers compared with the proposed better-performing classification model containing four residual blocks.

In addition to the aforementioned methods, other recent works [65, 75, 84, 85] have also performed well for this dataset, however the proposed approach has shown better performance metrics since the proposed classifier adopts recent advances in deep learning which enables the most appropriate features to be extracted automatically to perform the classification. In these other works, specific features are pulled out using the EDR or ECG signals, forcing the model to depend on manually extracted features. However, it should be emphasis that deep learning models need balanced datasets for optimal performance, unlike other classical machine learning models. Another possible negative point of the proposed methodology is that it requires conversion of the time-domain signal into two separate TFRs in order to be used in the model to perform its predictions.

The proposed CNN model performs comparatively well since it is trained using the whole dataset with 10-fold cross validation procedure. This technique avoids the model being overfitting for set of data and provides greater sample variation in the model training. In addition, the apnea images are randomly over sampled and subjected to image augmentation, which help to mitigate the class imbalance and improve the number of examples by creating modified versions of the images, (all the image types including fused images are subjected to this data over-sampling and augmenting). The augmented training dataset helps to create skillful models and improves unreliability of the model for unseen ECG segments. Most importantly, the fused images used in this work provide a nice blend of discriminative features where both CWT and STFT features are hybridized in an RGB image.

In contrast to other methods, note that the proposed CNN model classifies OSA ECG segments without separating the QRS complexes in ECG signals. This is quite a significant feature compared to other works which provides a robustness for detecting OSA. Moreover, the

Table 4.2: Performance comparison of proposed and previous methods for per-segment apnea detection for the same data set

Reference	Method	Validation	Acc (%)	Sen(%)	Spe(%)
Viswabhargav <i>et al.</i> (2019) [65]	SRE features with different dictionaries (SVM)	10-fold (SVM)	-	78	78.1
		Subject-Specific (SVM)	-	85.4	92.6
Tripathy <i>et al.</i> (2020) [84]	cardio-pulmonary signal / bivariate fast and adaptive EMD coupled with cross time-frequency analysis (SVM)	10-fold (SVM) subject-specific (SVM)	-	73.2	73.1
			-	82.3	78.7
Singh <i>et al.</i> (2020) [85]	Instantaneous amplitude and instantaneous frequency based features/ EDR and HRV signals	10-fold (SVM- RBF) Leave-One-Out(DNN)	-	82.4	79.7
Sharma <i>et al.</i> (2016) [75]	Hermite basis functions	10-fold (LS-SVM(RBFKernel))	83.8	79.5	88.4
Song <i>et al.</i> (2016) [81]	HMM-SVM	10-fold	86.2	82.6	88.4
Varon <i>et al.</i> (2015) [60]	LS-SVM	fixed-size method	84.7	84.7	84.7
Li <i>et al.</i> (2018) [82]	EDR signal NN and HMM	withheld dataset	84.7	88.9	82.1
Singh <i>et al.</i> (2019) [111]	scalogram images (Morle wavelet) , DNN and Decision fusion	withheld dataset	86.2	90	83.8
Wang <i>et al.</i> (2019) [151]	Time window and ANN	7-fold	87.3	85.1	88.7
proposed method	smoothed pseudo Wigner-Ville, CNN scalogram images (Morse wavelet), CNN spectrogram images, CNN fused images, CNN	10-fold	82.9	81.9	83.7
			90.5	91.04	90.2
			90.51	91.32	89.98
			92.4	92.3	92.6

CNN model can be adapted to detect apneic episodes using arbitrarily long ECG segments (e.g., 10 s or 20 s) because RGB images can be generated for any length ECG-segment and resized to (128, 128, 3) in order to be used in the model.

The purpose of this work is to implement a robust automatic OSA-detection classifier based on fused spectral images of TFRs based on TFRs. The proposed CNN model detects OSA episodes using images corresponding to one-minute ECG segments with TFRs including Wigner–Ville distribution, scalogram, spectrogram, or fused-image formats. The results shown for accuracy and other performance metrics clearly demonstrate that proposed model not only picks apneic episodes automatically but also outperforms previous works in the sense of automatic OSA classification. The proposed classification model achieved an overall accuracy of 92.4 % for fused spectral images generated from scalogram and spectrogram images. Another important aspect of this study is that the ability of employing arbitrary long ECG-segments since the ECG segments are converted image form before utilizing the classification model. Moreover, there is no manual feature extraction is needed, which depends on the experience and specific domain knowledge in the relevant fields. Since the proposed model based on a single-lead ECG channel, the model can be implemented in wearable electronics or a smart home-monitoring systems easily. Therefore, this methodology would be cheaper and more convenient than having to use a conventional sleep-study lab environment in the sense of both data collecting and analyzing the ECG signals. However, the proposed methodology has few limitations. Since the the PhysioNet Apnea-ECG database has only two types of annotations (apnea and normal), the proposed CNN model is not capable of classifying apnea sub types (e.g., hypopnea). In future works, it is expected to extend the proposed model to label these different types of apnea. In addition, proposed approach can be modified utilizing multiple apnea datasets. Investigating different fusing techniques is also a good option to improve the performance in the future works.

