Design of a Convolutional Neural Network for Classification of Physiological Signals

Submitted to

Graduate School of Information Science and Technology

Osaka University

January 2023

Taweesak EMSAWAS

List of Publications

Journal

- Emsawas T., Morita T., Kimura T., Fukui K., and Numao M. (2022) Multi-Kernel Temporal and Spatial Convolution for EEG-Based Emotion Classification. Sensors 2022, 22, 8250.

International Conference / Workshop (peer reviewed)

- Emsawas T., Kimura T., Fukui K., and Numao M. (2020) Comparative Study of Wet and Dry Systems on EEG-Based Cognitive Tasks, Brain Informatics. BI2020: Lecture Notes in Computer Science, vol 12241. Springer, Cham.
- Emsawas T., Kimura T., and Numao M. (2020) Feasible Affect Recognition in Advertising based on Physiological Responses from Wearable Sensors. In: Ohsawa Y. et al. (eds) Advances in Artificial Intelligence. JSAI2019. Advances in Intelligent Systems and Computing, vol 1128. Springer, Cham.
- Uraji H., Emsawas T., Hagad J., Fukui K. and Numao M. (2019). Analyzing the effect of video media on emotion using a VR headset platform and physiological data. Workshop on Computation: Theory and Practice of Computation. WCTP2018 (pp.151-158).

International Conference / Oral Presentations (without review)

- Emsawas T., Fukui K., and Numao M. (2019) Feasible Affect Recognition in Advertising based on Physiological Responses from Wearable Sensors, JSAI2019: The 33rd Annual Conference of Japanese Society for Artificial Intelligence.
- Emsawas T., Kimura T., Fukui K., and Numao M. (2019) TV commercial and emotion recognition using physiological data, The 22nd SANKEN International Symposium.

Abstract

Understanding human expressions through physiological signals using machine learning has recently become a hot topic. One or more modalities studies have investigated various medical and non-medical applications to recognize the signals corresponding to the desired action. Deep learning has been introduced for effective and reasonable analysis in feature extraction and classification tasks. However, the analysis remains challenging since it requires consideration of various architectural design components that influence the representational ability of extracted features. Hence, this study aims to specifically develop the classification model using an end-to-end convolutional neural network (ConvNet) to detect the responses from physiological signals. In particular, this study presents three findings of network architectures in physiological data, especially electroencephalography (EEG). Firstly, this study compares multiple modalities and a single modality of EEG signal in the affect recognition of an advertising video. The classification model applies sequence learning using long short-term memory (LSTM) network to deliver continuous emotion labels throughout the duration. The results show that the output sequence corresponds to the user's feeling scores over time. Secondly, this study investigates a comparative study of wet and dry systems of electrodes using two cognitive tasks: attention and music emotion. Deep and shallow convolution neural networks are applied as the feature extractor of temporal and spatial filtering from raw EEG signals. Additionally, transfer learning is used to improve the performance of the dry system by using transferred knowledge from the wet system. Thirdly, the multi-scale kernel is presented in the temporal filtering for EEG feature extraction to enable extracting a wide range of representations, covering short- to long-wavelength components. Compared with existing models on EEG-based emotion datasets, the proposed model outperforms with higher accuracy results and a few trainable parameters. Ultimately, adding the proposed multi-scale module to existing models of EEG-based convolution networks improves the learning ability and classification performance while preserving computation costs.

Acknowledgement

The dissertation would not have been possible without the support and helps of my professors, family, and friends. I would like to express my deepest gratitude to my advisor, Professor Masayuki Numao, for giving me the opportunity to study in Japan, advising me on research, providing the budget for extension, being a good listener, and kindly taking care of me. I would also like to express my very great appreciation to my advisor, Associate Professor Ken-ichi Fukui, for his patient guidance, helpful comments, and continual enlightenment regarding my research. I am also thankful to Assistant Professor Tsukasa Kimura for gracious advice, intended instruction, and persistent help throughout the experiment and research works. I am also grateful to Assistant Professor Takashi Morita for the valuable and constructive suggestions throughout this research work. My gratitude also goes out to the dissertation committees, Professor Yasumasa Fujisaki and Professor Jun Tanida, for their great comments and useful contributions toward my doctoral degree.

I would like to acknowledge the laboratory secretaries, Ms. Megumi Tanabe, Ms. Mitsuyo Otsuka, Ms. Akiko Yamamoto, Ms. Makiko Kuroki, Ms. Mamie Abe, Ms. Yumiko Morikawa, Ms. Yohko Okada, Ms. Asami Kasuga, and Ms. Kaori Nawa, for their accommodating and assisting me with all of the university affairs. My sincere thanks also go to Thai members of Numao laboratory, Dr. Nattapong Thammasan, Dr. Sirawit Sopchoke, Dr. Wasin Kalintha, Dr. Ekasit Phermphoonphiphat, Nattapat Boonprakong, Nat Pavasant, Pongpisit Thanasutives for their encouragement, guidance, and insightful advice during the student life. I would also like to thank other Numao laboratory students who have been substantially sharing their thoughts on my research, especially Juan Lorenzo Hagad, who has been working and discussing the research and life with me.

This endeavor would not have been possible without Professor Boonserm Kijsirikul, who opened and recommended me this pathway to be a student at Osaka University and for his compassionate guidance and support. I would not have had the honor of being a student of Osaka University without the generous scholarship of Japan International Cooperation Agency (JICA) under their Innovative Asia scholarship program for their support in terms of scholarship and pleasing guidance before and during the time in Japan. I would like express my gratitude and appreciation to the Center of Innovation program from Japan Science and Technology Agency (JST) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan for their continuous support in terms of scholarship. I would like to express my sincere appre-

ciation to Assistant Professor Marieke van Vugt, who provided the office during my work at the University of Groningen, and enlightened me on valuable knowledge regarding computational neuroscience.

I would be remiss in not mentioning Thai friends for their support and encouragement throughout my life in Japan, especially Dr. Naruchit Thanuthanakhun who provide the mental and moral supports. Lastly, I am forever grateful and thankful to my family, Kamonrat Emsawas, Raksak Emsawas, and Artid Emsawas, for always being supportive and encouraging me throughout my study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background to the Research

Recently, understanding human expressions regarding affect or emotion through physiological signals using computer interfaces has been a hot topic. The breakthrough is to develop a system for gathering and comprehending data from internal signals of human body. The physiological signals in response to the nervous systems of the human body [8] have several modalities, such as electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), and galvanic skin response (GSR). One or more modalities studies have investigated to recognize the signals corresponding to the desired action in various applications [40][51]. Regarding in recent years, the brain-computer interface (BCI) has become a technology that has gained wide attention. It acquires brain signals and interprets neuronal information into the desired action. BCI has been used in both medical and non-medical applications [3], such as assistive technology [44, 28], game playing [54, 53], and mental state recognition [47, 56]. There are several ways to record brain activity. One of the most popular modalities of BCI is EEG. The EEG method has been used to record electrical signals in the human brain by measuring tiny voltage fluctuations using electrode sensors. The EEG recording can be performed by attaching electrodes to the scalp surface without surgery and implantation. This non-invasive EEG system holds the promise of real-world BCI applications and is currently entering the mass market. Accordingly, two main types of electrode sensors are built to record EEG signals for different purposes [24]. The medical proposal mainly conducts the EEG recording using a wet system by placing electrodes on the scalp with a conductive gel [58]. The wet electrodes are typically used in expert

disciplines, such as clinical neurology, neurobiology, and medical physics. For example, EEG has been used to diagnose abnormalities of the human brain by detecting the presence of aberrant electrical activity [2, 5]. However, a wet system is a useful tool for these analyses, but it is meticulous and requires time and cost. Then, the dry sensors are generally designed based on wearable and mobility concepts and use in practical applications and non-medical tasks to investigate, entertain, or monitor brain activity [11][43]. In non-medical fields, attempts have been made to transduce EEG states from human players to control video games [57]. Moreover, the acquisition of EEG signals is also exploited to recognize human psychological states [36]. Nevertheless, the EEG recording using dry electrodes could be easily affected by various sources of noise, including eye movement, muscle contraction, and environmental settings, which present significant difficulties in EEG interpretation. [15][21].

Despite its several potential applications, the non-stationarity and high-dimensionality of EEG signals pose a challenge to reliable implementation. Because the brain is a complex organ with different parts that function and respond differently, to evaluate spatial brain activity, EEG data are generally recorded in more than one and up to 64 electrode positions increasing space dimensionality and complexity during feature extraction and analysis. Therefore, the capability of developed EEG methods should consider several aspects, including interpretability, performance, and usability for real-world applications. Ascertaining methods aim to understand complex EEG signals derived from the human brain extracting their informative features and classifying them. Conventionally, beneficial knowledge is extracted from raw EEG signals by manually defined features and reported through statistical reports. Nevertheless, these defined features are easily sensitive to noise, and their computation requires appropriate data cleansing, signal preprocessing, and expertise.

Recently, many studies have applied deep learning (DL) to extract and interpret EEG signals. An end-to-end convolutional neural network (ConvNet or CNN) is constructed on the basis of a shared-weight architecture that can reasonably learn joint optimization of feature extraction and classification. The appropriate design, such as the layer's depth and width, kernel size, and optimizer, becomes an essential consideration that significantly affects the representational ability and classification performance. It has been demonstrated that deep and shallow convolution

networks can automatically extract essential temporal and spatial information from raw EEG signals [49]. Nevertheless, as the feature extraction of these models is based on a single kernel size, their learning ability and EEG representational performance are limited. Larger ranges of signal transformations are needed to represent differences in slow and hyperactive EEG frequencies. Parameter tuning and optimization are also required for a particular task. In addition to the kernel, parameter dependence and computational cost in training and testing processes have to be taken into account [39]. A separation of network layers was suggested to help reduce the computational cost and model complexity. To further improve model efficiencies, the network architectures, including kernel sizes and parameter sets, should be rationally designed for each specific EEG analysis task.

1.2 Significance of the Research

By addressing the above concerns, this research aims to study the appropriate techniques for physiological signal analysis, especially EEG tasks, in various domains. First, this research presents the uses of modalities to recognize human affects using sequence learning in an advertising video. In particular, the experiments and results show the relative responses between human feelings and model outputs and investigate the informative modality for affective computing. Second, this study focuses on feature extraction and classification by comparing the traditional techniques and EEG-based end-to-end convolution neural networks. The experiments consist of multiple comparisons from various datasets, including self-recording of different electrode types and existing datasets. Lastly, the main contribution proposes an EEG-based emotion classification model called the multi-kernel temporal and spatial convolution network (MultiT-S ConvNet). The multi-scale kernels or filters are used in the proposed model to improve temporal representations by combining short- to long-wavelength patterns learned from raw EEG data. Furthermore, a lightweight gating mechanism enhances the extracted features enhancing both the temporal and spatial filters with global information. Subject-dependent and subject-independent experiments are conducted on EEG-based emotion datasets to validate the performance of the proposed MultiT-S ConvNet compared with existing methods.

This research investigation can shed light on the design of ConvNets models by improving

a wide range of representations with optimized trainable parameters. Furthermore, the proposed module can potentially enhance the learning ability and accuracy performance of any EEG-based ConvNet models. With the fabulous features automatically extracted from a limited dataset, ConvNet models are a promising tool in the EEG decoding toolbox in a variety of applications, both medical and non-medical tasks.

1.3 Organization

The rest of this dissertation is organized as follows: Chapter 2 is the related works of literature, including conventional analysis, machine learning models, and end-to-end convolutional neural network models. In Chapter 3, the multiple and single modalities are compared in the affect recognition of an advertising video using sequence learning. Chapter 4 presents a comparative study of wet and dry systems of electrodes using fashionable convolution neural networks to extract the temporal and spatial features. Then, Chapter 5 proposes the multi-kernel temporal and spatial convolution network (MultiT-S ConvNet) for enhancing the learning ability of EEG-based feature extraction. Lastly, Chapter 6 shows the discussion, conclusion, and future directions.

Chapter 2

Literature Review

2.1 Frontal Brain Electrical Activity

Frontal brain electrical activity is related to cognitive skills, such as emotional expression, working memory, and working memory capacity. The studies attempt to achieve cognitive recognition with various stimuli by focusing on frontal brain activity. For instance, Trainor [50] found that the pattern of asymmetrical frontal EEG activity distinguished the valence of the musical excerpts and higher relative left frontal EEG activity to joy and happy musical excerpts. Overall, brain activity in the frontal region correlates significantly with musical stimuli, consistent with many of the literature's findings [42]. The frontal theta activity plays an active role in working memory [19]. Theta activity increases while doing the working memory task [31]. To investigate an association between brain activity and working memory, Wayne Kirchner introduced the n-back task [35]. It is a performance task commonly used as an assessment in cognitive research. In the n-back task, the subject is requested to continuously memorize a single character's sequence on the screen. The subject task is to indicate if the current character is the same as the previous n-step.

2.2 Conventional EEG Analysis

In classifying emotions learned from EEG signals, feature extraction is an essential step in which these emotions are represented and categorized into the desired labels. Conventionally, EEG features can be categorized into three domains: time, frequency, and time-frequency domains [30]. The time domain observes time-series characteristics and variations using conven-

tional techniques, such as linear prediction and component analysis. The frequency or spectral domain is the standard method used for quantifying EEG signals by adapting the Fourier transform. Fourier analysis reflects the frequency content using sums of trigonometric functions and then distributes the average power into the Power spectral density (PSD). As shown in Figure 2.1a, the PSD in a human EEG signal is approximately divided into several ranges, including delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-32 Hz), and gamma (>32 Hz) bands. These frequency bands reflect brain activities through the strength of variation. High-frequency bands, such as gamma or beta waves, indicate hyperactive brain activity and alertness, and low-frequency bands, such as delta or theta waves, indicate deep meditation and relaxation. Theta (4-7 Hz) is a slow wave associated with the subconscious mind, deep relaxation, and meditation. For instance, frequency bands in slow to hyperactivity brain waves reflect the distinction in the intensity of emotions [50]. Positive emotions, such as joy or happiness, are relatively associated with the left frontal hemisphere. In contrast, negative emotions, such as sadness or fear, are relatively associated with the right frontal hemisphere [27, 29]. Changes in alpha (8-12 Hz) and beta (13-32 Hz) waves are the most discriminative for emotional states [50, 9]. Gamma (>33 Hz) is a hyperactivity wave associated with problem-solving and concentration and is related to positive and negative emotions but on different sides; left for negative and right for positive [6].

In addition to frequency or temporal features, spatial features can be extracted using multi-channel electrodes. Each part of the human brain can convey different information through the asymmetric hemisphere. For example, the frontal lobe is associated with reasoning, parts of speech, and emotion. The temporal lobe is associated with the perception and recognition of auditory stimuli, whereas the occipital lobe is responsible for vision. For this reason, EEG signals require a method that can manipulate temporal and spatial information for feature learning and classification.

A common spatial pattern (CSP) [48] is a mathematical method for computing the variance of features to discriminate window patterns or emotional classes. This method uses the simultaneous diagonalization of two covariance matrices to construct optimal spatial filters [4, 59]. CSP patterns can be used as features for machine learning (ML) to classify emotions [41]. However, the classification of CSP features requires a specific frequency range, which signif-

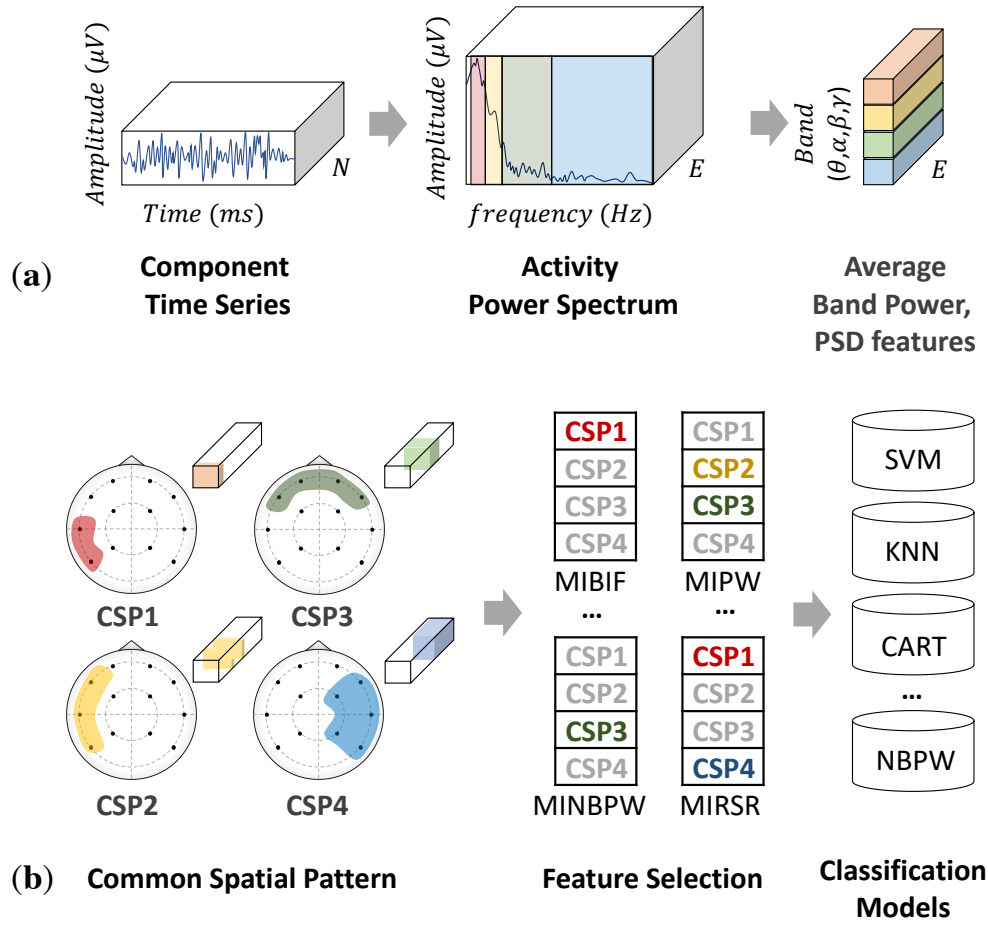

Figure 2.1: (a) Power spectral density (PSD) features extracted from raw EEG signals. (b) Filter bank common spatial pattern (FBCSP).

icantly depends on the subject or task. To address this problem, filter bank common spatial pattern (FBCSP) [4] has been proposed to perform an autonomous feature selection through temporal-spatial filtering. Compared with CSP methods, FBCSP, as shown in Figure 2.1b, adds two more processes to perform feature selection and classification, respectively. The first two stages perform temporal and spatial filtering to construct a filter bank of discriminative CSP features. Then, features are selected independently depending on the classifiers in the third stage. Popular techniques with good EEG classification accuracy include random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM), and fully connected network (FCN) [7, 51]. However, these techniques employ manual parameterization to classify EEG features, and the appropriate selection for a particular task relies on the complexity and cleanliness of data recording.

Additionally, the whole duration of EEG signals requires the necessary processing to separate them equally. Window segmentation is the technique that slices the raw signal into a similar window size and slides it with or without overlapping [30]. Accordingly, the cognitive stages can be observed by an individual sliding window. This continuous recognition while doing the task is an effective method for tracking the cognitive reporting oscillations, and the results provide an opportunity to understand human emotional processes better [55, 16]. However, the sequence output must consider the possibility of acquiring continuous emotional states.