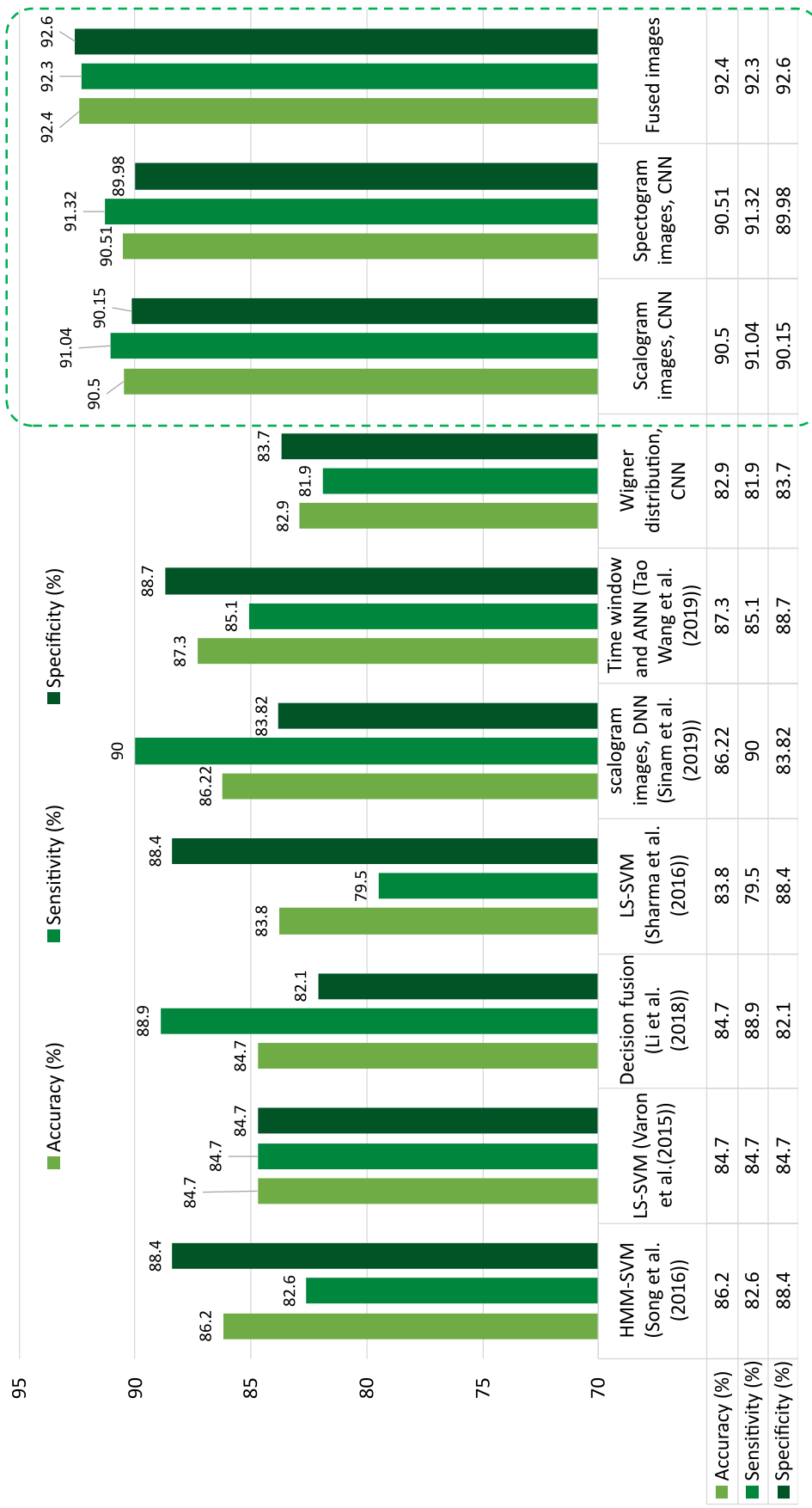


Figure 4.14: Results comparison with existing methods

5

Discussion

5.1 Sleep stage classification

The results of the experiments performed in three experiments confirmed that proposed CNN models can perform automatic sleep stage classification at good performance for 30 s of epochs. We observed that the performance of the proposed models may slightly vary in between the subjects. However, most of our experiments were performed with PSG data from sleep patients. Therefore, the models are likely to be performed well for sleep patients.

On the other hand, experiment II and experiment III models can be adapted for different channel configuration depending on the channel availability. Especially, alternative configurations of the models can be used to score sleep epochs, when PSG channels are not available or

affected due to the movement of the subjects. The performance are comparatively lower in alternative configurations like single channel model and three channel models.

However, the alternative models can be used as a computer assistance tool for sleep stage scoring. These models are trained without much pre-processing which make the models robust. Therefore, raw PSG data can be directly used in these models. All the models are vigorous since the feature are automatically extracted without using hand engineered features. The model also can be trained by other datasets with different data distribution since the features extraction sections can be retrained.

5.2 OSAS detection

The main aim of this work is to find a robust solution for automatic OSA detection method. In this work, convolution neural network model based on residual network is presented for detecting OSA in per segment basis employing fused spectral images created from different time frequency representations. The proposed residual CNN can label OSA episodes using either scalogram, spectrogram or Wigner-Ville based spectral images correspond to 1-minute ECG fragments. The overall accuracy and other performance metrics demonstrated that the proposed classification model can accurately identify apneic cases with comparison to previous studies works. The overall accuracy shown for fused images was 92.4 %.

This model can be adapted to used for arbitrary long ECG segments.this is possible because the ECG segment is converted to RGB image before feeding to the prediction model. Besides that, there is no manual feature extraction involved in this approach, which is susceptible to the experience and specific domain knowledge of the researcher. Since our model is based on single lead ECG channel, this model can be used with wearable electronics and smart devises as a home monitoring system which saves a lot of time and money compared to expensive conventional sleep study. However, our work still suffer from few limitations, since the PhysioNet Apnea-ECG dataset has only two annotations (apnea and normal).