2.3 EEG-based End-to-End Convolutional Neural Networks

Deep Convolutional Neural Networks (Deep ConvNets) have reformed the computer vision research field with the learning ability of hierarchical patterns [38]. Then, Deep ConvNets have been successfully applied and achieved in many research fields. In recent years, the use of end-to-end CNNs (ConvNets) has been introduced for effective and reasonable EEG analysis in feature extraction and classification tasks. First, the raw EEG signals are measured in microvolts (μV) and recorded in a 2D space, with temporal (T, time) and spatial dimensions (E, the number of electrode channels). Then, the convolutional operation is applied to extract informative features through three types of filtering, i.e., convolving across time, space, and both time and space, respectively, as shown in Figure 2.2. Temporal filtering convolves information across the time-space domain with a $1 \times t$ kernel size. The temporal content of each electrode channel is extracted by shared weighting over the time-space input. Spatial filtering proceeds with the filter matrix $E \times 1$ across all electrode channels, learning the variance of features. The temporal-spatial kernel with a $k_1 \times k_2$ 2D kernel size, where k represents a specific kernel size, is applied to both time and channel domains. In addition, there is a need to determine the suitable size for learning various representations and distinctions. Typically, the kernel size is manually defined and used in feature extractors in CNNs. A single-scale kernel might be trapped with a limited amount of representations. Especially for EEG feature extraction, learning representations needs to be diverse and capable of effectively extracting temporal and spatial features.

Apart from kernel designs, the network depth significantly influences the learning strategy, particularly low–high-level feature decoding. For example, a deep convolutional network (Deep

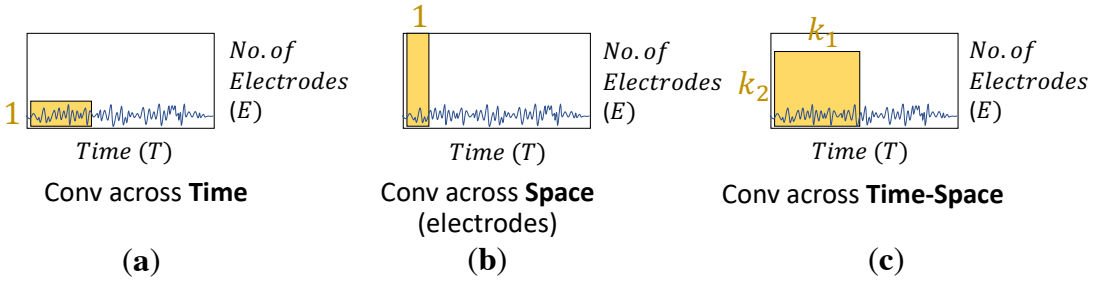

Figure 2.2: EEG convolution filter types include (a) temporal filter, (b) spatial filter, and (c) temporal-spatial filter. The yellow blocks denote the two-dimensional (2D) kernel for performing feature extraction.

ConvNet) has been used for EEG decoding with a single-layer temporal filter. Then, the output is fed into multi-layer spatial convolution and pooling layers, as shown in Figure 2.3a. The EEG classification results show that the Deep ConvNet outperforms the widely used FBCSP (mean decoding accuracy of 84.0% and 82.1%, respectively) [49]. Deep ConvNets can achieve competitive accuracy and be applied to general EEG decoding tasks. On the other hand, according to the FBCSP pipeline, a shallow convolutional network (Shallow ConvNet) is designed for tailoring decoding band power features, as shown in Figure 2.3b. The first layer performs a temporal convolution to simulate bandpass extraction. Then, the second layer performs a spatial convolution that analogizes the CSP spatial filter in FBCSP. The extracted features of Shallow ConvNet are related to log band power and are designed explicitly for oscillatory signal classification.

Moreover, another critical aspect that affects informative features and learning ability is reducing the network size while maintaining good performance. One of the simpler networks is a separable convolution that enables networks to construct informative features by fusing both spatial and channel information. The operation consists of spatial mapping independently performed over each input channel and feed-forward mapping to project the output onto a new feature space. These operations enable networks to construct informative features by fusing spatial and temporal information within local receptive fields at each layer. A compact CNN for EEG-based BCI, called EEGNet, is introduced to construct an EEG-specific model with separable convolution [39]. The number of trainable parameters in EEGNet is significantly less than that of Deep ConvNet and Shallow ConvNet (170 and 100 times, respectively), but EEGNet still achieves performance comparable to that of Deep ConvNet and Shallow ConvNet.

2.4 Dataset

This study introduces two open-access datasets: SEED (SJTU emotion EEG dataset) [59] and DEAP (dataset for emotion analysis using physiological signals) [37] datasets. These datasets have been widely used in emotion recognition using multimodal physiological signals. The SEED dataset [59] consists of 64 channels of EEG signals recorded from 15 subjects. All subjects were asked to watch 15 excerpts of movie clips for 3 trials. Each clip was approximately 4 min long, and the time interval between trials was one week or longer. The emotional labels corresponded to three types of movies to stimulate emotional states: positive, neutral, and negative. The recorded EEG signals were downsampled from 1000 to 200 Hz and applied a bandpass filter from 0-75 Hz.

The DEAP dataset [37] is a multimodal dataset that consists of EEG and peripheral physiological signals. The EEG signals of 32 subjects were recorded by 32 electrodes using the BioSemi ActiveTwo system; 40 one-min excerpts of music videos were used to stimulate emotional states. All subjects were asked to rate the emotion score (1-9) in the valence-arousal space. The EEG signal was recorded with a 512 sample rate. The signals were downsampled to 128 Hz, a bandpass filter was applied to them from 4 to 45 Hz, and EOG artifacts were removed. According to the review paper [51], the average accuracy of the SEED dataset was 90.0%, significantly higher than 83.6% of the DEAP dataset within-subject dependencies.

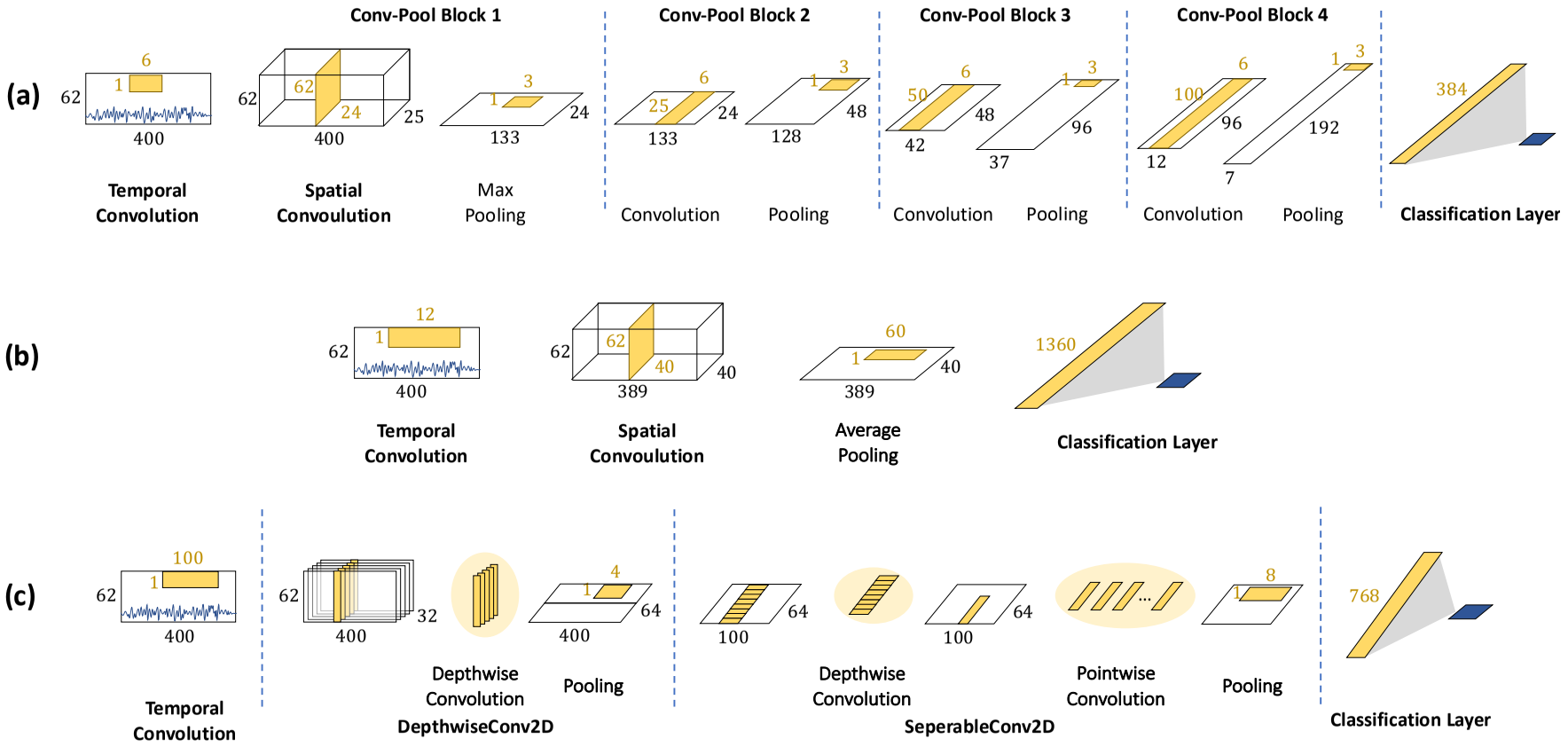

Figure 2.3: The network architectures. (a) Deep ConvNet (b) Shallow ConvNet and (c) EEGNet

Chapter 3

Feasible Affect Recognition in Advertising

3.1 Overview

Recent studies in affective computing have facilitated and stimulated the development of systems and sensors that can recognize and interpret human affects. Affective computing has been applied in various domains, and one of the applied domains is in the marketing area to increase the consumers' appeal and attraction. In particular, advertisements (ads) can convey amounts of information in a short time. Therefore, using physiological responses can help to acquire a user's feedback and obtain an advantage. This study proposes non-invasive affect recognition in each scene of an advertising video using electroencephalogram (EEG), electrocardiogram (ECG) and eye-tracking. The preliminary analysis of EEG shows the relationship between scene feeling score and emotional affects regarding physiological responses. Hence, this study also trained two types of recognition models: window-based learning and sequence learning. The models learned from the physiological responses and questionnaires on a user's preference in each ad scene.

3.2 Problem Statement

The measurement of marketing stimuli derived from subconscious deliberation has become a challenging topic this decade [18]. Acquiring unconscious mental processes using physiological sensors enables the marketer to discover what the customer thinks and feels about brand-related

messages while watching the advertising stimuli. Accordingly, the analysis based on physiological data comprises data collection, feature extraction, and classification to deliver customer responses such as emotional states. Furthermore, the responses can be a single emotion or a sequence of continuous emotions based on the research proposal and the advertising stimuli. The conventional techniques extract the frequency bands from the whole signal and use them as the features to train and classify the machine learning models such as multilayer perceptron (MLP) [46], and support vector machine (SVM)[23]. On the other hand, the studies focus on the continuous time interval of sequential behavior, or sequence learning [20] which recognizes and output the human affects throughout the duration. This study focuses on continuous ads-affect recognition that considers how the strength of changing states while watching ads. This work proposes subject-independent recognition using non-invasive wearable sensors, and a 15-second ad video was selected as stimuli. The data acquisition and preprocessing are described in Section 3.3, including electroencephalogram (EEG), electrocardiogram (ECG), and eye-tracking. This work studies the subject affects as negative or positive (low or high) emotional affects each scene interval. Section 3.4 presents our research methodology and experiment setting. The data trend is presented, which represents an observation of the relationship between physiological data and feeling score. Furthermore, The learning techniques, including MLP, SVM, and LSTM networks, were performed to recognize these emotional human affects. In Section 3.5, the result of the preliminary experiment shows a relation between the scene-feeling score of questionnaires and the physiological responses. Moreover, recognition and comparison results are shown in this section. The results of continuous affect recognition can support the interpretation of the scene interesting and how subjects interact with ad scenes. Finally, in Section 3.6. the discussion and conclusion are presented.

3.3 Data Acquisition and Preprocessing

This study used an undisclosed dataset provided from the company of toothpaste holder. The physiological dataset was collected by asking the subject to watch the advertising video while wearing the sensors. 130 subjects between 20 and 50 years of age including half of males and females participated in the experiments, and all subjects had never watched the ad video before,

to assure a genuine reactions. The 15-second ad video of toothpaste was chosen as stimuli to elicit emotional affect from the subjects. This ad video contains 11 scenes which respectively present tooth problems, bacteria illustration, brushing teeth, cleaning up and toothpaste products. The subjects were asked to watch the video while wearing Neurosky's sensors. Then, after watching the video, all subjects were asked to fill out the questionnaires concerning scene feeling based on 9 scales of preference (-4 to 4). This study represent the score of 0 to 4 and -1 to -4 as positive and negative labels respectively.

3.3.1 Electroencephalogram (EEG)

The frontal-brain activity plays a crucial role in emotion recognition [50] and distinguishes emotional affects, which were known to vary in affective valence such as positive and negative. In this research, EEG data were acquired from one electrode in the frontal-brain area. Various techniques have been proposed in the feature extraction step to estimate the emotional state from the acquisition signals [33][52]. Frequency domain features as power spectral density (PSD) seem to be the most commonly used method [3]. Consequently, PSD feature extraction was employed to extract features from EEG signal into five frequency bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta(13-30 Hz), and gamma (>30 Hz) from low to high frequencies. In this study, NeuroSky MindWave headset[1] is used as a portable EEG dry sensor to record the EEG signals. The reference and ground electrodes are on the ear clip, and the FP1 electrode is on the frontal sensor arm. Neurosky's sensor calculated and recorded five frequency bands with a sampling rate of 512 Hz.

3.3.2 Electrocardiogram (ECG)

CardioChip/BMD10X-based devices of Neurosky's sensors recorded and converted to ECG signals. ECG signal consists of P-QRS-T waves in each one cardiac cycle as shown in Figure 3.1. P-wave reflects atrial depolarization or activation. QRS complex appears after the P-wave, which represents ventricular depolarization and contraction. Within QRS rapid duration, a short complex implies the conduction system function properly, while the wide complex indicates dysfunction. T-wave represents the rapid repolarization of ventricles. The intervals

[1] https://neurosky.com/biosensors/eeg-sensor/biosensors/

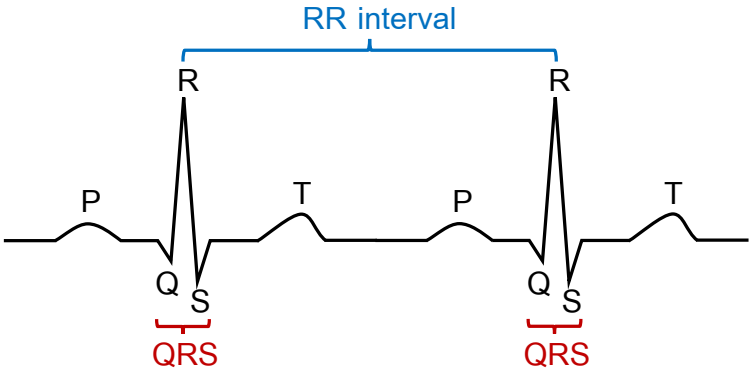

Figure 3.1: ECG curve with its most common P-QRS-T waveforms.

between successive heartbeats per minute of P-QRS-T waves are called RR intervals. In this study, the features were extracted throughout the whole P-QRS-T segment as the standard deviation of the RR intervals (SNDD), the number of pairs of successive RR intervals that differ by more than 50 ms (NN50) and the proportion of NN50 that divides by the total number of NN intervals (PNN50)[32, 14].

3.3.3 Eye-tracking

Video-based method of eye tracking is widely used in commercial eye trackers, which measure horizontal and vertical components of the movements of both eyes [10]. The study used the Neurosky sensors for detecting the distance of each X and Y pair and used them as an eye-tracking feature. To synchronize all physiological features, the lower-frequency feature was duplicated multiple times until it matched the timestamp with the higher-frequency feature. For example, in the case of all features, EEG and ECG features were duplicated until they matched with eye-tracking features. For the sake of simplicity, the targeted labels of affect recognition were extracted from questionnaires using mapping and rearrangement depending on the period of 11 scenes of advertising stimuli. Thus, each second which involves more than 2 scenes would calculate the average of time-weighting by

$$f_t = \sum_{i}^{N_t} (sf_i * t_i) \tag{3.3.1}$$

where f_t is feeling score at time t, sf_i is feeling score at scene i, t_i is time weight at scene i, N_t is the number of scene which relates to time t and t is observed timestamp $(1,2,3,...,15)$.

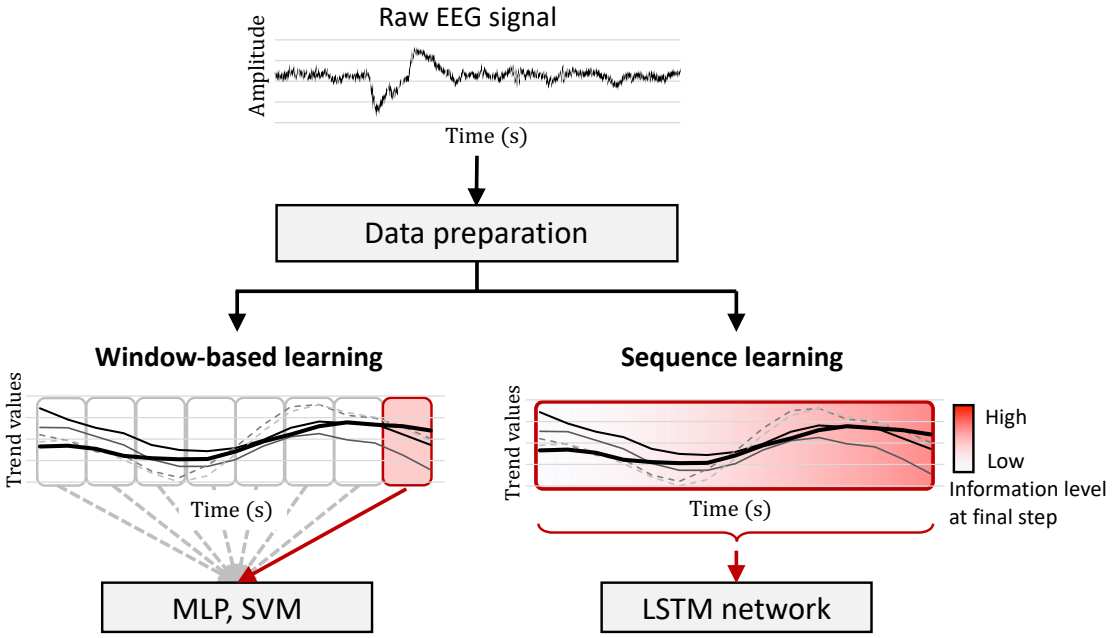

Figure 3.2: Information level of window-based learning and sequence learning techniques

3.4 Methodology

This study focused on the sequence-to-sequence learning between physiological signals and emotional affect. A preliminary experiment presented the trend value to comprehend the relation of physiological data over time. For affect recognition, the emotional affect was classified as positive and negative. The techniques can be divided into two types: window-based learning and sequence learning. The input and output data were constructed from extracted features and questionnaires, which are shown in Figure 3.2. This research also experiments to compare the features used as only EEG features and all features because some physiological responses, in 15-second ads-video watching tasks, can not gain explicit responses.

3.4.1 Data Trends

To observe the relationship of physiological data and feeling score, representation of the same axis is required to indicates whether each particular band is increasing or decreasing over time, and how strong each data values. The study utilized the Pearson correlation coefficient to represent the data trend or r. Each trend value is calculated by using the 4-second sliding window technique. The positive value is a direct variation of data values over time, and the negative

value is an inverse variation. The trend value can be calculated by

$$r = \frac{\sum_i^n (x_i - \bar{x})(t_i - \bar{t})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \sqrt{\sum_i^n (t_i - \bar{t})^2}} \tag{3.4.1}$$

where n is sliding window size, x is data value, t is time in second and i is running number of sliding window. These trend values were only applied for PSD features of EEG data. While ECG and Eye-tracking, their features were directly used as the input of classifier.

3.4.2 Window-based learning

window-based learning recognizes emotional affects each second in which this study applied the 4-second sliding window for segmentation technique to analyze temporal data and recognize emotional affects over time. The model input was constructed from the trend values in Section 3.4.1, and the shifting size is 1 second. Then, the values were mapped to the time axis for data and label consistency. Meanwhile, the label or output was the feeling score value at each second. In the attempt to compare with the standard models, MLP and SVM are ubiquitous techniques were chosen. Two layers of feed-forward neural networks were trained with 100 hidden units and softmax activation in each layer. Besides, SVM models were trained with a linear kernel. For illustrating the data representation, Figure 3.2 (left) shows the window as red-rectangle and shading, which is the information level of the sliding window at the final time step. The classifiers merely learn the information from the individual window. However, these techniques are not able to transmit useful information between each window entirely but can learn along with the values according to the sliding window size.

3.4.3 Sequence learning

This study also applied sequence learning [20] as LSTM networks [25] which improve from a recurrent neural network (RNN). The main idea of the RNN is to memorize previous information in the network's internal states by using self-connections. The benefit of an RNN is the ability to use contextual information from the whole data sequence, and the dependencies of inputs can remain in the network. Figure 3.2 (right) shows the information level of learning techniques. The LSTM architecture preserves information by using memory cells and gate

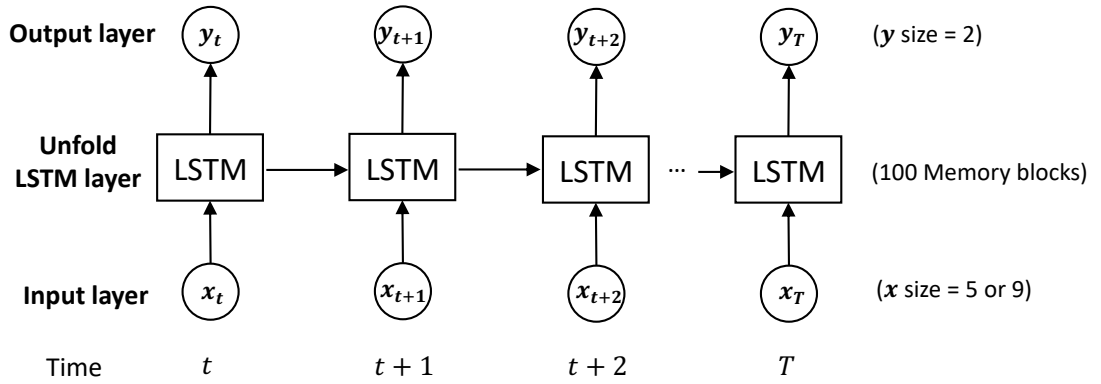

Figure 3.3: The unfold structure of the LSTM networks. The first layer size (below) is based on the input vector of the experiment. The hidden layer (center rectangles) consists of 100 memory blocks. The final layer output the vector of size 2 for positive and negative classes.

units which allows the exhibiting of temporal dynamic behavior for a time series data. In the experiment of sequence learning, the LSTM networks were used to learn the dependencies of physiological responses. After constructing from the trend values in Section 3.4.1, these experiments perform a 1-second sliding window that contains the data values, and each feeling score in a second was used as the label. The sliding window is shifted along the data waves to construct an input vector for each window. Then, only EEG features experiment constructed the fully connected LSTM networks with 5 input nodes according to frequency domains, 100 memory blocks, and 2 output nodes according to the target output. Besides, the input layer were adjusted to 9 input nodes for all features experiment. The network setup for both only EEG features and all features used experiments are shown in Figure 3.3. After having been trained, the network can be used to predict the label sequence of feeling score.

3.5 Experiments and Results

3.5.1 Preliminary Experiment

In the preliminary experiment, the trends were calculated and investigated considering the five frequency bands of only EEG data over time. The subject's averages between EEG data over time are shown in Figure 3.4, and the average time-weighting of the feeling score is shown in Figure 3.5, respectively. Considering these illustrations, the evidence is found that shows the relation between EEG responses and scene feeling scores. Fortunately, this study can indicate

the unusual point around 3 to 7 seconds, which is the disgusting scene of bacteria. Hence, the investigation continued to examine the recognition experiment based on physiological responses using machine learning techniques.

3.5.2 Recognition Experiment

This section is divided according to the features used, into two experiments: only EEG features and all features. These two experiments applied MLP, SVM and LSTM network to recognize the output label or feeling score each second. 5-fold cross-validation was performed for all subjects. The trend values and feeling score over time were used as the dataset and labels. This dataset was proportionally divided into 5 fold and used for train and test the models. In addition, MLP and LSTM networks need the validation set in the training process, so the training sets of each fold were divided into the training set and validation set with 80% and 20% respectively. The results are shown in Figure 3.6. In the experiment of only EEG features, LSTM network achieved the highest accuracy at 76.4%, and SVM and MLP achieved 70.3% and 68.8% respectively. Unfortunately, all features experiment achieved an accuracy of 72.8%, 69.2%, 61.9% respectively.

To illustrate the comprehensive of continuous recognition, two user results of LSTM net-

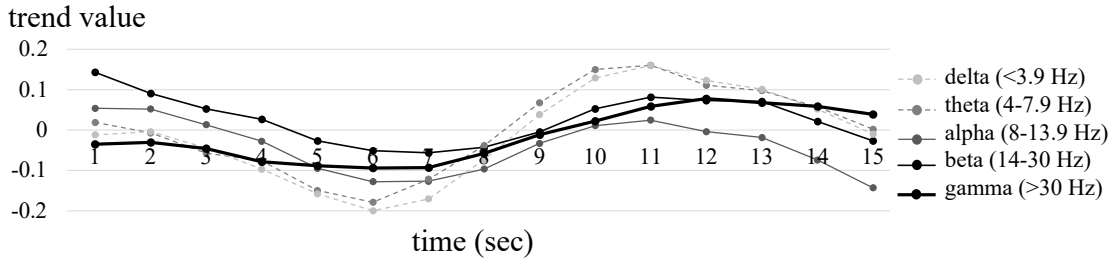

Figure 3.4: Average data trends of EEG data over time for all subjects

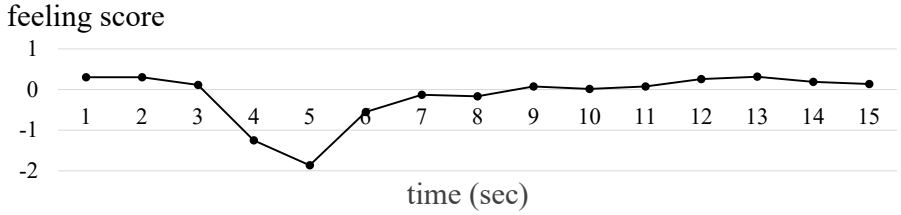

Figure 3.5: Average feeling score overtime for all subjects

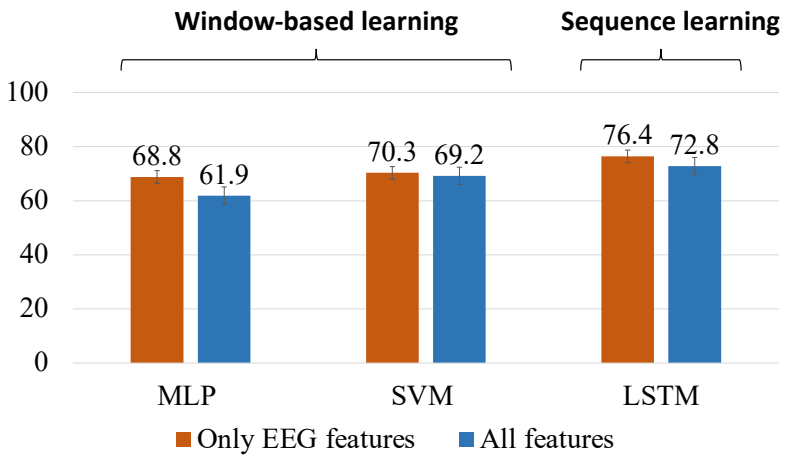

Figure 3.6: Accuracy of affect recognition

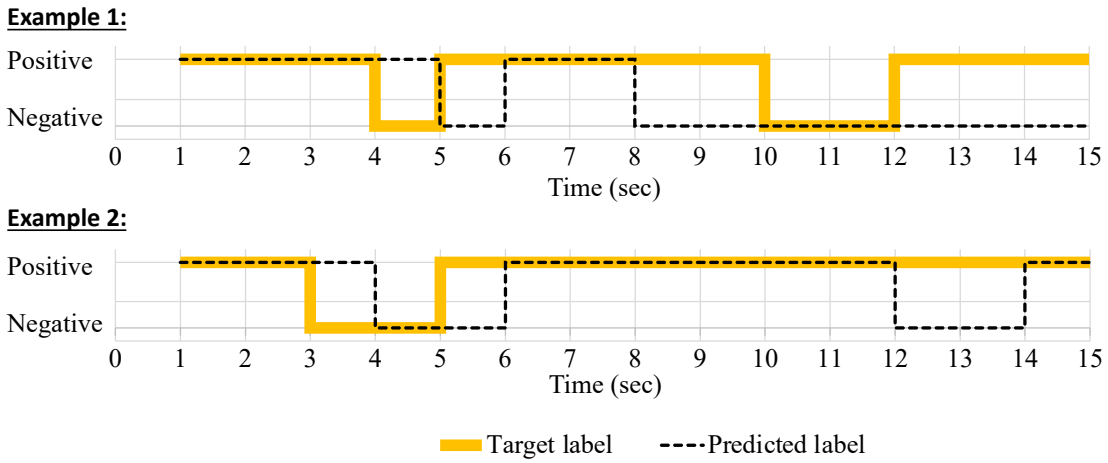

Figure 3.7: The examples of target and predicted results using LSTM network along the duration

work prediction are plotted, which are shown in Figure 3.7. The orange bold-line shows the target feeling score, and the black dot-line shows the predicted feeling score. The x-axis represents the time domain, and the y-axis represents the binary emotional affects: positive and negative. The results can appropriately capture the emotional affect over time.

3.6 Discussion

This chapter present a study of feasible emotional affect recognition based on physiological responses from wearable sensors. The models constructed through this study approach can be used for detecting the emotional affects change over time. For using physiological signals, based on the recognition results, the best results are achieved in recognition when only EEG features

are used. Considering ECG features and according to previous researches, this study found that ECG features could not perform good results in a short duration. While eye-tracking features in this experiment could not denote the emotional affects, the subjects just watched the interesting points in each scene. The recognition techniques present general recognition and sequence learning techniques. The sequence learning outperformed accuracy through the learning of data sequences. The results could gain a continuous affect along with the duration and apply to comprehend the feedback and obtain an advantage from the interesting points in ads. Machine learning techniques can detect the user's response to ads watching by using wearable sensors. The acquired responses can be utilized as informative knowledge without concealing.

Chapter 4

Comparative Study of Wet and Dry Systems

4.1 Overview

Brain-Computer Interface (BCI) is a communication tool between humans and systems using electroencephalography (EEG) to predict certain cognitive state aspects, such as attention or emotion. For brainwave recording, there are many types of acquisition devices created for different purposes. The wet system conducts the recording with electrode gel and can obtain high-quality brainwave signals, while the dry system expressly proposes the practical and ease of use. This study investigates a comparative study of wet and dry systems using two cognitive tasks: attention and music-emotion. The 3-back task is used as an assessment to measure attention and working memory in attention studies. Comparatively, the music-emotion experiments are used to predict the emotion according to the subject's questionnaires. This analysis shows the similarities and differences between dry and wet electrodes by calculating the statistical values and frequency bands. Besides, this chapter further study the relative characteristics by conducting the classification experiments. This study proposed the end-to-end models of EEG classification, which are constructed by combining EEG-based feature extractors and classification networks. A deep convolution neural network (Deep ConvNet) and a shallow convolution neural network (Shallow ConvNet) were applied as the feature extractor of temporal and spatial filtering from raw EEG signals. The extracted feature is then forwardly conveyed to a long short-term memory (LSTM) to learn the dependencies of convolved features and classify attention states or emo-

tional states. Additionally, transfer learning was utilized to improve the performance of the dry system by using transferred knowledge from the wet system. The model was applied not only on the dataset but also on the existing dataset to verify the model performance compared with the baseline techniques and the-state-of-the-art models. Using this proposed model, the result shows the significant differences between accuracy and chance level in attention classification (92.0%±6.8%) and SEED dataset's emotion classification (75.3%±9.3%).

4.2 Problem Statement

This chapter contributes to the data analysis of wet and dry systems and the EEG-based classification model. The novelty of this study is the first attempt to transfer knowledge from wet to dry system to improve the performance of the dry system. Two cognitive experiments were conducted to study a comparison of wet and dry systems. The attention experiment assessed using a 3-back task and learned the attention/non-attention classes of EEG signals. For more difficult BCI-task, this study learned the positive and negative emotions of EEG signals by stimulating with the MIDI song. In terms of classification, the proposed method utilizes Deep ConvNet and Shallow ConvNet to perform feature extractor and LSTM networks to perform classification learning. These classification results also show the dependencies among datasets and systems comparatively. To improve dry system performance, transfer learning was applied by freezing the ConvNet of the wet system and used it as a feature extractor of the dry system. Because the dataset is considerably small, collecting from 5 subjects, so the model performance was also verify by implementing on the existing datasets. DEAP [37] and SEED datasets [59] are top 2 citation datasets on EEG-based emotion recognition [3]. Thus, this study is expected to contribute to developing the combination of non-invasive BCI with a dry electrode and machine learning techniques for a future brainwave translation tool.

4.3 Data Acquisition and Preprocessing

4.3.1 Self-collected dataset Acquisition

This study collected dataset from 5 healthy subjects, which are 4 males and 1 female. They are graduate students of Osaka University, between 20 and 30 years of age. The study was approved

by the local ethical committee, and all subjects or their legal representatives provided written informed consent to participate in the study. The subjects were asked to finish two different cognitive tasks: attention experiments using the 3-back task and music-emotion experiments using the brAInMelody application [45]. For acquiring physiological signals in this study, brAInMelody is a recording system demonstrated by wearing the portable sensor and listening to automatic composition. Meanwhile, the EEG signals were recorded while doing these cognitive tasks. The experiments were conducted twice in the corresponding task and setting but properly using different dry and wet systems.

Sensor Device

In this study, the experiments were conducted with two systems comparatively. The wet device is Polymate AP1532 (Miyuki Giken, Japan), an electrode cap (Easycap GmbH, Germany) and Ag/AgCl electrodes with a cable connection to AP monitor program (Miyuki Giken, Japan). The sampling rate is set to 1000 Hz. On the other hand, the dry device is an imec sensor named EEG brAInMelody Gen-3 compatible[1] with brAInMelody application and Nyx software. The sampling rate of the sensor is 256 Hz. EEG placement is in accordance with the international 10-20 system. Ten electrodes (Fz, Fpz, T3, F7, F3, Fp1, Fp2, F4, F8, and T4.) are placed in the frontal brain region. The Fpz and Fz are set as ground and reference electrodes. Consequently, The total signals gained are 8 signals.

Attention Experiment Setting using 3-Back Task

In the experiments, the subject performed the 33 repetitions of a 3-back task, which is memorizing the 3-back character. Firstly, the subject was introduced by instructions and then wore the sensor. Before doing the task, the '+' character appeared in the middle of the screen for 20 seconds. At the same time, the brain signals began to be recorded as a baseline. Then, the 33 repetitions of the 3-back task started sequentially. Each repetition consists of a 0.5-second character and a 2-second blank screen. Additionally, the experiment conducted in a closed room with minimal noise, and the subject was asked to minimize their movement to avoid the noises. To compare 2 systems, all subjects performed the same task 2 times, with a different random set

[1] https://www.imec-int.com/en/articles/amorepacific-uses-imecs-eeg-headset-neuromarketing-research

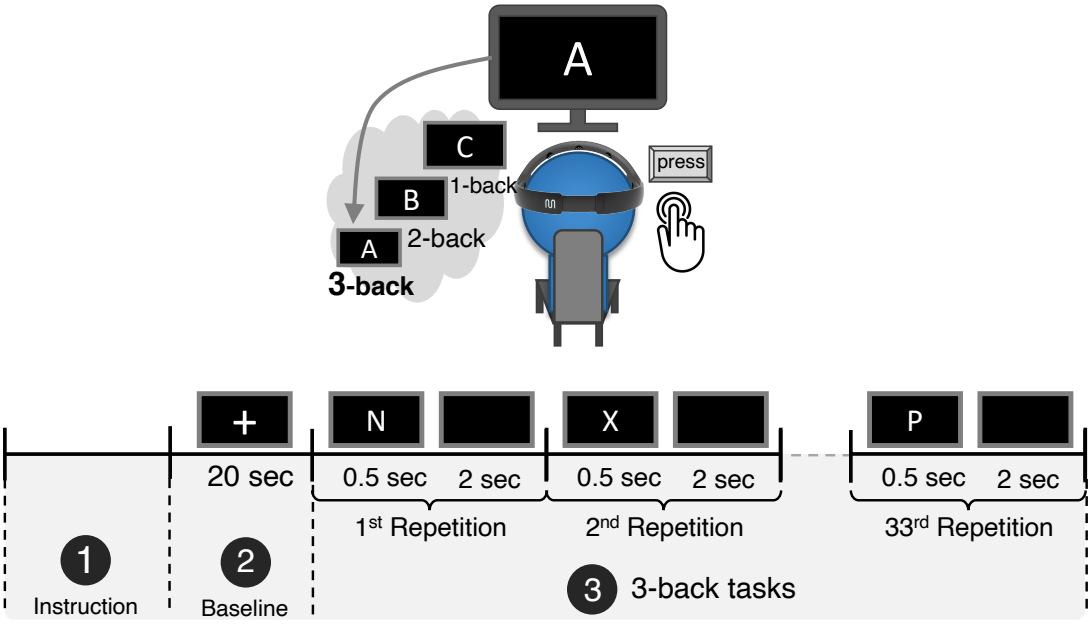

Figure 4.1: Attention experiment setting with 3-back task

of characters, in dry and wet settings. Figure 4.1 shows the illustration and detail of this experiment. This study classifies the recorded signals into 2 classes: attention and non-attention. The recorded signal of the 3-back task is an attention class, and the baseline signal is a non-attention class.