6

Conclusion and future works

6.1 Main contributions

The main contributions of this thesis works are published in a refereed journal and international conferences. The publications that comprise this study within the scope of this dissertation are listed as follows:

1. *Novel multi-channel 2D convolutional neural network approach for sleep stage classification*

2D CNN based image classification principles were adapted for processing multi-channel 1D PSG signals. A multi-branch approach is proposed to perform effective feature extraction.

2. *Novel 1D CNN-RNN deep neural architectures based on multi resolution feature extraction approach for 4-stage classification*

A 1D CNN-RNN model is proposed to perform automatic 4-stage classification. The model can be adapted for multi-channel and single channel configurations. Combination of several CNN branches followed by RNN layers were used to perform the sleep stage classification.

3. *Novel 1D CNN-RNN deep neural architectures based on multi resolution feature extraction approach for 4-stage and 5-stage classification*

A 1D CNN-RNN model is proposed to perform automatic sleep classification for both 4-stage and 5-stage classification. The model also can be adapted for multi-channel and single channel configurations. Combination of several CNN branches followed by RNN layers were used to perform the the feature extraction with an extra outer CNN layer.

4. *Novel method for OSAS detection based on deep neural networks and fused Time-Frequency representations of electrocardiogram signals*

A novel fused-image-based technique that detects OSAS using only a single-lead ECG signal was proposed. In the proposed approach, a CNN extracts Time-Frequency features automatically from fused spectral images created with one-minute ECG segments. In this study, three time–frequency representations, namely the scalogram, the spectrogram, and the Wigner–Ville distribution, were used to investigate the effectiveness of the fused-image-based approach. We found that blending scalogram and spectrogram images has the best performances as it has discriminative characteristics compared to its normal form.

The presented main contributions were published in a refereed journal and two international conferences. The publications that comprise the work carried out within the scope of this dissertation are as follows:

- *Automatic Sleep Stage Classification Based on Convolutional Neural Networks* (Chapter 2 and Chapter 3). 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech), Osaka, Japan, 2019, pp. 275-276, doi: 10.1109/LifeTech.2019.8883961. [152].
- *Sleep Stage Classification Based on EEG, EOG, and CNN-GRU Deep Learning Model* (Chapter 2 and Chapter 3). 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-7, doi: 10.1109/ICAwST.2019.8923359.[153]
- *A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network* (Chapters Chapter 2 and Chapter 4). PLOS One, 2021, <https://doi.org/10.1371/journal.pone.0250618> [154].

6.2 Sleep stage classification

To address practical complications in manual sleep stage scoring, an accurate and generalized automatic sleep stage scoring system is required. One of the main objective of this thesis is to develop a computer assisting tool for sleep stage scoring. Mainly, three approaches are presented to predict sleep stages for over night PSG recordings. Some models are trained with healthy subjects and some are trained for subjects who has sleep disorders. All three proposed models can be used with multiple electrodes configuration and two of them can be used with single channel configurations. All the model has less parameters compared to the other methods. Therefore these models can be easily implemented on small devices such as smart phones.

In order to improve the performance even more, I anticipate to conduct more experiments with other data sets. It is expected to utilize time frequency analysis together with deep learning to improve the performance further more. On the other hand, I expect to implement dedicated models for single channel configuration. And also, there is a need for developing real time sleep stage detection system for immediate clinical diagnosis.

6.3 OSAS detection

The main aim of this work is to find a robust solution for automatic OSA detection method. In this work, Convolution neural network model based on residual network is presented for detecting OSA in per segment basis employing fused spectral images created from different time frequency representations. The proposed residual CNN can label OSA episodes using either scalogram, spectrogram or Wigner-Ville based spectral images correspond to 1-minute ECG fragments. The overall accuracy and other performance metrics demonstrated that the proposed classification model can accurately identify apneic cases with comparison to previous studies works. The overall test accuracy shown for the fused images was 92.4 %.

Based on this analysis and the knowledge formed in this study, the model can be adapted for arbitrary lengths of ECG segments as explained in the discussion section of Chapter 4. Similarly, this kind of techniques can be adopted for other classification models where other types of physiological signals are involved (i.e EEG EOG) as the fused spectral image can be generated with any length of signal segment.

Most importantly, the fused image based approach removes the need of generating manual feature extraction which is not realistic to use for large data sets. Basically, the main underlying problems have been emphasized in the Chapter 4, especially the increasing need of accurate sleep apnea detection technology. On the other hand, the proposed method does not rely on the sleep-expert variability in visual OSA detection based on PSG. This is a common limitation seen in the conventional OSA detection based on PSG studies producing subjective and unreliable results in detecting sleep related issues.

However, this work still needs future implementation to provide improved performance in terms of detecting more apnea types as detailed in Chapter 4 since the PhysioNet Apnea-ECG dataset provide only apneic and non apneic events. Therefore, our model can not perform multi-type sleep apnea detection (i.e. hypopnea and apnea), unfortunately. In future works, I anticipate to improve the CNN model so that it can identify other types of apnea episodes. Furthermore, the proposed approach can be improved using multiple apnea data sets. Inves-

Investigating new fusion algorithms also will be a good option to improve the performances of the proposed method.

Acknowledgments

I am grateful to my supervisor, Senior Associate Professor. Zhu Xin, for his tremendous support and understanding throughout this study. And also I wish to express sincere appreciation to my co-supervisor, Professor Chen Wenxi for his valuable guidance and cooperation. And also, I would like to express gratitude to rest of my dissertation committee members, Professors Truong Cong-Thang and Okuyama Yuichi for providing their perception and ideas into my thesis study.