Music-Emotion Experiment Setting

For the music-emotion experiment, the introductory and environment are the same as the previous attention experiment. Each subject was asked to listen to 5 MIDI songs. Each song was added a 20-second baseline with no sound before the music started. After listening, the subjects immediately filled in the questionnaires, consisting of valence and arousal level between -1 to 1. They performed the same playlist for 2 times in a dry setting and a wet setting, respectively. The subjects also filled in the questionnaires for 2 times because their feelings might slightly change in the second round. Moreover, the subject is required to close the eyes while recording a baseline and listening to music. Figure 4.2 shows the illustration and detail of this experiment.

The music-emotion experiment investigates the emotional state from the EEG signals. The target labels or emotions obtained from the subject's questionnaires are valence and arousal values from -1 to 1. Because the emotion label is ambiguous, the study merely classifies 2

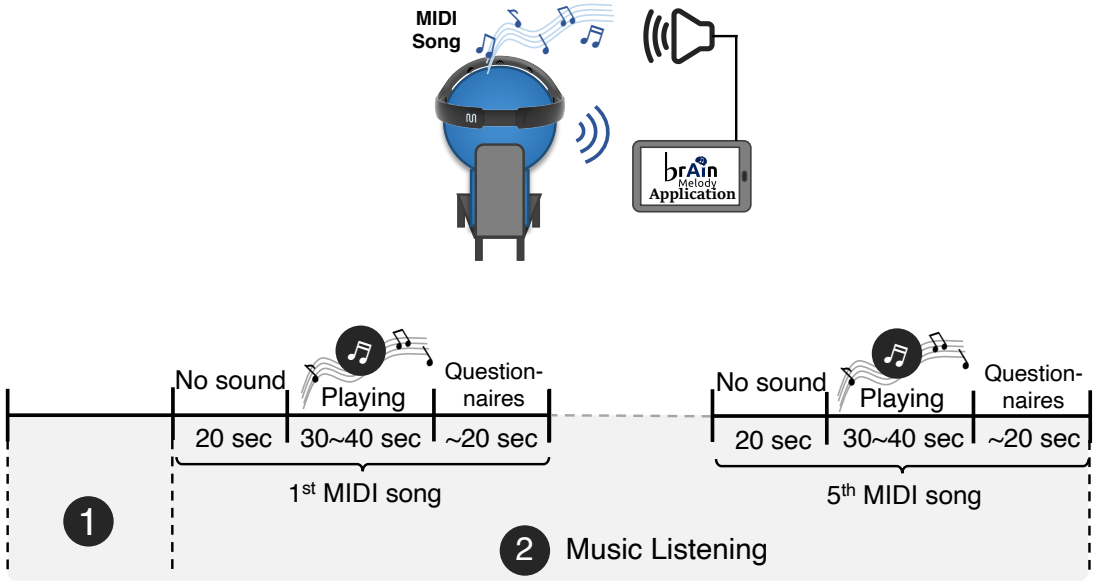

Figure 4.2: Music-emotion experiment setting

classes: positive and negative. The positive class represents positive valence and arousal, and the rest is a negative class.

4.3.2 Existing Datasets

This study is not only interested in the quality of recorded signals, but the quantity of data in BCI research is also a tremendous problem according to the cost and time manipulation, especially the wet system. Fortunately, the quantity of this research may not be at a reliable level. The well-known and open-access datasets are further studied to increase the reliability of the proposed method.

Data Preprocessing

In recorded signals, there are missing values caused by the device connection. So, the signals were manipulated by duplicating the previous 8 samples. Then, a notch filter was applied to remove the 60 Hz power line noise and applied artifact removal by using the EEGLAB [13] toolbox to avoid the severe contamination of EEG signals automatically. The sampling rate of wet sensors was downsampled into 256 Hz, which is equivalent to the dry sensors. A sliding window technique was applied with 2-second window size and 1-second overlapping for seg-

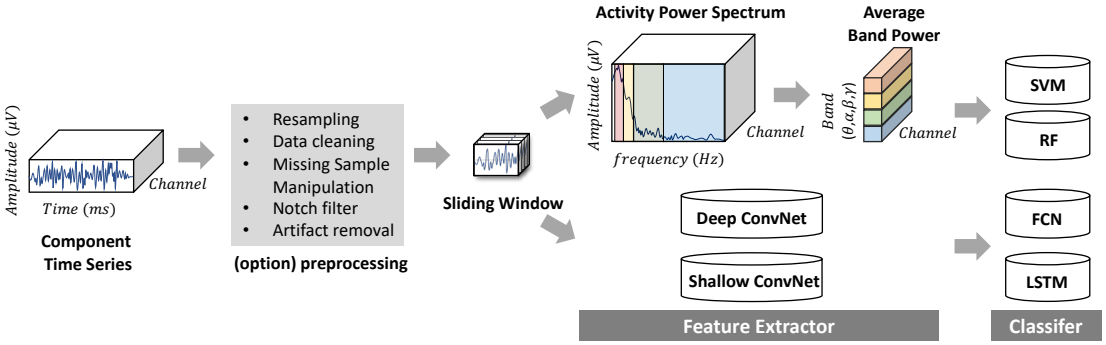

Figure 4.3: Data preprocessing and feature extraction

mentation and sequential analysis. For the baseline classification, the PSD was calculated from each window sequentially as the extracted features: theta band (4-7 Hz.), alpha band (8-12 Hz.), and beta band (13-32 Hz.). The process is shown at the top of Figure 4.3. While ConvNet's feature extraction of each window, the convolution is applied to the raw signals for learning the temporal and spatial filtering automatically and directly without any signal preprocessing. The learning ability of a neural network can also deal with noises or artifacts from unprocessed signals.

In the attention dataset, the data sequence individually consists of 84 pairs of input data and labels obtaining from 19 baseline windows and 65 doing-task windows. Each system entirely contains 420 pairs of window input and label output. Thus, there are 25 pairs of train and test samples for each system, but the input data was sliced into sliding windows, and all of them were paired with the same emotion label. Each system of music-emotion dataset totally consists of 870 pairs of input windows and labels.

In DEAP and SEED dataset, the input was initially processed by the provider, as mentioned in Section 2.4. A 2-second window sliding was applied with a 1-second overlap to slice the raw data into window input for classification tasks. To simplify the DEAP's continuous scales, the rating was classified into two classes of emotion: the positive class is positive arousal and valence, and the rest is the negative class. While in the SEED dataset, the positive and negative classes were selected. Thus, all experiments are a binary classification with different chance levels according to the dataset.

4.4 Statistical Analysis

This section shows the difference between dry and wet electrodes by calculating the statistical values and frequency bands. Firstly, this study investigates the raw signal through the range, mean, and standard deviation, which shows the statistical report of voltages and features calculated from all subjects data in Figure 4.4, 4.5 and Table 4.1. These can easily notice the diversity of both systems.

Table 4.1 shows the significant wet-dry relationship and also the relationship between attention and music-emotion tasks. The wide range of dry system on both tasks shows signal instability and susceptibility of the sensor, but the wet system is more stable and sensible. While considering the task relevances, the attention experiments, the subject is required to do the 3-back task intently with prompt and continuous response. It caused the signal stability, which can observe from the shallow range. In contrast, in the music-emotion experiment's recording session, the subject did not move while listening to music, but the higher range is acquired. These mentions that music listening varies the raw data or gain high frequency than the attention task.

Considering the example of attention experiment, Figure 4.6 shows the topoplots and histograms of subject 4. As the related research about the frontal brain activity, the attention experiment shows the relative result. These findings are consistent with the results of frontal brain activity reported by many studies. The attention class is associated with increasing low-frequency (theta) and mid-frequency bands (alpha), especially left-right electrodes: T3, F7, F8, and T4. Whereas, the high-frequency band, the beta band, decreases when doing the task. These frequencies reflect attentional requests such as alertness and memorization. Unfortunately, there is an inconsistent part of the increasing and decreasing between baseline and task. These might be affected by the baseline focusing, which the subject might highly pay attention to the '+' character. On the other hand, considering the system comparison, it could see both systems' tendencies fortunately. Notably, the theta and beta band of the left-right frontal brain identifies the increase while doing the 3-back task in both wet and dry systems. For the beta band, it also has decrease tendencies comparatively.

On the other hand, in the example of the music-listening task, which shows in Figure 4.7, the

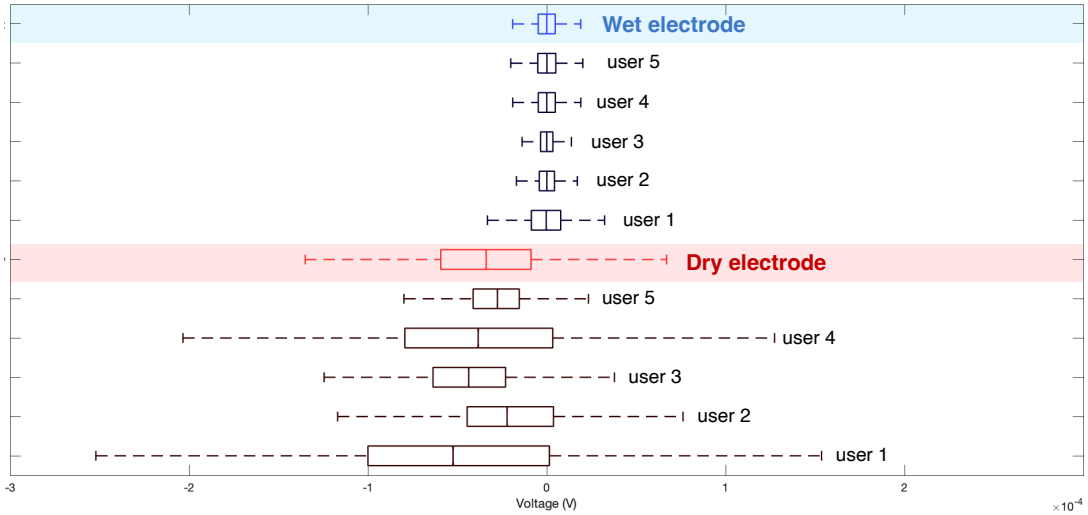

Figure 4.4: Box plot shows the variation of raw data signals (voltage) and comparison between dry and wet system in music-emotion analysis.

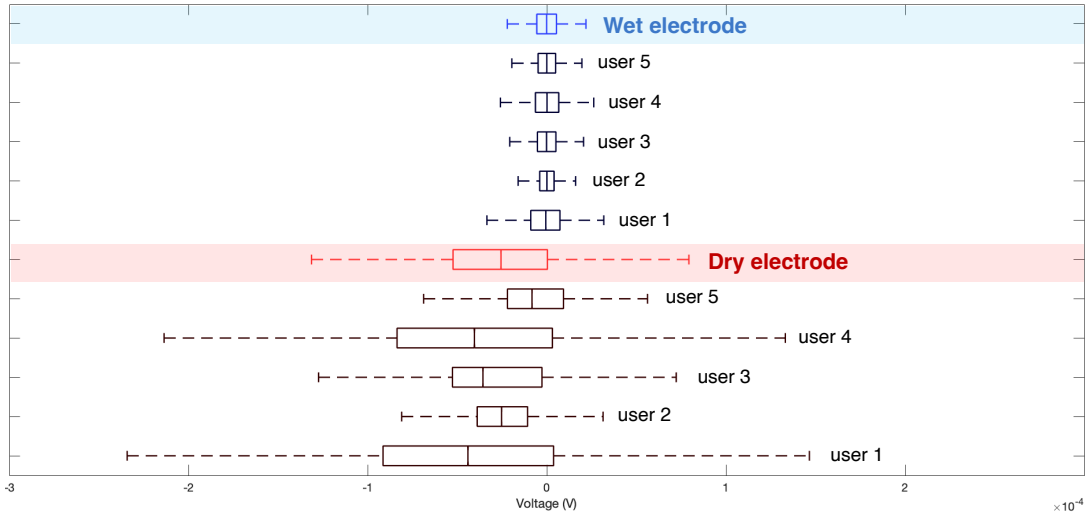

Figure 4.5: Box plot shows the variation of raw data signals (voltage) and comparison between dry and wet system in 3-back task.

Table 4.1: The statistical report of voltages and 3 frequency band features in 4 collected datasets

Task	Type	Raw data (μV)			PSD features ($\times 10^{-11}$)					
					Theta		Alpha		Beta	
		Range	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Emotion-Music	dry	21764.39	-34.65	68.65	9.65	6.21	5.22	3.48	5.61	2.32
	wet	1783.34	0.01	11.06	0.09	0.05	0.08	0.05	0.13	0.07
3-Back Task	dry	727.47	-26.25	39.07	8.64	3.46	4.51	1.40	5.54	1.21
	wet	486.42	-0.04	12.54	0.11	0.06	0.07	0.04	0.18	0.05

PSD features of subject 2 is plotted. The subject questionnaire is positive valence and arousal, which is a positive class. The related research associates the observation that the left frontal EEG activity is greater relative to happy and joyful music. This study observe that theta and alpha bands have a similarity in increasing while the beta band is contrast. The high-frequency features might be sensitive and contaminated by the environment or unrelated noises.

4.5 Methodology

The proposed model utilizes ConvNet and LSTM network for binary classification of cognitive tasks. For feature extraction, the convolution network is implemented for the brain-signal decoding method. Deep ConvNet and Shallow ConvNet were applied in this experiment to show the decoding. The original Deep ConvNet and Shallow ConvNet performs Fully Connected Networks (FCN) as a classifier in the final layer to discriminate the classes [49]. This work applied the LSTM network instead to perform the convolution features. Firstly, the raw signals are reshaped into the 2D shape, consisting of time and electrode channels, as the input for feature learning layers.

The setting of Deep Convnet is shown in Figure 4.8 (a). The input is convolved with 24 temporal filters, and then 24 spatial filters are sequentially performed. The deep features of temporal-spatial filtering are extracted by 3 blocks of convolution and pooling layers with twice the number of filters in each block (48, 96, 192). the batch normalization and activation with an exponential linear unit (ELU) are applied for each convolution and pooling layer. On the other hand, the setting of Shallow ConvNet and LSTM networks is shown in Figure 4.8 (b). This network is more compatible and less trainable parameters than the previous one and can effectively learn the EEG features. The model contains 40 temporal and spatial filtering constructed by convolution networks continuously. Then, the network applies the batch normalization layer and ELU activation after convolving the features. Finally, the values are averaged by the pooling layer before sending it to the classification layers. Consequently, the extracted features of Deep ConvNet or Shallow ConvNet are conveyed to two layers of 50 blocks LSTM networks for learning dependencies of feature sequence and summarized into predicted class. In the final classification layer, the output is activated with a sigmoid function. Their kernel size and pool

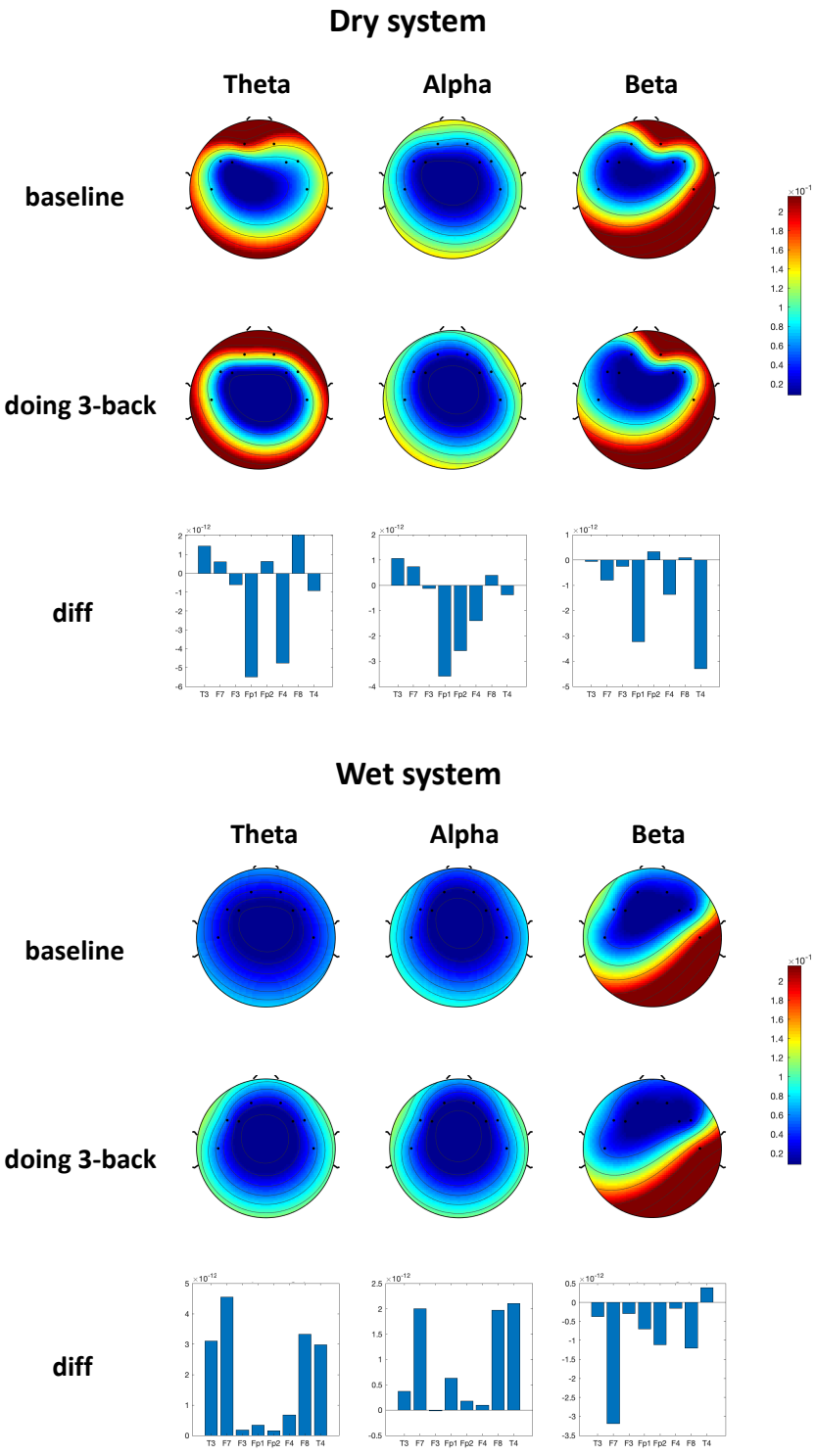

Figure 4.6: Topoplots and histograms to show a comparison between dry and wet systems on the 3-back task of subject 4. The rows denote the frequency bands: theta, alpha, and beta respectively. The columns denote the baseline and 3-back task sessions, and the last column is different between these sessions.