I gratefully acknowledge the financial support of the Ministry Education of Japan, The university of Aizu, and Japan student service Organization (JASSO) . I would also like to thank doctors, nurses, and technicians at Fukushima Otsuki Clinic, Fukushima, Japan for their valuable efforts in collecting and annotating the data used in the study. I am grateful to all citizens of Japan who paid taxes to provide the finance support for the research community.

Most importantly I appreciate the help from friends and colleagues, for enriching my academic life as well as my social life in past three years. Especially I respectfully, thank to my old friend Dr.Isuru Jayarathne for his priceless support during my stay in Japan. Moreover, I would like to acknowledge the staff of The university of Aizu for their tremendous and kind support in various issues, and the Japanese friends from Aizu-wakamatzu city for making my life memorable and comfortable during my stay in Aizu.

Finally, I am indebted to my parents and fiancée Samadhie for all the love and support given to me, and for their patience. Finally, I would like to dedicate this work to my parents and

Samadhi.

Senevirathna Mudiyansele Isuru Niroshana

14/04/2021

References

- [1] H. R. Smith, C. L. Comella, and B. Högl, *Sleep medicine*. Cambridge University Press, 2008.
- [2] C. Iber, S. Ancoli-Israel, A. L. Chesson, S. F. Quan *et al.*, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine Westchester, IL, 2007, vol. 1.
- [3] J. V. Rundo and R. Downey III, “Polysomnography,” *Handbook of clinical neurology*, vol. 160, pp. 381–392, 2019.
- [4] J. M. Krueger, D. M. Rector, S. Roy, H. P. Van Dongen, G. Belenky, and J. Panksepp, “Sleep as a fundamental property of neuronal assemblies,” *Nature Reviews Neuroscience*, vol. 9, no. 12, pp. 910–919, 2008.
- [5] R. Wolk, A. S. Gami, A. Garcia-Touchard, and V. K. Somers, “Sleep and cardiovascular disease,” *Current problems in cardiology*, vol. 30, no. 12, pp. 625–662, 2005.
- [6] S. Najdi, A. A. Gharbali, and J. M. Fonseca, “Feature transformation based on stacked sparse autoencoders for sleep stage classification,” in *Doctoral Conference on Computing, Electrical and Industrial Systems*. Springer, 2017, pp. 191–200.
- [7] B. M. Altevogt, H. R. Colten *et al.*, *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006.
- [8] Z.-J. Cai, “The functions of sleep: further analysis,” *Physiology & behavior*, vol. 50, no. 1, pp. 53–60, 1991.
- [9] N. C. Rattenborg, H. O. de la Iglesia, B. Kempnaers, J. A. Lesku, P. Meerlo, and M. F. Scriba, “Sleep research goes wild: new methods and approaches to investigate the ecology, evolution and functions of sleep,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1734, p. 20160251, 2017.
- [10] J. W. Shepard, D. J. Buysse, A. L. Chesson, W. C. Dement, R. Goldberg, C. Guilleminault, C. D. Harris, C. Iber, E. Mignot, M. M. Mitler *et al.*, “History of the development of sleep medicine in the United States,” *Journal of clinical sleep medicine*, vol. 1, no. 01, pp. 61–82, 2005.
- [11] T. Penzel, P.-G. Behler, M. Von Buttlar, R. Conradt, M. Meier, A. Moller, and H. Danker-Hopfe, “Reliability of visual evaluation of sleep stages according to Rechtschaffen and Kales from eight polysomnographs by nine sleep centres,” *SOMNOLOGIE-BERLIN*, vol. 7, no. 2, pp. 49–58, 2003.

- [12] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S.-L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn *et al.*, “Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders,” *Journal of sleep research*, vol. 13, no. 1, pp. 63–69, 2004.
- [13] R. S. Rosenberg and S. Van Hout, “The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring,” *Journal of clinical sleep medicine*, vol. 9, no. 01, pp. 81–87, 2013.
- [14] J. A. Hobson, “A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: A. Rechtschaffen and A. Kales (Editors). (Public Health Service, US Government Printing Office, Washington, DC, 1968, 58 p., \$4.00),” 1969.
- [15] S.-L. Himanen and J. Hasan, “Limitations of rechtschaffen and kales,” *Sleep medicine reviews*, vol. 4, no. 2, pp. 149–167, 2000.
- [16] B. J. Swihart, B. Caffo, K. Bandeen-Roche, and N. M. Punjabi, “Characterizing sleep structure using the hypnogram,” *Journal of Clinical Sleep Medicine*, vol. 4, no. 4, pp. 349–355, 2008.
- [17] A. Domingues, T. Paiva, and J. M. Sanches, “Hypnogram and sleep parameter computation from activity and cardiovascular data,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1711–1719, 2014.
- [18] M. J. Lavery, C. Stull, M. O. Kinney, and G. Yosipovitch, “Nocturnal pruritus: the battle for a peaceful night’s sleep,” *International journal of molecular sciences*, vol. 17, no. 3, p. 425, 2016.
- [19] R. Boostani, F. Karimzadeh, and M. Nami, “A comparative review on sleep stage classification methods in patients and healthy individuals,” *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.
- [20] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
- [21] B. Koley and D. Dey, “An ensemble system for automatic sleep stage classification using single channel EEG signal,” *Computers in biology and medicine*, vol. 42, no. 12, pp. 1186–1195, 2012.
- [22] L. J. Herrera, C. M. Fernandes, A. M. Mora, D. Migotina, R. Largo, A. Guillén, and A. C. Rosa, “Combination of heterogeneous EEG feature extraction methods and stacked sequential learning for sleep stage classification,” *International journal of neural systems*, vol. 23, no. 03, p. 1350012, 2013.
- [23] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, “Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier,” *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

- [24] A. Krakovská and K. Mezeiová, “Automatic sleep scoring: A search for an optimal combination of measures,” *Artificial intelligence in medicine*, vol. 53, no. 1, pp. 25–33, 2011.
- [25] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, and F. Chapotot, “Feature selection for sleep/wake stages classification using data driven methods,” *Biomedical Signal Processing and Control*, vol. 2, no. 3, pp. 171–179, 2007.
- [26] B. Koley and D. Dey, “An ensemble system for automatic sleep stage classification using single channel EEG signal,” *Computers in biology and medicine*, vol. 42, no. 12, pp. 1186–1195, 2012.
- [27] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, “A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal,” *Expert Systems with Applications*, vol. 104, pp. 277–293, 2018.
- [28] M. Sharma, D. Goyal, P. Achuth, and U. R. Acharya, “An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank,” *Computers in biology and medicine*, vol. 98, pp. 58–75, 2018.
- [29] S. Crisler, M. J. Morrissey, A. M. Anch, and D. W. Barnett, “Sleep-stage scoring in the rat using a support vector machine,” *Journal of neuroscience methods*, vol. 168, no. 2, pp. 524–534, 2008.
- [30] S. Güneş, K. Polat, and Ş. Yosunkaya, “Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 7922–7928, 2010.
- [31] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [32] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, “Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier,” *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.
- [33] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, “Classification of sleep stages using multi-wavelet time frequency entropy and LDA,” *Methods of information in Medicine*, vol. 49, no. 03, pp. 230–237, 2010.
- [34] F. Chapotot and G. Becq, “Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules,” *International Journal of Adaptive Control and Signal Processing*, vol. 24, no. 5, pp. 409–423, 2010.
- [35] S. Özşen, “Classification of sleep stages using class-dependent sequential feature selection and artificial neural network,” *Neural Computing and Applications*, vol. 23, no. 5, pp. 1239–1250, 2013.

- [36] M. E. Tagluk, N. Sezgin, and M. Akin, “Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG,” *Journal of medical systems*, vol. 34, no. 4, pp. 717–725, 2010.
- [37] C.-S. Huang, C.-L. Lin, L.-W. Ko, S.-Y. Liu, T.-P. Sua, and C.-T. Lin, “A hierarchical classification system for sleep stage scoring via forehead EEG signals,” in *2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*. IEEE, 2013, pp. 1–5.
- [38] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, “A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.
- [39] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
- [40] Z. Cui, X. Zheng, X. Shao, and L. Cui, “Automatic sleep stage classification based on convolutional neural network and fine-grained segments,” *Complexity*, vol. 2018, 2018.
- [41] S. Mousavi, F. Afghah, and U. R. Acharya, “SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach,” *PloS one*, vol. 14, no. 5, 2019.
- [42] N. Michielli, U. R. Acharya, and F. Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals,” *Computers in biology and medicine*, vol. 106, pp. 71–81, 2019.
- [43] V. Bajaj and R. B. Pachori, “Automatic classification of sleep stages based on the time-frequency image of EEG signals,” *Computer methods and programs in biomedicine*, vol. 112, no. 3, pp. 320–328, 2013.
- [44] J. Shi, X. Liu, Y. Li, Q. Zhang, Y. Li, and S. Ying, “Multi-channel EEG-based sleep stage classification with joint collaborative representation and multiple kernel learning,” *Journal of neuroscience methods*, vol. 254, pp. 94–101, 2015.
- [45] B. Koley and D. Dey, “An ensemble system for automatic sleep stage classification using single channel EEG signal,” *Computers in biology and medicine*, vol. 42, no. 12, pp. 1186–1195, 2012.
- [46] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, and C.-Y. Hsu, “Automatic sleep stage recurrent neural classifier using energy features of EEG signals,” *Neurocomputing*, vol. 104, pp. 105–114, 2013.
- [47] Ş. Yücelbaş, C. Yücelbaş, G. Tezel, S. Özşen, and Ş. Yosunkaya, “Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal,” *Expert Systems with Applications*, vol. 102, pp. 193–206, 2018.

- [48] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, "Sleep stages classification based on heart rate variability and random forest," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 624–633, 2013.
- [49] K. Kesper, S. Canisius, T. Penzel, T. Ploch, and W. Cassel, "ECG signal analysis for the assessment of sleep-disordered breathing and sleep pattern," *Medical & biological engineering & computing*, vol. 50, no. 2, pp. 135–144, 2012.
- [50] H. Yoon, S. H. Hwang, J.-W. Choi, Y. J. Lee, D.-U. Jeong, and K. S. Park, "REM sleep estimation based on autonomic dynamics using R–R intervals," *Physiological measurement*, vol. 38, no. 4, p. 631, 2017.
- [51] S. J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 4, pp. 245–256, 2007.
- [52] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini-Strambi, S. Cerutti, and A. Bianchi, "Sleep staging from heart rate variability: time-varying spectral features and hidden Markov models," *International Journal of Biomedical Engineering and Technology*, vol. 3, no. 3-4, pp. 246–263, 2010.
- [53] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiological measurement*, vol. 36, no. 10, p. 2027, 2015.
- [54] V. F. Helland, A. Gapelyuk, A. Suhrbier, M. Riedl, T. Penzel, J. Kurths, and N. Wessel, "Investigation of an automatic sleep stage classification by means of multiscored hypnogram," *Methods of information in medicine*, vol. 49, no. 05, pp. 467–472, 2010.
- [55] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. Van Huffel, B. Haex, and J. Vander Sloten, "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 661–669, 2013.
- [56] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients," *Sleep medicine*, vol. 2, no. 5, pp. 389–396, 2001.
- [57] S. Quan, J. C. Gillin, M. Littner, and J. Shepard, "Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. editorials," *Sleep (New York, NY)*, vol. 22, no. 5, pp. 662–689, 1999.
- [58] N. M. Punjabi, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 136–143, 2008.
- [59] J. Gubbi, A. Khandoker, and M. Palaniswami, "Classification of sleep apnea types using wavelet packet analysis of short-term ECG signals," *Journal of clinical monitoring and computing*, vol. 26, no. 1, pp. 1–11, 2012.