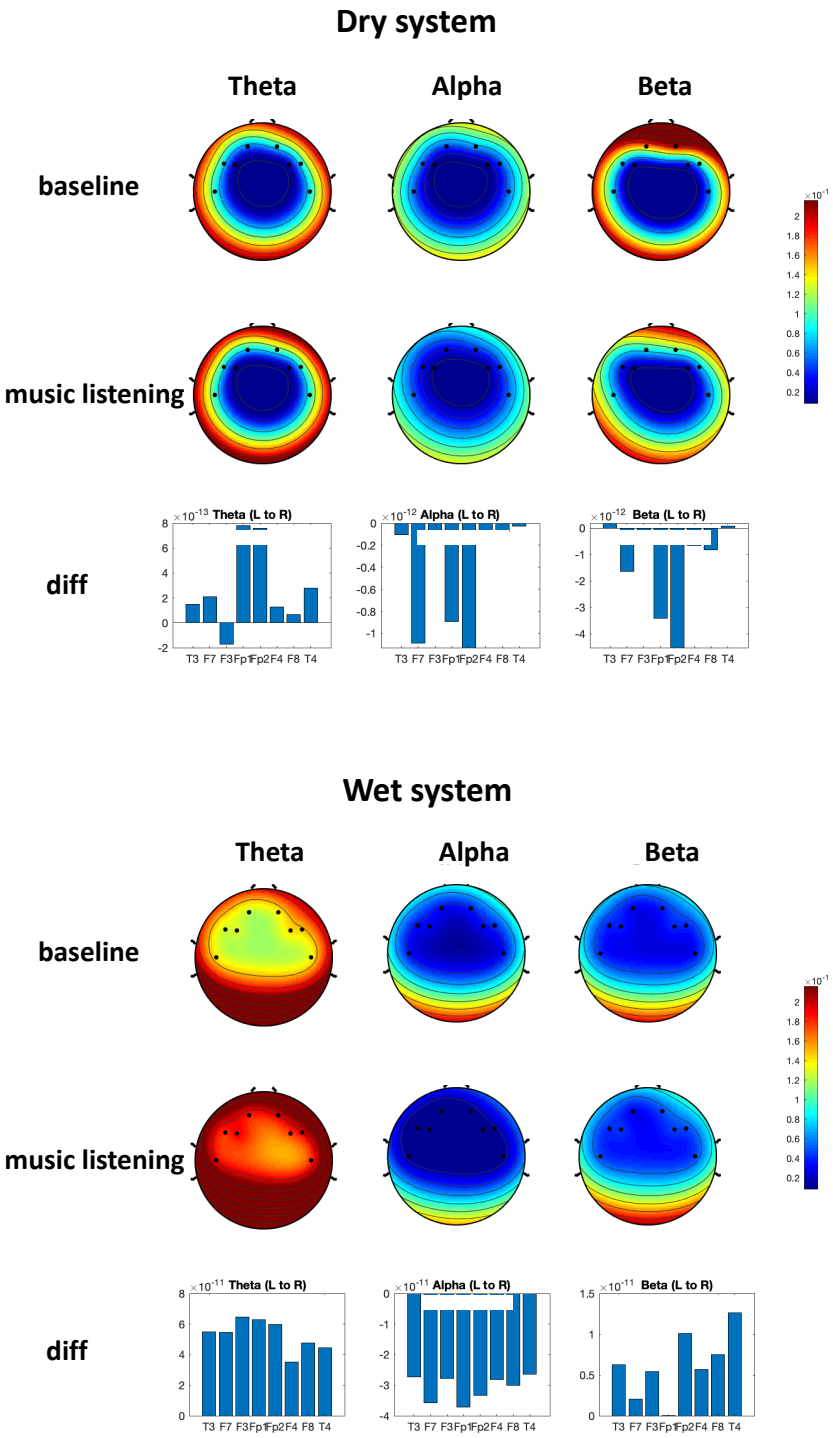

Figure 4.7: Topoplots and histograms to show a comparison between dry and wet systems on the music-listening task of subject 2. The rows denote the frequency bands: theta, alpha, and beta respectively. The columns denote the baseline and music-listening sessions, and the last column is different between these sessions.

size are adjusted based on each dataset's window size and sampling rate. Figure 4.8 (a) and (b) show the setting of the dataset which is 512 window size, 256 sampling rate, and 8 electrode channels.

For transfer learning, this study transfers the knowledge of feature extraction from the wet system to the dry system. The implementation is the freezing of feature extractor or ConvNet in the wet system and then apply it to the dry system. Then, the LSTM layers are retrained to classify the attention or emotion classes.

4.6 Experiments and Results

This section presents the binary classification learning implemented on self-collected datasets and existing-datasets. The baseline models are random forest (RF) and SVM with radius basis function (RBF) classifiers combined with PSD features. The state-of-the-art models from [49] are Deep ConvNet and Shallow ConvNet with FCNs. Moreover, the proposed models are Deep ConvNet and Shallow ConvNet with LSTM networks shown in Section 2.3. However, these small datasets may cause an overfitting problem, and it can not transcend the reliability of data quantity. All techniques on the SEED and DEAP dataset further are applied to prove to model performance. All models implemented the leave-one-out (LOO) cross-validation for all subjects to investigate the subject-independent performance. Self-implementation of all techniques was performed in the same environment and technical setting. For baseline settings, the grid search was applied for hyper-parameter tuning to select the setting in these experiments. All neural network architectures were fitted using Adam optimizer [34] to minimize the binary cross-entropy loss function with 100 epochs and 16 batch size for self-collected datasets but 64 batch size for existing datasets. The evaluation metric is the accuracy of binary classification even there is the problem of an unbalanced class. To investigate the learning performance in this problem, this study present the chance level, which is the majority class's accuracy when predicting constantly. A paired t-test is a statistical procedure used to determine the significant difference between the corresponding models. All the results are shown as the accuracy and standard deviation across all subjects in Table 4.2 and the significant comparison by histograms in Figure 4.9.

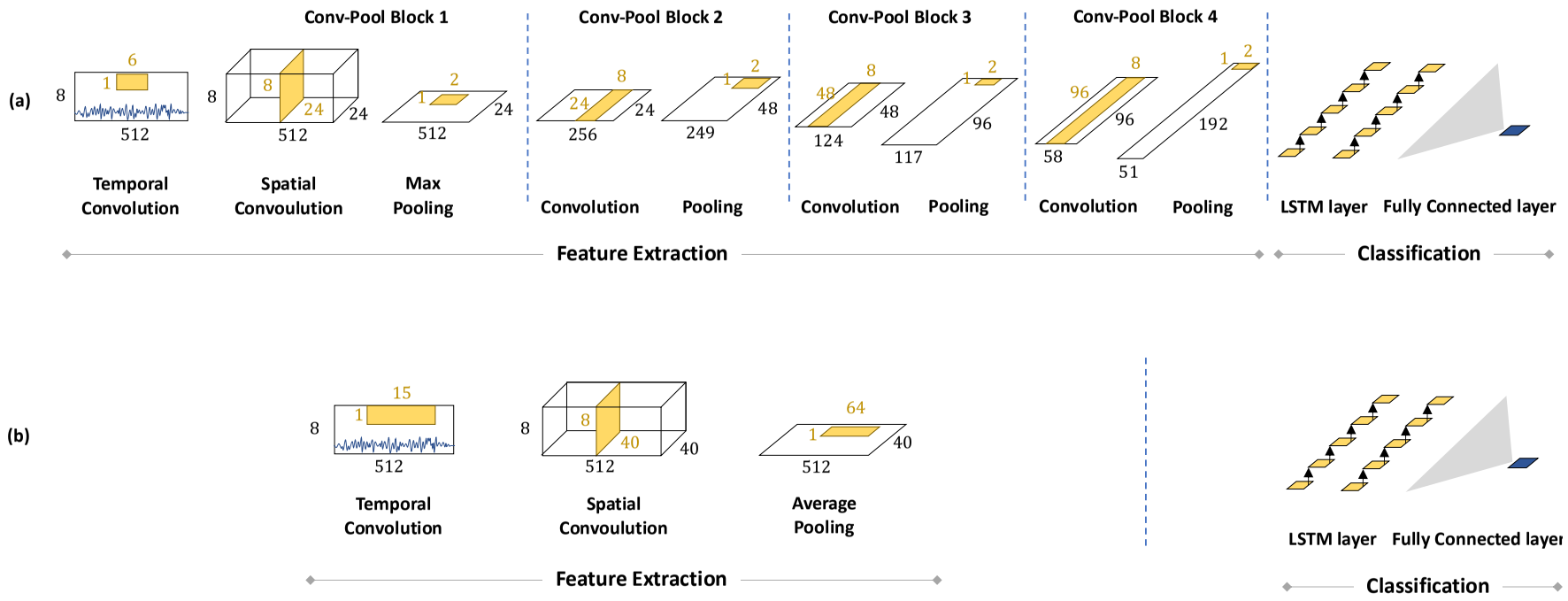

Figure 4.8: The network structures. (a) Deep ConvNet and LSTM networks (b) Shallow ConvNet and LSTM networks

Table 4.2: Classification results determine with accuracy and standard deviation of all subjects. The bolds denote the highest accuracy among the corresponding dataset.

		Baseline with PSD		ConvNet with FCN		ConvNet with LSTM	
Dataset	Chance Lvl	RBF-SVM	RF	Deep	Shallow	Deep	Shallow
Attention (Dry)	77.4±0.0	81.8±2.1	81.8±2.1	82.5±12.2	84.6±7.7	81.6±14.0	**90.0±4.5**
Attention (Wet)	77.4±0.0	84.2±1.0	84.2±1.0	88.5±3.4	90.5±4.4	90.0±6.0	**92.0±6.8**
Music-Emotion (Dry)	70.2±16.2	68.1±23.4	69.1±19.5	62.4±17.3	59.25±24.8	**71.7±15.5**	68.0±19.3
Music-Emotion (Wet)	71.5±20.4	69.5±23.0	**70.3±22.5**	67.4±21.2	68.7±24.7	67.7±21.2	67.4±20.9
DEAP [37]	64.2±11.2	53.4±7.2	**60.4±10.1**	56.4±10.0	58.8±13.7	59.8±12.1	**60.4±12.4**
SEED [59]	50.0±0.0	51.0±0.6	74.0±8.2	72.0±7.8	74.6±9.3	73.0±7.8	**75.3±9.3**

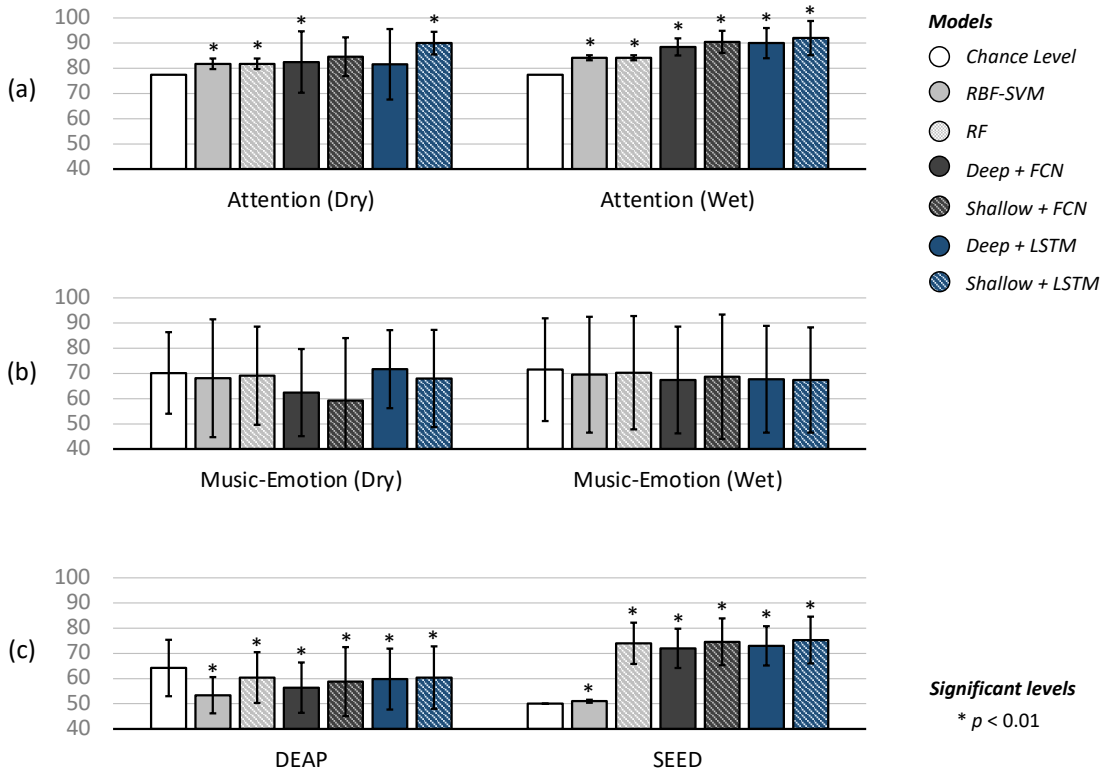

Figure 4.9: Classification results; error bars denote the standard deviation for all subjects and stars indicate a significant improvement compared to chance levels.

4.6.1 Self-collected dataset Classification

Based on the overall results in 5 subjects cross-validation, the attention classification is easier classified than music-emotion classification, which is observed by the average accuracy and standard deviation. A high standard deviation can show the subject variation and complexity.

Attention Classification

In the attention results shown in Figure 4.9 (a), there are the same chance levels of 77.5% $\pm$0.0% accuracy in both wet and dry systems. All classification models achieved significantly higher results than the chance level ($p<0.01$), except for Shallow ConvNet with FCN and Deep ConvNet with LSTM. The proposed models, Shallow ConvNet with LSTM network, achieved the highest accuracy in both dry and wet attention experiments: 90.0%$\pm$4.5% and 92.0%$\pm$6.8%, respectively. The study observed that the accuracies of Shallow ConvNet with LSTM are higher than the chance level ($p<0.01$) and baseline techniques ($p<0.05$). Moreover, this experiment shows the improvement of replacing FCN with LSTM network on attention classification, espe-

cially combining of Shallow ConvNet and LSTM network. Considering the system comparison, overall performances can learn with higher consistency observed in the models' accuracies and standard deviations.

Music-Emotion Classification

While in music-emotion experiments, all results, except Deep ConvNet with LSTM in dry system, can not significantly outperform the chance levels ($p>0.05$), which are of $70.2\%\pm16.2\%$ of dry system and $71.5\%\pm20.4\%$ of wet system. Classification results are lower than the attention results explicitly. All results are shown in Figure 4.9 (b) and found that no significant difference in performance across all models ($p>0.05$). The highest performance is the Deep ConvNet with LSTM network in dry system ($71.7\%\pm15.5\%$) and RF in wet system ($70.3\%\pm22.5\%$).

4.6.2 Wet-Dry Transfer Learning

Applying the transfer learning strategy, this experiment focused on improving the dry system using the transfer learning from the wet system. In the attention experiment, the new accuracies of Deep ConvNet and Shallow ConvNet with retained LSTM networks are $88.2\%\pm3.7\%$ (+6.6% from $81.6\%\pm14.0$) and $90.0\%\pm4.5\%$ (+0.0% from $90.0\%\pm4.5\%$), respectively. While in the music-emotion experiment, the new accuracies are $67.1\%\pm18.0\%$ (-4.6% from $71.7\%\pm15.5\%$) and $69.2\%\pm19.7\%$ (+1.2% from $68.0\%\pm19.3\%$), respectively. The results show the small improvement of Deep ConvNet in the attention experiment and Shallow ConvNet in the music-emotion experiment without significance level ($p>0.05$).

4.6.3 Existing-Dataset Classification

In the DEAP dataset results, the study classified the binary classes with $64.2\%\pm11.2\%$ chance level among 32 subjects. Fortunately, all classification models could not outperform the chance level significantly ($p>0.05$). However, Shallow ConvNet and RF similarly achieved the highest accuracy with different standard deviations at $60.4\%\pm10.1\%$ and $60.4\%\pm12.4\%$, respectively.

For the SEED dataset, the chance level is $50.0\%\pm0.0\%$, which is calculated from the balanced-label stimulus on 15 subjects. All models were superior to the chance level significantly ($p<0.01$). Again, the Shallow ConvNet achieved the highest accuracy with $75.3\%\pm9.3\%$

accuracy. Replaced FCN with LSTM network of ConvNet-based models can improve the performance of both datasets, especially the Shallow ConvNet with LSTM network outperformed on both existing datasets.

4.7 Discussion

This chapter discuss on 3 aspects consisting of system comparison, dataset comparison, and model comparison.

4.7.1 Wet and Dry systems

To acquire and analyze brain activity, a sensor device to record EEG brain-signal is needed. This research conducted the experiments with the dry and wet systems on the same tasks to investigate two types of electrode. According to the signal quality, the wet systems are better than dry systems obviously. For classification results, the wet systems also outperformed dry systems, which were observed by the average accuracy. Even though there is a data quantity problem in this study, the statistical analysis in Section 4.4 shows the relationship between wet and dry systems. Fortunately, the classification results also show related tendencies and characteristics following the corresponding tasks. From the wet-dry transfer learning experiment, this study inconsiderably discovered the possibility of using a wearable sensor in a real-world application by transferring the knowledge from laboratory sensors.

4.7.2 Dataset Comparison

Considering the dataset comparison, this study recorded the EEG signals with two different cognitive tasks and two different systems from 5 subjects. After performing classification with machine learning techniques, the study realized that the self dataset is too small for machine learning, and it is not reliable to verify the model performance. We further applied the proposed model to the existing datasets, which are more subjects and more stimuli. Based on cognitive tasks, all datasets were divided into attention and emotion datasets. The attention dataset is only the 3-back task, which classifies the attention and non-attention classes, while the rest are emotion classification with different stimuli. In the self dataset, the subject is stimulated

by music listening. DEAP's stimuli are music video excerpts, and SEED's stimuli are movie excerpts. The study found that subjects' emotions were elicited from different sources, and then their labels were obtained from the questionnaires after performing the task. The emotional scale might be different depending on the individual subject's feelings. The variety can be occurred by emotion, subject, or questionnaires individually. However, SEED provided the labels derived specifically from the movie excitation, not from the actual subject's emotion. These may generalize the complexity and affect the classification performance. The results show that the music-emotion datasets are not explicit separable and more difficult tasks than the attention experiments. Meanwhile, the DEAP dataset is more complicated than the SEED dataset.

4.7.3 EEG Classification Model

Analyzing the baseline techniques, SVM can not perform properly because of the sensitivity to the noise, while the ensemble modeling, like RF, performs good results in all datasets with multiple decision tree models. It can learn to reduce the generalization error of the prediction, and also can deal with subject-independence of the EEG classification. Additionally, the raw signals might include environmental noise, muscle artifact, or eye-movement artifact. PSD features were directly extracted from these raw signals with predefined and unlearnable preprocessing. It means that the EEG signals with more noise directly and easily affect the learning performance, and it is not appropriate to BCI practical applications. In contrast, considering the ConvNet models, training with neural networks can automatically learn additional preprocessing without predefined knowledge. The architecture of the ConvNets is used as a feature extractor or EEG decoder, and then it needs the added layer for discriminative learning. Instead of using the simple fully connected layer, which individually and independently learns the feature perception on convolved features, the LSTM network can interpret the dependencies across time steps efficiently. Based on the classification results, it is helpful for temporal-spatial learning like multivariate time-series classification, such as brainwave signals. Considering the results, it might not see the significant differences between these two ConvNet, but the Shallow ConvNet applied less parameter than the Deep ConvNet. Moreover, adding the sample can retrain

the model with new additional data with pre-trained knowledge from the existing datasets or subjects.