- [60] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ECG," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2269–2278, 2015.
- [61] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE transactions on information technology in biomedicine*, vol. 13, no. 1, pp. 37–48, 2008.
- [62] J. Coleman, "Complications of snoring, upper airway resistance syndrome, and obstructive sleep apnea syndrome in adults." *Otolaryngologic Clinics of North America*, vol. 32, no. 2, pp. 223–234, 1999.
- [63] F. J. Nieto, T. B. Young, B. K. Lind, E. Shahar, J. M. Samet, S. Redline, R. B. D'agostino, A. B. Newman, M. D. Lebowitz, T. G. Pickering *et al.*, "Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study," *Jama*, vol. 283, no. 14, pp. 1829–1836, 2000.
- [64] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American journal of epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.
- [65] C. S. Viswabhargav, R. Tripathy, and U. R. Acharya, "Automated detection of sleep apnea using sparse residual entropy features with various dictionaries extracted from heart rate and EDR signals," *Computers in biology and medicine*, vol. 108, pp. 20–30, 2019.
- [66] W. R. Ruehland, P. D. Rochford, F. J. O'Donoghue, R. J. Pierce, P. Singh, and A. T. Thornton, "The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index," *sleep*, vol. 32, no. 2, pp. 150–157, 2009.
- [67] H. D. Nguyen, B. A. Wilkins, Q. Cheng, and B. A. Benjamin, "An online sleep apnea detection method based on recurrence quantification analysis," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1285–1293, 2013.
- [68] N. Sadr and P. de Chazal, "A comparison of three ECG-derived respiration methods for sleep apnoea detection," *Biomedical Physics & Engineering Express*, vol. 5, no. 2, p. 025027, 2019.
- [69] A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. Clifford, "A review of signals used in sleep analysis," *Physiological measurement*, vol. 35, no. 1, p. R1, 2013.
- [70] S. Boudaoud, H. Rix, O. Meste, C. Heneghan, and C. O'Brien, "Corrected integral shape averaging applied to obstructive sleep apnea detection from the electrocardiogram," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, 2007.

- [71] T. Penzel, J. McNames, P. De Chazal, B. Raymond, A. Murray, and G. Moody, "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," *Medical and Biological Engineering and Computing*, vol. 40, no. 4, pp. 402–407, 2002.
- [72] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [73] L. Bacharova, E. Triantafyllou, C. Vazaios, I. Tomeckova, I. Paranicova, and R. Tkacova, "The effect of obstructive sleep apnea on QRS complex morphology," *Journal of electrocardiology*, vol. 48, no. 2, pp. 164–170, 2015.
- [74] S. Gupta, B. Cepeda-Valery, A. Romero-Corral, A. Shamsuzzaman, V. K. Somers, and G. S. Pressman, "Association between QRS duration and obstructive sleep apnea," *Journal of Clinical Sleep Medicine*, vol. 8, no. 6, pp. 649–654, 2012.
- [75] H. Sharma and K. Sharma, "An algorithm for sleep apnea detection from single-lead ECG using Hermite basis functions," *Computers in biology and medicine*, vol. 77, pp. 116–124, 2016.
- [76] L. Almazaydeh, K. Elleithy, M. Faezipour, and A. Abushakra, "Apnea detection based on respiratory signal classification," *Procedia Computer Science*, vol. 21, pp. 310–316, 2013.
- [77] A. Smruthy and M. Suchetha, "Real-Time Classification of Healthy and Apnea Subjects Using ECG Signals With Variational Mode Decomposition," *IEEE Sensors Journal*, vol. 17, no. 10, pp. 3092–3099, 2017.
- [78] N. Sezgin and M. E. Tagluk, "Energy based feature extraction for classification of sleep apnea syndrome," *Computers in biology and medicine*, vol. 39, no. 11, pp. 1043–1050, 2009.
- [79] C. Guilleminault, R. Winkle, S. Connolly, K. Melvin, and A. Tilkian, "Cyclical variation of the heart rate in sleep apnoea syndrome: mechanisms, and usefulness of 24 h electrocardiography as a screening technique," *The Lancet*, vol. 323, no. 8369, pp. 126–131, 1984.
- [80] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE transactions on information technology in biomedicine*, vol. 13, no. 1, pp. 37–48, 2008.
- [81] C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian, "An Obstructive Sleep Apnea Detection Approach Using a Discriminative Hidden Markov Model From ECG Signals," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1532–1542, 2016.