Chapter 5

Multi-kernel Temporal and Spatial Convolution Network

5.1 Overview

Deep learning using an end-to-end convolutional neural network (ConvNet) has been applied to several electroencephalography (EEG)-based brain-computer interface tasks to extract feature maps and classify the target output. However, the EEG analysis remains challenging since it requires consideration of various architectural design components that influence the representational ability of extracted features. This study proposes an EEG-based emotion classification model called the multi-kernel temporal and spatial convolution network (MultiT-S ConvNet). The multi-scale kernel is used in the model to learn various time resolutions, and separable convolutions are applied to find related spatial patterns. In addition, this study enhanced both the temporal and spatial filters with a lightweight gating mechanism. To validate the performance and classification accuracy of MultiT-S ConvNet, subject-dependent and subject-independent experiments are conducted on EEG-based emotion datasets: DEAP and SEED. Compared with existing methods, MultiT-S ConvNet outperforms with higher accuracy results and a few trainable parameters. Moreover, the proposed multi-scale module in temporal filtering enables extracting a wide range of EEG representations, covering short- to long-wavelength components. This module could be further implemented in any model of EEG-based convolution networks, and its ability potentially improves the model's learning capacity.

5.2 Problem Statement

For learning the feature representation, the design of ConvNets is essential to achieve a promising objective. The improvement of learning capability can be considered in several ways, such as increasing the number of layers and modifying kernel. The kernel or filter acts as a regularizer to improve the generalization at the cost of more complicated training. It performs the dot product with the sub-region of input and outputs the matrix of dot products. ConvNet model traditionally and generally applied the single-size kernel as a convolution operator to extract the feature from the raw EEG signals. The larger kernel needs a high computational cost, while the smaller kernel captures lower representations. However, achieving good EEG data analysis performance requires a wide range of useful representations based on the functional characteristics of frequency bands. Therefore, the learning model should consider the learning capability of kernel size while preserving optimized trainable parameters. This study proposes an architecture called multi-kernel temporal and spatial convolution for EEG-based emotion classification. This study presents (1) multi-kernel learning for temporal convolution and (2) filter recalibration with a lightweight gating mechanism. The proposed model makes the classification process more efficient by factoring the kernel into a series of operations to capture various short and long patterns. Then, separable convolution is applied for spatial learning by considering each channel separately and convolving over electrode channels. Moreover, weights are recalibrated after temporal and spatial convolving to utilize the limited data available. MultiT-S ConvNet is compared with existing techniques and investigate task accuracy by conducting classification experiments on BCI datasets.

5.3 Multi-kernel Temporal and Spatial Convolution Networks

This study proposes an end-to-end CNN called multi-kernel temporal and spatial convolution (MultiT-S ConvNet) for EEG-based emotion recognition. The proposed model enables multi-scale representation learning and improves classification performance. The model consists of three parts, namely, a temporal learner, spatial learner, and classifier, which simultaneously learn discriminative representations in the time and electrode channel dimensions. The model

architecture is depicted in Figure 5.1, and the details are described in the following sections.

5.3.1 Multi-kernel Temporal Processing

This study hypothesizes that multi-kernel filters can improve temporal representations learned from raw EEG data. Considering the kernel design of temporal convolution, the larger size can learn higher resolutions in the time domain but necessitates high computational costs. On the other hand, a smaller kernel size essentially learns low-level features or shorter temporal patterns in the time domain and can reduce computational costs. Multi-scale convolution kernels combining long and short patterns are applied by factoring a kernel into various kernel sizes that would independently convolve and map them in order. Learning a wide range of representations while avoiding high computational costs can improve the performance of EEG classification. The main advantage of this architecture is significant quality gain at a modest increase in computational requirements compared with shallower and narrower architectures. Accordingly, the study specifically adopted four kernels of length 25 ms (= 5 samples under the sampling rate of 200 Hz), 50 ms (= 10 samples), 100 ms (= 20 samples), and 200 ms (= 40 samples). These choices were based on four frequency bands, namely, theta (4–7 Hz), alpha (8–12 Hz), beta (13–32 Hz), and gamma (>33 Hz), extensively used to characterize brain states/activities. The sampling rate is also considered because wavelengths can be captured with the same time scale, even from different datasets. These multi-kernels can capture an extensive range of representations in EEG signals. The larger kernel size can learn various informative features, along longer temporal patterns. Meanwhile, the smaller kernel size specifically extracts shorter temporal patterns.

The raw EEG signals in Figure 5.1 can be represented as 2D time series whose dimensions are time (T) and electrode channels (E). The window size is set as 2 s (T is 400 data points for the SEED dataset) based on the change in emotion over time. This study duplicates the input into four modules and convolve them using four different kernel sizes: (1×5), (1×10), (1×20), and (1×40). Each module is padded with zeros to retain the same size. Along the multi-kernel temporal convolution, a tensor of size $(T' \times E \times F^{te})$ is produced, where $T' \times$ denotes the temporal dimensionality after convolution and F^{te} denotes the total number of temporal features

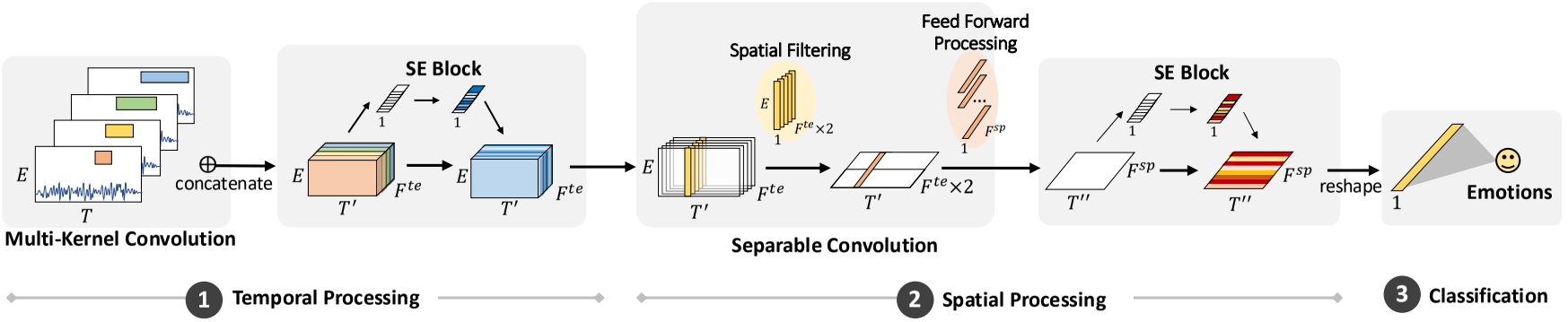

Figure 5.1: MultiT-S ConvNet architecture. The model comprises temporal filtering, spatial filtering, and classification layers. The first layer employs four types of temporal convolutions to learn multiple temporal features from raw EEG signals. Then, the separable convolution is used to learn temporal-specific spatial filters across electrode channels. The SE block is used to recalibrate the features of both filters. The final classification layer is used to discriminate the emotional labels.

after concatenation. All convolutions, including those modules, apply rectified linear activation (ReLU) and average pooling to data before sending them to the next layer.

5.3.2 Remaining Modules

Spatial Processing

For each time step (T') and temporal feature (F^{te}), the study collected information across electrodes using two full-length filters. A separable convolution was applied for spatial feature learning and to reduce computational costs. Figure 5.2 depicts the separable convolution and the virtualization of topoplots. It consists of two steps: spatial filtering and feed-forward processing. Spatial filtering connects each temporal feature map individually to learn specific spatial filters across electrodes. This study manually set the feature multiplier to 2 to increase the number of parameters in the neural network to learn more traits better. The network gathers data with the $F^{te} \times 2$ of the respective filter $(1 \times E)$ and stacks the outputs into feature maps $(T' \times 1 \times (F^{te} \times 2))$. The interpretation of this step is to build multiple filters that can learn informative features across all electrode channels individually. Then, feed-forward processing continues to learn and optimizes a temporal-–spatial summary for each feature map with independent filters $(1 \times 1 \times (F^{te} \times 2))$ across the spatial feature maps. Here the study learns how to weigh the useful set of filters from the previous extraction. After this spatial processing, temporal-–spatial feature maps $(T'' \times 1 \times F^{sp})$ are extracted, where T'' represents the time after temporal-–spatial processing and pooling.

Channel Recalibration

From the outputs, after each temporal and spatial processing, this study needs to utilize the limited dataset as much as possible regarding the time-cost consumption of EEG recording. The squeeze-and-excitation block (SE block) is further applied to adaptively recalibrate the gathering of informative feature responses by explicitly modeling interdependencies between feature maps [26]. Adding SE blocks after the convolutional layers can help model weighting that learns to emphasize and dismiss informative features. The weights of each feature map are equally informative after performing temporal convolution in Section 5.3.1. The temporal

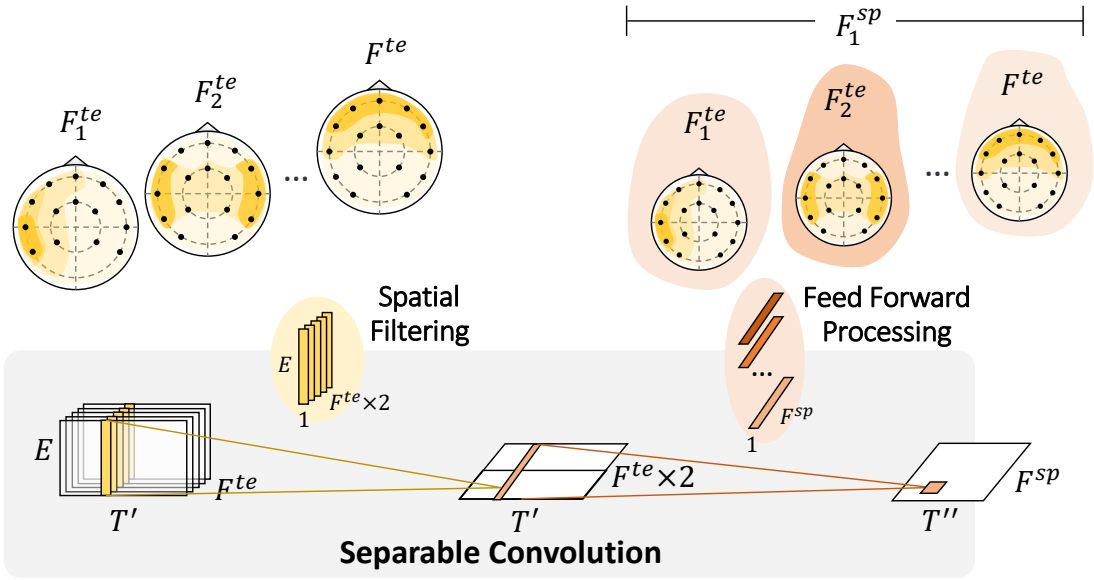

Figure 5.2: Spatial Convolution. Spatial processing separately convolves across the electrodes on the temporal feature maps. Feedforward processing subsequently weights the feature maps to extract temporal-–spatial features. Shading represents how the network emphasizes or understates the corresponding feature maps.

SE block is briefly described in two steps. First, in the squeezing step, global information is compressed into a feature descriptor $(1 \times 1 \times F^{te})$ using average pooling. Next, a dense layer is added with ReLU activation to reduce the channel complexity by a ratio (r) while r is set to 16. Then, another dense layer with sigmoid activation is added to give each channel a smooth gating function. Finally, the excitation step obtains the aggregated information and fully captures the channel-wise dependencies. According to the temporal feature maps, the features are multiplied by the weights of the temporal SE block and output $(T' \times E \times F^{te})$.

Moreover, after spatial processing in Section 5.3.2, the spatial SE is applied to adaptively recalibrate temporal-spatial feature responses. The squeeze step is to compress global temporal-spatial information into a feature descriptor $(1 \times 1 \times F^{sp})$. Next, two dense layers reduce the channel complexity $(r = 16)$. In the squeezing step, the temporal-–spatial feature maps are multiplied by the weights of the spatial SE block. After performing these convolutions, the extracted features are generated and transmitted to the classification layer.

Classification

The outputs from the previous module are flattened into one vector representing all extracted features and linearly transformed into classification logits with an FCN. For binary classification, the final layer unit is 1. The loss function is binary cross-entropy, and the activation function is a sigmoid function. While 3 classes classification, the final layer unit is 3. The loss function is categorical cross-entropy, and the activation function is a softmax function. The final output is a discrete emotion label with probability.

5.4 Experiment Settings

To validate the method's performance, feature extractability, and classification accuracy, these studies examined and compared it with existing methods. The experiments and results consist of subject-dependent and independent classification on the SEED and DEAP datasets.

5.4.1 Dataset

This study conducted SEED experiments in a three-class classification, including positive, neutral, and negative labels, and a binary classification with less complexity and fewer data considered positive and negative labels. On the other hand, DEAP employed 2D models rated by subjects caused discrepancies between movie types and real subject emotions. To simplify the problem, this study investigates a binary classification of valence and arousal scores with positive and negative labels.

5.4.2 Baseline and Existing Models

In baseline experiments, the PSD with four bands was used as extracted features, along with ML classifiers. The baseline classifiers related to EEG analysis comprise the KNN, RF, and FCN. These models were tuned by grid search with the optimal hyperparameters selected, resulting in the most accurate performance. Accordingly, the number of neighbors for the KNN classification is set to 21. The number of RF trees is set to 100. The FCN has 2 hidden layers with 100 nodes, which are then forwarded to the output layer. The activation function is the logistic sigmoid function. Adam optimization is applied while training the FCN model.

Moreover, this study compares the performance of MultiT-S ConvNet against existing models, including the Deep ConvNet, Shallow ConvNet, and EEGNet models. The Deep ConvNet architecture as shown in Figure 2.3 (a) consists of one temporal convolution layer, four spatial convolution and pooling layers, and a classification layer in order. The Shallow ConvNet architecture as shown in Figure 2.3 (b) consists of single temporal and spatial convolution layers. Then, it sequentially passes data to a pooling layer and a classification layer. For the EEGNet architecture, this study refers to the original network as shown in Figure 2.3 (c), which consists of temporal convolution and a separable convolution layer followed by a classification layer. The settings of Deep ConvNet, Shallow ConvNet, and EEGNet are shown in Figure 5.3, Figure 5.4, and Figure 5.5, respectively. The MultiT-S ConvNet setting is depicted in Figure 5.6. The number of filters and trainable parameters are shown in Table 5.1. This study applied the Adam optimizer to all DL models in the training process. The loss function for binary classification is binary cross-entropy, whereas three-class classification is categorical cross-entropy. Each model was trained for 200 epochs with 100 batch sizes. These were trained on GPU, NVIDIA Quadro P6000 (Santa Clara, United States), using Tensorflow [1] and Keras [12].

5.4.3 Comparison Approaches

All models were examined on the subject-dependent and independent classification of the SEED and DEAP datasets. In the subject-dependent experiments, 5-fold cross-validation was applied, experimenting with different partitions of 80% training data and 20% test data; 30% of the training data were held out for validation of free parameters, and the model parameters were optimized on the basis of the remaining 70%. Moreover, the validation set was randomly picked on the desired subject, and the 5-fold cross-validation was applied to all subjects. For the subject-independent experiment, this study applied the leave-one-out cross-validation to access model performance. The model was trained using data from all but one subject and tested on the held-out subject; 30% of the training data were randomly selected for free-parameter validation without balancing among subjects.

The performance metric was classification accuracy. Moreover, the experiment reports show the chance level, which is the obtained accuracy when consistently predicting the majority

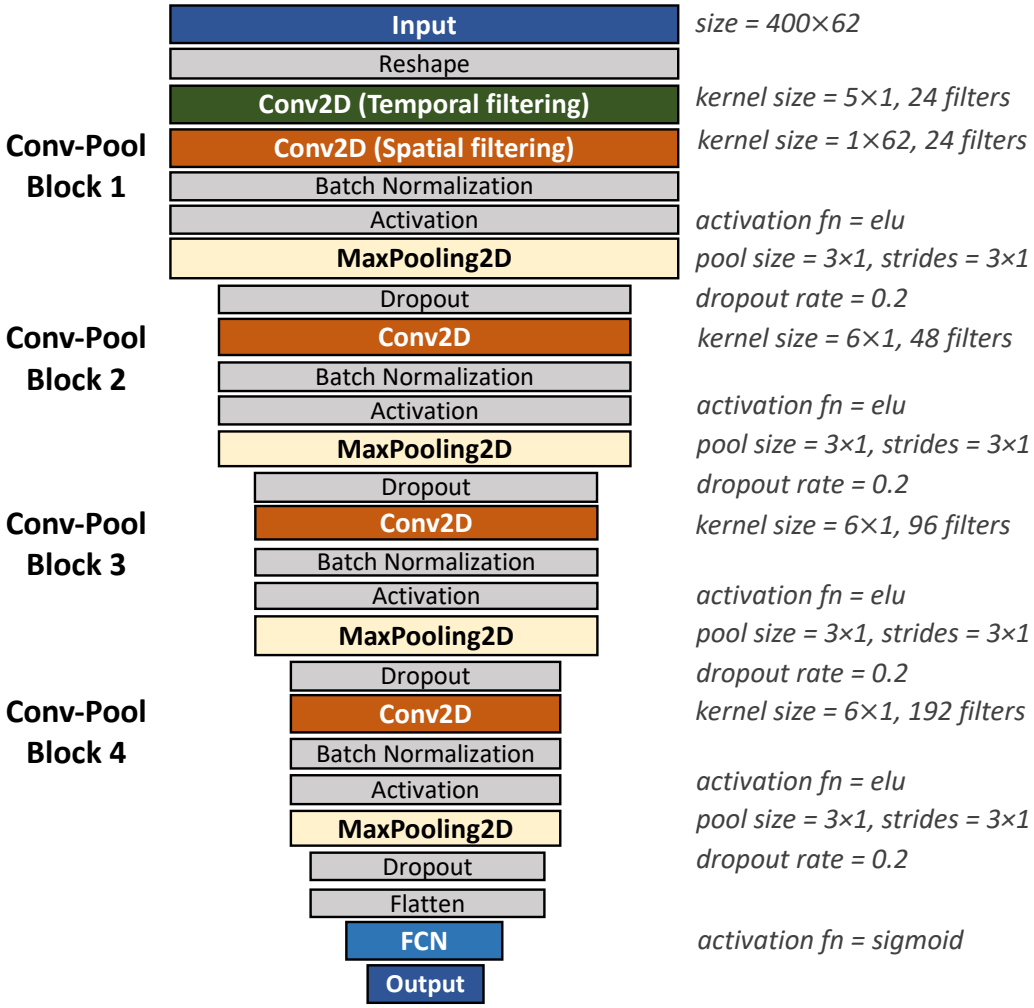

Figure 5.3: Deep ConvNet model setting.

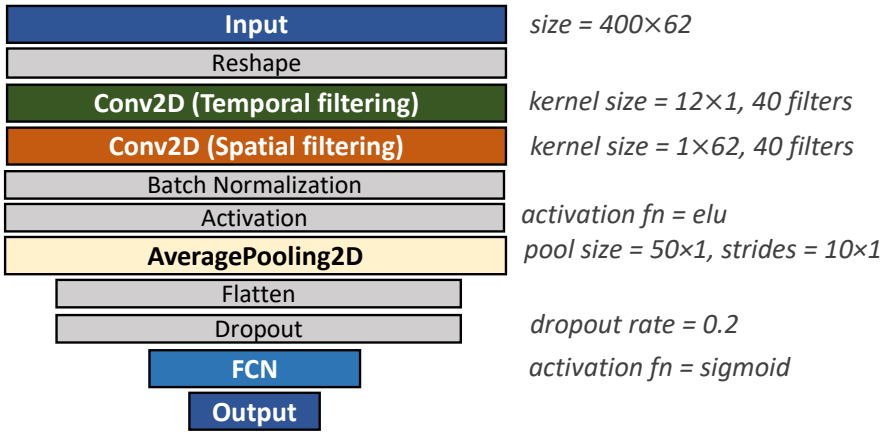

Figure 5.4: Shallow ConvNet model setting.

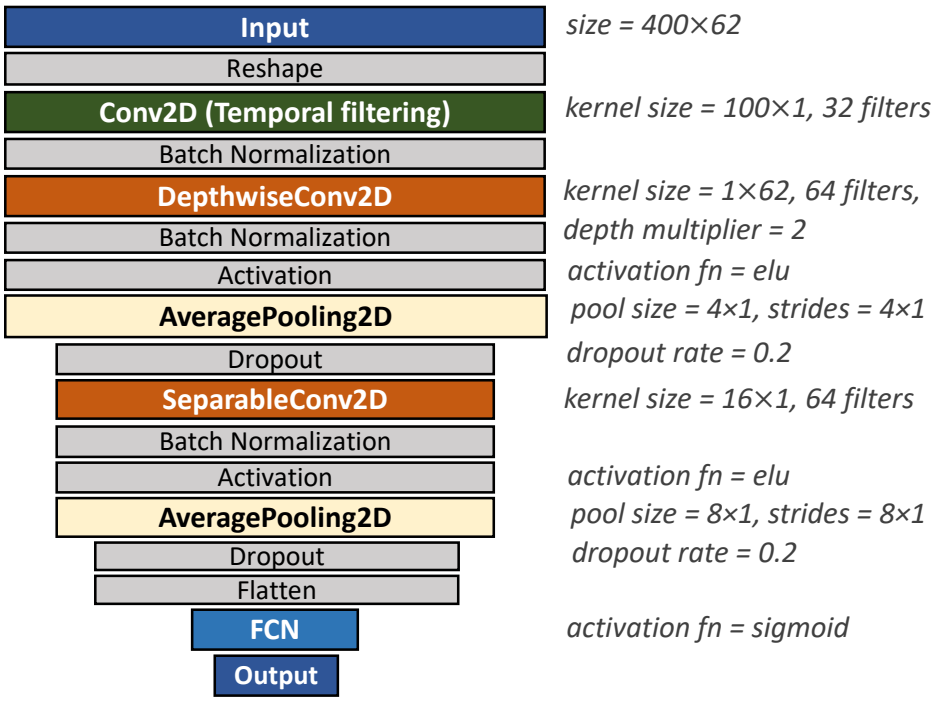

Figure 5.5: EEGNet model setting.

class. To verify multiple comparisons, analysis of variance (ANOVA) was applied to determine whether the means in the desired group were significantly different. Then, Dunnett's test was used to observe the many-to-one comparisons with the proposed method.

5.5 Results on SEED dataset

For the SEED dataset experiments, this study performed two-class and three-class classification to study both subject-dependent and subject-independent results. The two-class experiment uses two-thirds of the three-class dataset to simplify the problem's complexity and additional investigation. Their accuracy is shown in Figure 5.7 and Table 5.2. For the performance results, all models can achieve the 50% and 33.3% chance levels in the subject-dependent and subject-independent experiments, respectively. This work used one-way ANOVA to compare differences in the means of all models, and the study found that the SEED dataset's results are significant. The post hoc Dunnett's test is then conducted to compare every result with that of MultiT-S ConvNet for observing the significant difference. RF, Deep ConvNet, Shallow ConvNet, EEGNet, and MultiT-S ConvNet significantly outperform the chance levels ($p < 0.01$).

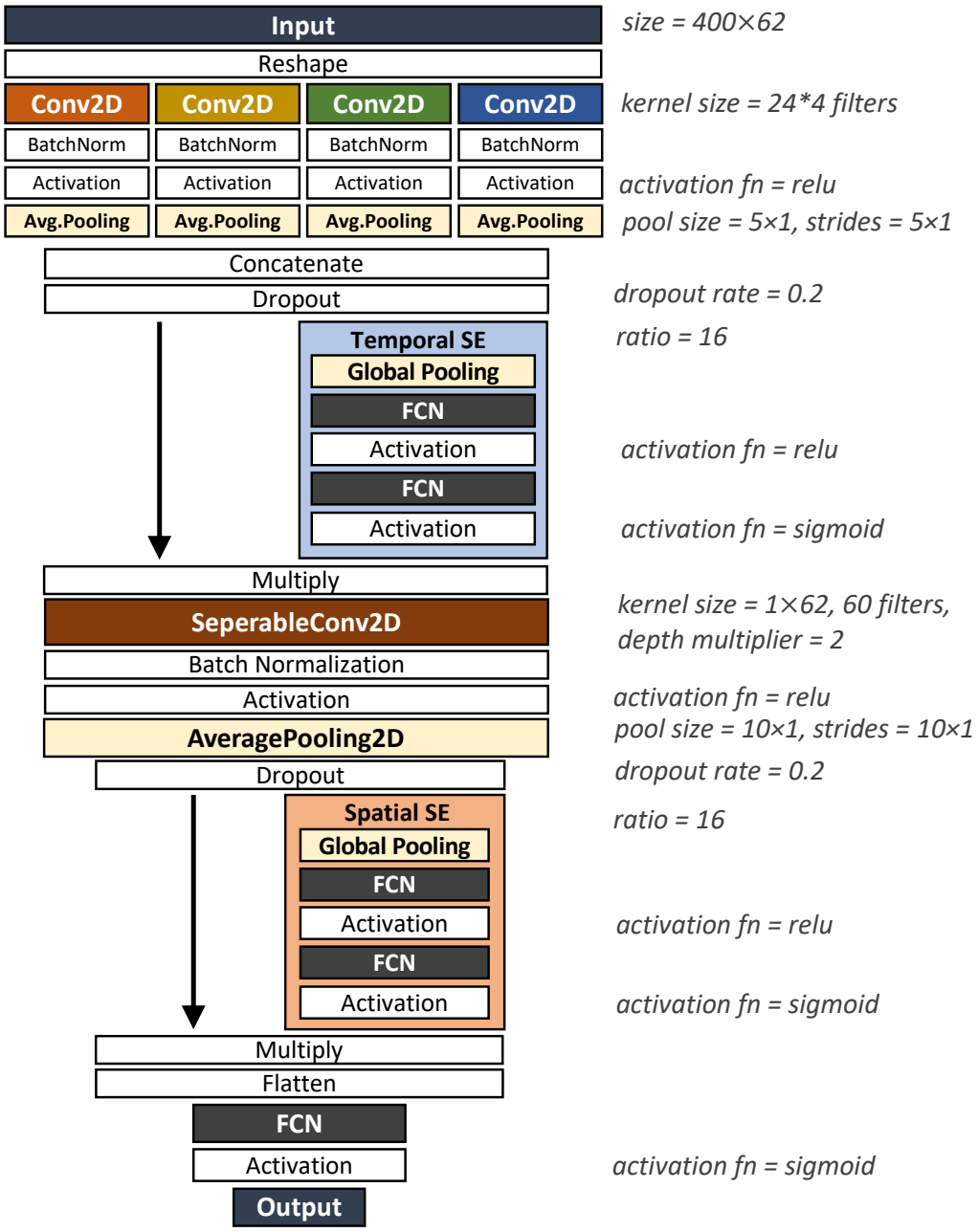

Figure 5.6: MultiT-S ConvNet model setting.

Moreover, MultiT-S ConvNet achieves the highest accuracy in all experiments. For two-class and three-class classifications, the subject-dependent results are 95.2%±3.2% and 86.0%±5.3% respectively, whereas the subject-independent results are 75.1%±8.4% and 54.6%±6.8% respectively. RF outperforms other baseline techniques by a significant level ($p < 0.05$). The accuracy of all DL models is significantly higher than KNN ($p < 0.01$) and FCN ($p < 0.05$) models. However, there is no statistical difference among deep learning models ($p > 0.05$), except

in three-class and subject-independent experiments, where the accuracy of MultiT-S ConvNet is significantly higher than that of EEGNet ($p < 0.05$).

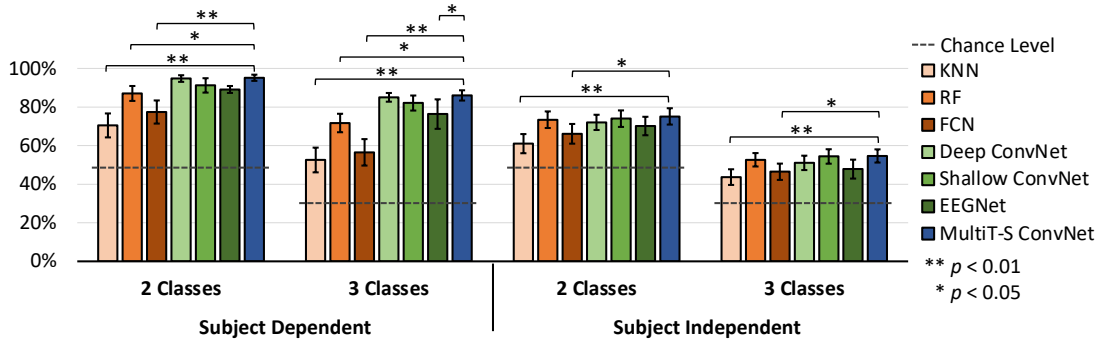

Figure 5.7: Average classification accuracy for all subjects on the SEED dataset. Error bars denote a 95% confidence interval (CI) computed from all subject's means. The horizontal dashed lines indicate the chance level. The stars indicate significant differences compared with MultiT-S ConvNet.

5.6 Results on DEAP dataset

Using the DEAP dataset, this study conducted four experiments of valence and arousal binary classification on both subject dependence and subject independence. The classes were defined by emotion scores, where 1 to 4 is a negative class, and 5 to 9 is a positive class. DEAP dataset's results are shown in Figure 5.8 and Table 5.2. Similar to the SEED dataset, one-way ANOVA with Dunnett's post hoc is applied to investigate significant differences among all models. The chance level of valence is 56.6% $\pm$ 9.2%, whereas the chance level of arousal is 58.9% $\pm$ 15.5%. For subject-dependent experiments, EEGNet outperforms the other models with a valence classification accuracy of 79.2% $\pm$ 5.8%, and MultiT-S ConvNet outperforms the other models with an arousal classification accuracy of 82.2% $\pm$ 6.5%. In both valence and arousal experiments, all CNN models explicitly and significantly outperform the baseline models ($p < 0.01$). Nevertheless, these chance levels could not be achieved in the subject-independent experiments.

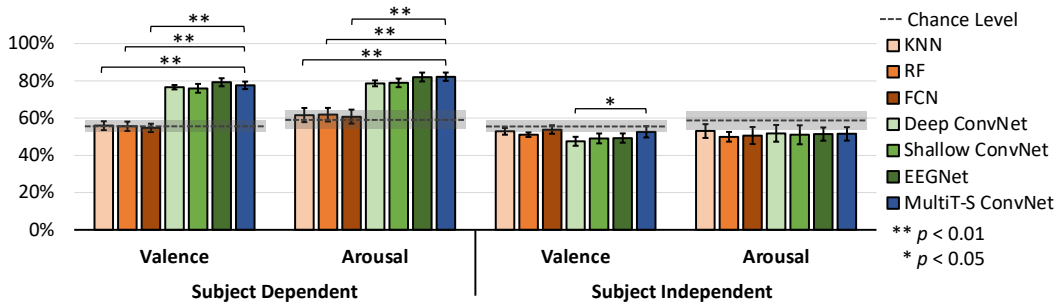

Figure 5.8: Average classification accuracy for all subjects on the DEAP dataset. Error bars denote a 95% CI computed from all subject's means. The horizontal dashed lines indicate the chance level, and the grey areas indicate the CI of chance levels. The stars indicate significant differences compared with MultiT-S ConvNet.

5.7 Ablation Study

To verify the effectiveness of temporal filtering, the studies further investigates the effect of temporal convolution on the other CNN models. Based on the first blocks of temporal convolution of each model in Figure 2.3, this ablation study reforms a temporal convolution with multi-kernel convolution. Meanwhile, the spatial filters were halved to preserve the number of parameters and computational cost. Here, this study consider a two-class classification with a subject-independent setting on the SEED dataset, and the results are shown in Table 5.1. In addition, a paired t-test is used to observe the difference between the corresponding models. Deep ConvNet with multi-kernel temporal convolution consists of 12×4 temporal (band) filters, 12 spatial filters, and 168 temporal-spatial filters. As a result, the number of trainable parameters decreases from 182,497 to 73,777. It significantly outperforms the original Deep ConvNet with 73.96% accuracy from 72.04% ($p < 0.05$). For EEGNet with multi-kernel temporal convolution, the number of trainable parameters increases from 13,537 to 16,177, but it achieves a significantly better accuracy of 74.00% than the original EEGNet of 70.15% ($p < 0.01$). However, there is no significant difference between Shallow ConvNet with and without multi-kernel convolution ($p > 0.05$). Moreover, this study trained the MultiT-S ConvNet model without SE blocks to observe their effects on the learning performance. The accuracy of proposed model without SE Blocks decreases from 75.13% to 73.95% ($p < 0.05$).

5.8 Discussion

In this study, the accuracy performance was examined using SEED and DEAP, which are well-known EEG emotion datasets. In addition, existing and proposed models were implemented with similar settings, including data segmentation, pre-processing, parameter tuning, and evaluation metrics, to ensure the reliability of model performance comparison. The experiments and results demonstrate that all ConvNets with appropriate design choices can outperform traditional approaches in terms of accuracy performances in all classification experiments. Moreover, this study compares MultiT-S ConvNet against existing ConvNet models in terms of accuracy performance in EEG-based emotion classification, as shown in Table 5.3. The comparison shows that the overall performance of MultiT-S ConvNet outperforms Deep ConvNet, Shallow ConvNet, and EEGNet with higher accuracies of 2.2%, 2.3%, and 3.7%, respectively.

According to the result, the SEED dataset is more uncomplicated to classify than the DEAP dataset because it contains different sources of emotional labels and the cleanliness of the signals. For the emotion states, the SEED dataset uses the movie types, whereas the DEAP dataset uses a subject questionnaire. This ambiguity directly affects the classification performance and trustworthiness of each dataset. Based on the SEED dataset results, all the obtained accuracy reached the chance level ($p < 0.01$) in both subject-dependent and independent experiments. For the DEAP dataset, only the ConvNet-based techniques reached the chance level ($p < 0.01$) in the subject-dependent experiments. In contrast, for the DEAP dataset, none of the models could surpass the chance level in subject-independent experiments. Many studies using the DEAP dataset have mainly conducted subject-dependent experiments because the data distribution of each subject in this dataset is highly diverse and varied. Observing the topoplots of PSD features in 4 frequency bands by human eyesight shows that the SEED dataset plots can be distinguished, whereas the DEAP dataset plots are very similar. In terms of subject-dependent and subject-independent learning, subject-dependent performance outperforms subject-independent performance significantly in all experiments. Consequently, the individual distribution directly affects learning performance. Moreover, the proposed model results are comparable and consistent with existing studies in terms of average accuracy [51].

5.8.1 Effect of Multi-Kernel Convolution

For several decades, ML has been applied in EEG analysis for individual modules combined with prior knowledge. Most previous studies applied manual parameterization for feature extraction and then transmitted them to an ML classifier [5, 36]. These traditional techniques require background knowledge for signal processing, noise reduction, and data manipulation. From the results in terms of the PSD feature, the classifiers are possibly influenced and affected by the noise and complexity of multi-channel EEG signals, resulting in poor performance. However, the ensemble model, RF, performs well overall and significantly outperforms all baseline techniques on the SEED dataset. The RF model potentially learns how to reduce the generalization error of the prediction and deal with various EEG classification tasks.

On the other hand, ConvNet models are a better choice in EEG analysis than manual handcrafted features. Numerous ConvNets with end-to-end architecture have been proposed to learn informative features and deal with the variation of noises in EEG analysis automatically and efficiently. For instance, attention classification using Shallow ConvNet and long short-term memory (LSTM) network on a three-back task [17], and emotion classification using subject-invariant bilateral variational domain adversarial neural network [22]. However, EEG features need to be extracted into various representations that improve the learning effect, especially for EEG analysis. Here, this study proposed the multi-kernel convolution that helps the model achieve better results than the single-kernel convolution. As a result, the module applies to and obtains better accuracy than existing ConvNets while preserving the model size and computation costs. In this study, all ConvNets models outperform the baseline techniques in all experiments, except in subject-independent experiments on the DEAP dataset. The number of parameters verifies that the learned features are useful for classification. It effectively estimates missing data and maintains good accuracy even when a large proportion of the data is missing. With significantly higher accuracy performance, MultiT-S ConvNet has approximately six times and three times fewer trainable parameters than Deep ConvNet and Shallow ConvNet, respectively. Despite having fewer parameters than the proposed model, EEGNet performance is unstable in terms of the average accuracy and standard deviation of the SEED and DEAP datasets. Moreover, the additional variations of features extracted by multi-kernel reduce over-

fitting or vanishing gradient problems while training a model.

Table 5.1: The number of filters, trainable parameters, and accuracies before and after applying multi-kernel convolution on the SEED dataset.

	No. of Temp. Filter	**No. of Spat. Filter**	**No. of Temp-Spat Filter**	**Trainable Parameters**	**Acc.(%)**
Deep ConvNet	24	24	48 + 96 + 192 = 336	182,497	72.04
Deep ConvNet (+)	$12 \times 4 = 48$	12	24 + 48 + 96 = 168	73,777	73.96*
Shallow ConvNet	40	40		101,281	73.96
Shallow ConvNet (+)	$20 \times 4 = 80$	20		101,721	73.19
EEGNet	32	64	64	13,537	70.15
EEGNet (+)	$16 \times 4 = 64$	32	32	16,177	74.00**
MultiT-S ConvNet	$24 \times 4 = 96$	64		30,313	75.13

Note: (+) indicates reforming of temporal filters and halving of spatial filters. Stars indicate paired t-tests compared to the corresponding model. (** is $p < 0.01$ and * is $p < 0.05$)

Table 5.2: The accuracy results of subject-dependent and subject-independent classification on SEED and DEAP datasets. The bottom row indicates the proposed MultiT-S ConvNet results. Bold indicates the highest accuracy.