- [82] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, "A method to detect sleep apnea based on deep neural network and hidden markov model using single-lead ECG signal," *Neurocomputing*, vol. 294, pp. 94–101, 2018.
- [83] J. Hayano, E. Watanabe, Y. Saito, F. Sasaki, K. Fujimoto, T. Nomiya, K. Kawai, I. Kodama, and H. Sakakibara, "Screening for obstructive sleep apnea by cyclic variation of heart rate," *Circulation: Arrhythmia and Electrophysiology*, vol. 4, no. 1, pp. 64–72, 2011.
- [84] R. Tripathy, P. Gajbhiye, and U. R. Acharya, "Automated sleep apnea detection from cardio-pulmonary signal using bivariate fast and adaptive EMD coupled with cross time–frequency analysis," *Computers in Biology and Medicine*, vol. 120, p. 103769, 2020.
- [85] H. Singh, R. K. Tripathy, and R. B. Pachori, "Detection of sleep apnea from heart beat interval and ECG derived respiration signals using sliding mode singular spectrum analysis," *Digital Signal Processing*, vol. 104, p. 102796, 2020.
- [86] P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automatic classification of sleep apnea epochs using the electrocardiogram," in *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*. IEEE, 2000, pp. 745–748.
- [87] P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 6, pp. 686–696, 2003.
- [88] P. de Chazal, T. Penzel, and C. Heneghan, "Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram," *Physiological measurement*, vol. 25, no. 4, p. 967, 2004.
- [89] S. Babaeizadeh, D. P. White, S. D. Pittman, and S. H. Zhou, "Automatic detection and quantification of sleep apnea using heart rate variability," *Journal of electrocardiology*, vol. 43, no. 6, pp. 535–541, 2010.
- [90] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG," *IEEE transactions on information technology in biomedicine*, vol. 15, no. 3, pp. 416–427, 2010.
- [91] B. Xie and H. Minn, "Real-time sleep apnea detection by classifier combination," *IEEE Transactions on information technology in biomedicine*, vol. 16, no. 3, pp. 469–477, 2012.
- [92] D. Liu, X. Yang, G. Wang, J. Ma, Y. Liu, C.-K. Peng, J. Zhang, and J. Fang, "HHT based cardiopulmonary coupling analysis for sleep apnea detection," *Sleep medicine*, vol. 13, no. 5, pp. 503–509, 2012.
- [93] K. Kesper, S. Canisius, T. Penzel, T. Ploch, and W. Cassel, "ECG signal analysis for the assessment of sleep-disordered breathing and sleep pattern," *Medical & biological engineering & computing*, vol. 50, no. 2, pp. 135–144, 2012.

- [94] N. Sadr and P. De Chazal, "Automated detection of obstructive sleep apnoea by single-lead ECG through ELM classification," in *Computing in Cardiology 2014*. IEEE, 2014, pp. 909–912.
- [95] H. D. Nguyen, B. A. Wilkins, Q. Cheng, and B. A. Benjamin, "An online sleep apnea detection method based on recurrence quantification analysis," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1285–1293, 2013.
- [96] A. R. Hassan, "Automatic screening of obstructive sleep apnea from single-lead electrocardiogram," in *2015 international conference on electrical engineering and information communication technology (ICEEICT)*. IEEE, 2015, pp. 1–6.
- [97] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ECG," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 9, pp. 2269–2278, 2015.
- [98] A. R. Hassan, "Computer-aided obstructive sleep apnea detection using normal inverse Gaussian parameters and adaptive boosting," *Biomedical Signal Processing and Control*, vol. 29, pp. 22–30, 2016.
- [99] G. Surrel, A. Aminifar, F. Rincón, S. Murali, and D. Atienza, "Online obstructive sleep apnea detection on medical wearable sensors," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 4, pp. 762–773, 2018.
- [100] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, "A method to detect sleep apnea based on deep neural network and hidden markov model using single-lead ECG signal," *Neurocomputing*, vol. 294, pp. 94–101, 2018.
- [101] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," *PeerJ*, vol. 7, p. e7731, 2019.
- [102] B. S. Chandra, C. S. Sastry, and S. Jana, "Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 3, pp. 710–717, 2018.
- [103] A. Isin and S. Ozdalili, "Cardiac arrhythmia detection using deep learning," *Procedia computer science*, vol. 120, pp. 268–275, 2017.
- [104] M. Hammad, A. M. Ilyasu, A. Subasi, E. S. Ho, and A. A. Abd El-Latif, "A Multitier Deep Learning Model for Arrhythmia Detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2020.
- [105] A. Sedik, A. M. Ilyasu, A. El-Rahiem, M. E. Abdel Samea, A. Abdel-Raheem, M. Hammad, J. Peng, A. El-Samie, E. Fathi, A. A. A. El-Latif *et al.*, "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, no. 7, p. 769, 2020.

- [106] M. Hammad, M. H. Alkinani, B. Gupta, and A. A. Abd El-Latif, "Myocardial infarction detection based on deep neural network on imbalanced data," *Multimedia Systems*, pp. 1–13, 2021.
- [107] A. Alghamdi, M. Hammad, H. Ugail, A. Abdel-Raheem, K. Muhammad, H. S. Khalifa, and A. A. Abd El-Latif, "Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities," *Multimedia tools and applications*, pp. 1–22, 2020.
- [108] Ö. Türk and M. S. Özerdem, "Epilepsy detection by using scalogram based convolutional neural network from EEG signals," *Brain sciences*, vol. 9, no. 5, p. 115, 2019.
- [109] Y.-H. Byeon, S.-B. Pan, and K.-C. Kwak, "Intelligent deep models based on scalograms of electrocardiogram signals for biometrics," *Sensors*, vol. 19, no. 4, p. 935, 2019.
- [110] J. McNames and A. Fraser, "Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram," in *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*. IEEE, 2000, pp. 749–752.
- [111] S. A. Singh and S. Majumder, "A novel approach osa detection using single-lead ECG scalogram based on deep neural network," *Journal of Mechanics in Medicine and Biology*, vol. 19, no. 04, p. 1950026, 2019.
- [112] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Mach. Learn. Technol.*, vol. 2, 2008.
- [113] A. J. Boe, L. L. M. Koch, M. K. O'Brien, N. Shawen, J. A. Rogers, R. L. Lieber, K. J. Reid, P. C. Zee, and A. Jayaraman, "Automating sleep stage classification using wireless, wearable sensors," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–9, 2019.
- [114] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, "On Empirical Comparisons of Optimizers for Deep Learning," *arXiv preprint arXiv:1910.05446*, 2019.
- [115] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [116] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
- [117] T. Wang, C. Lu, and G. Shen, "Detection of Sleep Apnea from Single-Lead ECG Signal Using a Time Window Artificial Neural Network," *BioMed research international*, vol. 2019, 2019.
- [118] L. J. Herrera, A. M. Mora, C. Fernandes, D. Migotina, A. Guillén, and A. C. Rosa, "Symbolic representation of the EEG for sleep stage classification," in *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, 2011, pp. 253–258.

- [119] Y. Li, F. Yingle, L. Gu, and T. Qinye, "Sleep stage classification based on EEG Hilbert-Huang transform," in *2009 4th IEEE Conference on Industrial Electronics and Applications*. IEEE, 2009, pp. 3676–3681.
- [120] A. R. Hassan, S. K. Bashar, and M. I. H. Bhuiyan, "On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram," in *2015 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2015, pp. 2238–2243.
- [121] F. Ebrahimi, M. Mikaeili, E. Estrada, and H. Nazeran, "Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 1151–1154.
- [122] H. Phan, Q. Do, T.-L. Do, and D.-L. Vu, "Metric learning for automatic sleep stage classification," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 5025–5028.
- [123] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.
- [124] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [125] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.
- [126] J. Brownlee, "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks," Jul 2019. [Online]. Available: <https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/>
- [127] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
- [128] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [129] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*. IEEE, 2000, pp. 255–258.
- [130] L. Almazaydeh, K. Elleithy, and M. Faezipour, "Detection of obstructive sleep apnea through ECG signal features," in *2012 IEEE International Conference on Electro/Information Technology*, 2012, pp. 1–6.

- [131] P. de Chazal, T. Penzel, and C. Heneghan, “Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram,” *Physiological measurement*, vol. 25, no. 4, p. 967, 2004.
- [132] E. Sejdic, I. Djurovic, and L. Stankovic, “Quantitative Performance Analysis of Scalogram as Instantaneous Frequency Estimator,” *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3837–3845, 2008.
- [133] A. Ullah, S. M. Anwar, M. Bilal, and R. M. Mehmood, “Classification of Arrhythmia by Using Deep Learning with 2-D ECG Spectral Image Representation,” *Remote Sensing*, vol. 12, no. 10, p. 1685, 2020.
- [134] J. Huang, B. Chen, B. Yao, and W. He, “ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network,” *IEEE Access*, vol. 7, pp. 92 871–92 880, 2019.
- [135] T. K. Hon and A. Georgakis, “Enhancing the resolution of the spectrogram based on a simple adaptation procedure,” *IEEE transactions on signal processing*, vol. 60, no. 10, pp. 5566–5571, 2012.
- [136] T. MathWorks, *Wavelet Toolbox*, Natick, Massachusetts, United State, 2020. [Online]. Available: <https://www.mathworks.com/help/wavelet/>
- [137] T. MathWorks, *Signal Processing Toolbox*, Natick, Massachusetts, United State, 2020. [Online]. Available: <https://www.mathworks.com/help/signal/>
- [138] B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press, 2015.
- [139] A. Khan, A. Sohail, U. Zahoor, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.
- [140] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [141] A. Khan, A. Sohail, U. Zahoor, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.
- [142] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [143] P. Kumar and V. K. Sharma, “Detection and classification of ECG noises using decomposition on mixed codebook for quality analysis,” *Healthcare Technology Letters*, vol. 7, no. 1, pp. 18–24, 2020.

- [144] Y. Liu and J. Lin, “A general-purpose signal processing algorithm for biological profiles using only first-order derivative information,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019.
- [145] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, “Deep sequential image features on acoustic scene classification,” in *Proc. DCASE Workshop, Munich, Germany*, 2017, pp. 113–117.
- [146] S. Jayalakshmy and G. F. Sudha, “Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks,” *Artificial Intelligence in Medicine*, vol. 103, p. 101809, 2020.
- [147] T. MathWorks, *Deep Learning Toolbox*, Natick, Massachusetts, United State, 2020. [Online]. Available: <https://www.mathworks.com/help/deeplearning/>
- [148] R. O. Duda, P. E. Hart *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.
- [149] S. S. Xu, M. W. Mak, and C. C. Cheung, “Towards End-to-End ECG Classification With Raw Signal Extraction and Deep Neural Networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1574–1584, 2019.
- [150] D. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” *Mach. Learn. Technol.*, vol. 2, 01 2008.
- [151] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson, “Automatic sleep staging using support vector machines with posterior probability estimates,” in *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-LAWTIC’06)*, vol. 2. IEEE, 2005, pp. 366–372.
- [152] S. I. Niroshana, X. Zhu, Y. Chen, and W. Chen, “Automatic Sleep Stage Classification Based on Convolutional Neural Networks,” in *2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, 2019, pp. 275–276.
- [153] I. N. SM, X. Zhu, Y. Chen, and W. 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)*. IEEE, 2019, pp. 1–7.
- [154] S. I. Niroshana, X. Zhu, K. Nakamura, and W. Chen, “A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network,” *Plos One*, vol. 16, no. 4, p. e0250618, 2021.