Models	SEED - 2 Classes		SEED - 3 Classes		DEAP - Valence		DEAP - Arousal	
	Subj. Dep.	Subj. Indep.	Subj. Dep.	Subj. Indep.	Subj. Dep.	Subj. Indep.	Subj. Dep.	Subj. Indep.
Chance Level	50.0 ± 0.0	50.0 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	56.6 ± 9.2	**56.6 ± 9.2**	58.9 ± 15.5	**58.9 ± 15.5**
KNN	70.5 ± 12.3	72.0 ± 7.8	52.5 ± 12.7	43.7 ± 8.0	55.9 ± 6.9	52.9 ± 5.2	61.6 ± 11.0	53.0 ± 10.7
RF	87.0 ± 7.7	73.4 ± 8.5	71.7 ± 9.5	52.7 ± 6.9	55.6 ± 7.3	51.1 ± 3.6	61.8 ± 10.6	50.0 ± 7.5
FCN	77.4 ± 11.9	66.1 ± 10.1	56.5 ± 13.6	46.4 ± 8.4	54.8 ± 6.4	53.9 ± 6.6	60.8 ± 10.9	50.6 ± 13.2
Deep ConvNet	94.8 ± 3.4	72.0 ± 7.8	85.0 ± 4.5	51.1 ± 7.4	76.6 ± 3.4	47.5 ± 7.0	78.6 ± 4.6	51.8 ± 13.0
Shallow ConvNet	91.2 ± 7.4	74.0 ± 8.5	82.1 ± 7.7	54.4 ± 7.4	76.0 ± 6.9	49.1 ± 7.7	78.9 ± 6.7	51.0 ± 14.7
EEGNet	89.1 ± 3.6	70.1 ± 9.5	76.3 ± 15.1	47.8 ± 9.7	**79.2 ± 6.3**	49.3 ± 7.2	82.1 ± 6.9	51.4 ± 10.3
MultiT-S ConvNet	**95.2 ± 3.2**	**75.1 ± 8.4**	**86.0 ± 5.3**	**54.6 ± 6.8**	77.6 ± 5.8	52.6 ± 8.8	**82.2 ± 6.5**	51.5 ± 10.4

Table 5.3: The comparison of model performance between the proposed MultiT-S ConvNet and existing ConvNets. The accuracy (%) is the average from all experiments in this study. It calculates differences in the accuracy between MultiT-S ConvNet and other models, and the positive mark denotes a higher accuracy of the MultiT-S ConvNet. The minimum frequency (Hz) covered by each model is displayed. The bottom row computes the number of trainable parameters in dependence on the sampling rate of the SEED dataset. The negative and positive marks denote a decrease and increase of trainable parameters required by the MultiT-S ConvNet, compared with indicated models.

	MultiT-S ConvNet	**vs. Deep ConvNet**	**vs. Shallow ConvNet**	**vs. EEGNet**
Accuracy (%)	71.9	+2.2	+2.3	+3.7
Minimum freq. covered	5 Hz	40 Hz	17 Hz	2 Hz
Number of Trainable Parameters	30,313	−152,184	−43,464	+14,136

Chapter 6

Conclusion and Future Work

6.1 Summary

This research addressed a well-designed end-to-end convolution network in classification based on physiological signals, especially EEG signals, to increase feature extractability. First, this research aimed to compare physiological responses in feasible emotional affect recognition using sequence learning. Based on the results, only EEG features outperform the combination of modalities. Considering ECG features, the signals could not interpret in a short duration of input. While eye-tracking features could not denote the emotional affects, the subjects just watched the interesting points in each scene. In recognition techniques, sequence learning using the LSTM network outperforms the traditional techniques. The results could gain a continuous affect along with the duration and apply to comprehend the feedback and obtain an advantage from the interesting points ads. The results show that the output sequence corresponds to the user's feeling scores over time.

Second, this research presents the comparative study of wet and dry systems by performing two cognitive tasks and proposes the EEG-based classification model by constructing the convolution neural network with LSTM networks. The analysis shows the tendency relationship between dry and wet systems, not only in the statistical reports but also in classification results. Besides, the frontal brain activity occurs while performing the attention and music-emotion trials, similar to literature evidences. Overall results show that ConvNet with LSTM networks outperformed the baseline with PSD and ConvNet with FCNs on all datasets significantly. Unfortunately, our studies are limited to small imbalanced-dataset problems. It directly

caused reliability and the overfitting problem when classified by machine learning techniques. The experiments on the existing dataset verify that the proposed model performance relieves this problem.

Third, this study proposed multi-kernel filtering to increase the variation of temporal representations and further recalibrate both temporal and spatial features using the lightweight gating mechanism. The results show that our MultiT-S ConvNet outperforms the traditional and existing models. The number of parameters verifies that the learned features are helpful and informative for EEG-based emotion classification. ConvNets effectively estimates missing data and maintains good accuracy even when a large proportion of the data is missing. MultiT-S ConvNet has fewer trainable parameters than Deep ConvNet and Shallow ConvNet with significantly higher accuracy performance. Moreover, the multi-kernel reduced overfitting or vanishing gradient problems by extracting the variations of features. Ultimately, adding the proposed multi-scale module to existing models of EEG-based convolution networks improves the learning ability and classification performance while preserving computation costs. However, there are some limitations in the current study that could be addressed in future research to diversify the research problem and applications.

6.2 Future Research Directions

The findings in this research may lead to practical applications in EEG classification. In the future, the applicability of the MultiT-S network will be explored in brain disease recognition, such as epilepsy and Alzheimer's disease. The end-to-end model classifies the raw signal without pre-defined knowledge and expertise. Furthermore, the classification result potentially supports a specialist's diagnosis for more promising treatment from training large amounts of data. In addition, this proposed model could be further developed to allow detecting and analyzing dynamic changes in EEG signals over time. The model could perform a real-time interaction and outputs a continuous sequence along the task, such as game playing or anomaly detection during activity. Finally, the proposed module could be added to any model of EEG-based convolution networks, and its ability could improve the learning effect of other existing models when there is a limited dataset. Moreover, the multi-scale can be further developed by varying the kernel size

for other signals to extract various representations while preserving the computational costs. In the future BCI application, this research strongly supposes that a non-invasive electrode with a dry system combined with an end-to-end model may become a future brainwave translation tool.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] U. R. Acharya, F. Molinari, V. S. Subbhuraam, S. Chattopadhyay, N. Kh, and J. Suri. Automated diagnosis of epileptic EEG using entropies. *Biomedical Signal Processing and Control*, 7:401–408, 07 2012.

[3] A. N. Al-Nafjan, M. I. Hosny, Y. Al-Ohali, and A. Al-Wabil. Review and classification of emotion recognition based on EEG brain-computer interface system research: A systematic review. *Applied Sciences*, 7:1239, 2017.

[4] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, 6:39, 2012.

[5] S. Asadzadeh, T. Yousefi Rezaii, S. Beheshti, A. Delpak, and S. Meshgini. A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities. *Journal of Neuroscience Methods*, 339:108740, 2020.

[6] D. O. Bos. EEG-based emotion recognition. *The influence of visual and auditory stimuli*, 56(3):1–17, 2006.

[7] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su. Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In *Proceedings of 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7250–7253, 2015.

[8] W. Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, 39:106–124, 1927.

[9] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[10] H. Chennamma and X. Yuan. A survey on eye-gaze tracking techniques. *Indian Journal of Computer Science and Engineering*, 4, 2013.

[11] Y. M. Chi, Y. Wang, Y. Wang, C. Maier, T. Jung, and G. Cauwenberghs. Dry and non-contact EEG sensors for mobile brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(2):228–235, March 2012.

[12] F. Chollet et al. Keras, 2015. Available at: https://github.com/fchollet/keras.

[13] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134:9–12, 01 2004.

[14] N. Dewangan. A survey on ECG signal feature extraction and analysis techniques. *International Journal of Innovative Research in Electrical, Electronics, Instrumentatin and Control Engineering*, 3:12–19, 06 2017.

[15] M. Duvinage, T. Castermans, T. Dutoit, M. Petieau, T. Hoellinger, C. D. Saedeleer, K. Seetharaman, and G. Cheron. A p300-based quantitative comparison between the Emotiv Epoc headset and a medical EEG device. *Biomedical Engineering*, 765(1):2012–2764, 2012.

[16] T. Emsawas, K. Fukui, and M. Numao. Feasible affect recognition in advertising based on physiological responses from wearable sensors. In *Proceedings of Advances in Artificial Intelligence*, pages 27–36, Cham, 2020. Springer International Publishing.

[17] T. Emsawas, T. Kimura, K. Fukui, and M. Numao. Comparative study of wet and dry systems on EEG-based cognitive tasks. In *Proceedings of International Conference on Brain Informatics*, pages 309–318. Springer, 2020.

[18] V. C. R. Fortunato, J. d. M. E. Giraldi, and J. H. C. de Oliveira. A review of studies on neuromarketing: Practical results, techniques, contributions and limitations. *Journal of Management Research*, 6(2):201, 2014.

[19] A. S. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush. Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human factors*, 40 1:79–91, 1998.

[20] A. Graves. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer, 2012.

[21] C. Guger, G. Krausz, B. Allison, and G. Edlinger. Comparison of dry and gel based electrodes for p300 brain–computer interfaces. *Frontiers in Neuroscience*, 6:60, 2012.

[22] J. L. Hagad, T. Kimura, K. Fukui, and M. Numao. Learning subject-generalized topographical EEG embeddings using deep variational autoencoders and domain-adversarial regularization. *Sensors*, 21(5):1792, 2021.

[23] M. A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[24] L. R. Hochberg and J. P. Donoghue. Sensors for brain-computer interfaces. *IEEE Engineering in Medicine and Biology Magazine*, 25(5):32–38, 2006.

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 1997.

[26] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[27] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan. Asymmetric spatial pattern for EEG-based emotion detection. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2012.

[28] N. Jamil, A. N. Belkacem, S. Ouhbi, and A. Lakas. Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: A systematic literature review. *Sensors*, 21(14):4754, 2021.

[29] N. Jatupaiboon, S. Pan-ngum, and P. Israsena. Emotion classification using minimal EEG channels and frequency bands. In *Proceedings of 10th international joint conference on Computer Science and Software Engineering (JCSSE)*, pages 21–24, 2013.

[30] R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

[31] O. Jensen and C. Tesche. Frontal theta activity in humans increases with memory load in a working memory task. *The European journal of neuroscience*, 15:1395–9, 05 2002.

[32] S. Karpagachelvi, M. Arthanari, and M. Sivakumar. ECG feature extraction techniques - a survey approach. *International Journal of Computer Science and Information Security*, 8, 2010.

[33] M. Kim, M. Kim, E. Oh, and S.-P. Kim. A review on the computational methods for emotional state estimation from the human EEG. *Computational and mathematical methods in medicine*, 2013:573734, 2013.

[34] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[35] W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4):352, 1958.

[36] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2):169–195, 1999.

[37] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3:18–31, 12 2011.

[38] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[39] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[40] N. Lee, A. J. Broderick, and L. Chamberlain. What is 'neuromarketing'? a discussion and agenda for future research. *International Journal of Psychophysiology*, 63(2):199–204, 2007.

[41] M. Li and B.-L. Lu. Emotion classification based on gamma-band EEG. In *Proceedings of 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1223–1226, 2009.

[42] Y. Lin, C. Wang, T. Jung, T. Wu, S. Jeng, J. Duann, and J. Chen. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, July 2010.

[43] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu. Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors (Basel, Switzerland)*, 13:10273–86, 08 2013.

[44] J. d. R. Millan, R. Rupp, G. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb, C. Neuper, K.-R. Müller, and D. Mattia. Combining brain–computer interfaces and assistive technologies: State-of-the-art and challenges. *Frontiers in neuroscience*, 4:161–175, 09 2010.

[45] M. Numao. Emotion-driven creative music composition in brainmelody®. In *Proceedings of 10th International Conference on Computational Creativity (ICCC)*, pages 356–357, 2019.

[46] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis. Multilayer perceptron and neural networks. *World Scientific and Engineering Academy and Society (WSEAS) Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[47] T. S. Rached and A. Perkusich. Emotion recognition based on brain-computer interface systems. *Brain-computer interface systems-Recent progress and future prospects*, pages 253–270, 2013.

[48] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

[49] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[50] L. A. Schmidt and L. J. Trainor. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition & Emotion*, 15(4):487–500, 2001.

[51] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors (Switzerland)*, 18(7):2074, 2018.

[52] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing*, 12(4):857–869, 2019.

[53] A. K. Singh, Y.-K. Wang, J.-T. King, and C.-T. Lin. Extended interaction with a BCI video game changes resting-state brain activity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(4):809–823, 2020.

[54] M. W. Tangermann, M. Krauledat, K. Grzeska, M. Sagebaum, C. Vidaurre, B. Blankertz, and K.-R. Müller. Playing pinball with non-invasive BCI. In *Proceedings of the 21st*

International Conference on Neural Information Processing Systems, NIPS'08, page 1641–1648, Red Hook, NY, USA, 2008. Curran Associates Inc.

[55] N. Thammasan, K. Moriyama, K. Fukui, and M. Numao. Continuous music-emotion recognition based on electroencephalogram. *IEICE Transactions on Information and Systems*, E99.D(4):1234–1241, 2016.

[56] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo. EEG-based BCI emotion recognition: A survey. *Sensors*, 20(18):5083, 2020.

[57] M. van Vliet, A. Robben, N. Chumerin, N. V. Manyakov, A. Combaz, and M. M. Van Hulle. Designing a brain-computer interface controlled video-game using consumer grade EEG hardware. In *Proceedings of 2012 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–6, 2012.

[58] I. Volosyak, D. Valbuena, T. Malechka, J. Peuscher, and A. Gräser. Brain–computer interface using water-based electrodes. *Journal of Neural Engineering*, 7(6):066007, nov 2010.

[59] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

Appendix A

The visualization of SEED and DEAP datasets

A.1 The examples of topoplots extracted from PSD features

According to the complexity of SEED and DEAP datasets, these figures visualize the PSD features of theta, alpha, beta, and gamma bands. The topoplots of the SEED dataset are shown in Figure A.1 and A.2. The difference between the positive, neutral, and negative classes can be observed and related to the literature studies [50, 42, 31]. On the other hand, Figure A.3 and A.4 show the topoplots of the DEAP dataset. The differences between classes are complicated to distinguish, which caused the classification analysis hard.

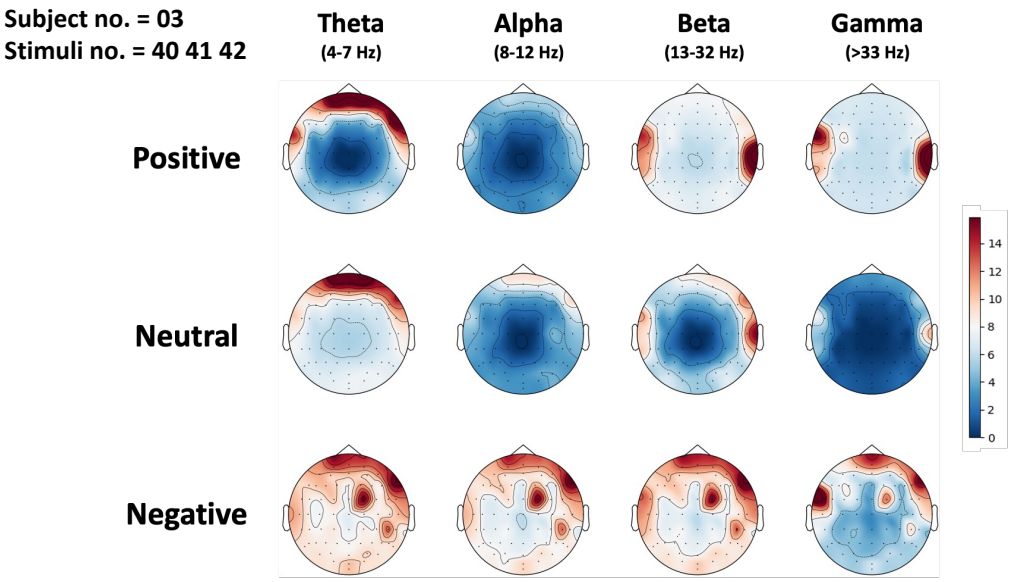

Figure A.1: Visualization of power spectrum features derived from subject no.3. Each of the 3 rows shows the positive neutral and negative.

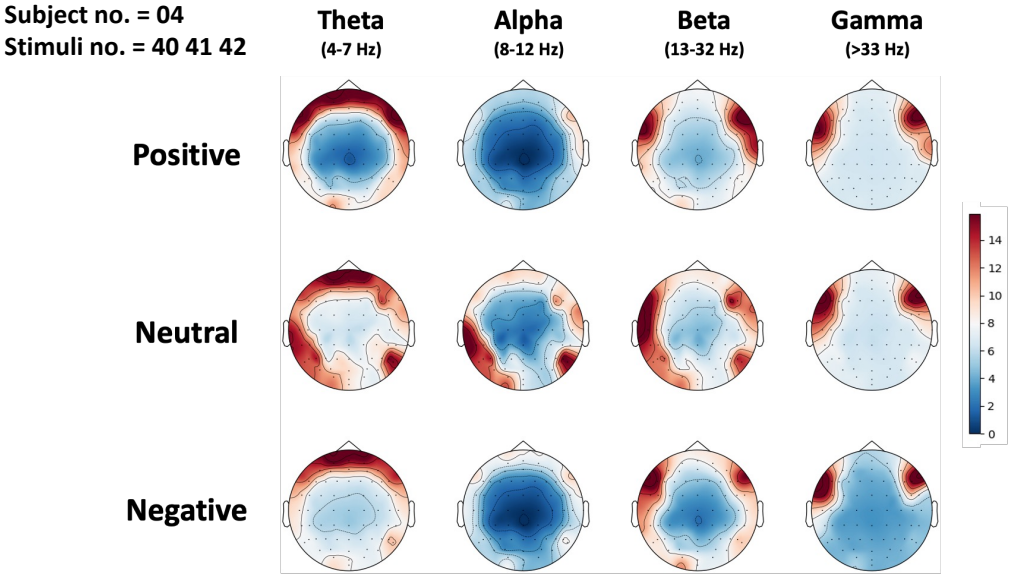

Figure A.2: Visualization of power spectrum features derived from subject no.4. Each of the 3 rows shows the positive neutral and negative.

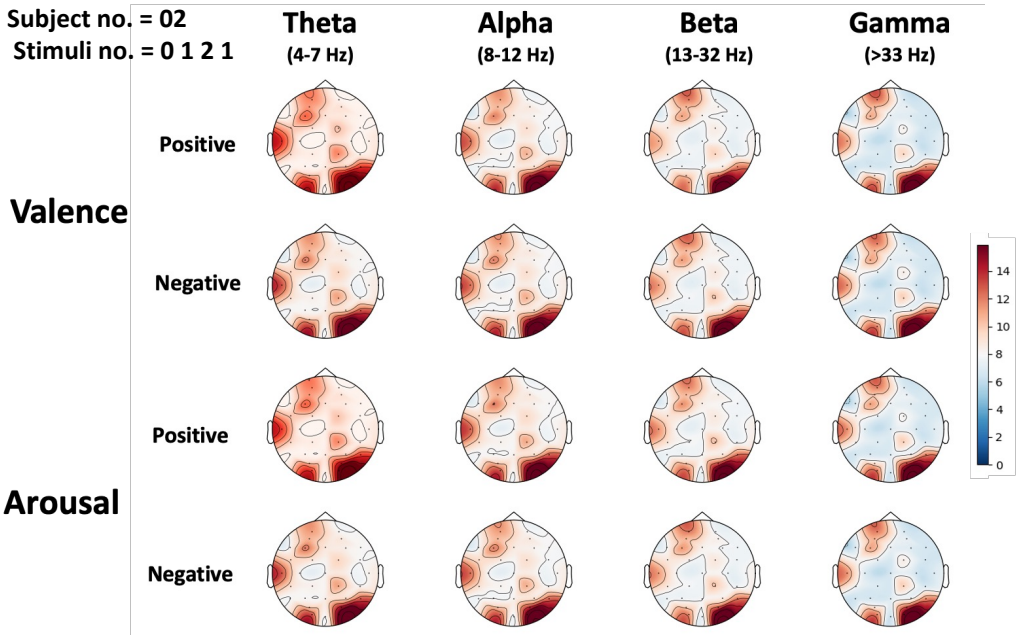

Figure A.3: Visualization of power spectrum features derived from subject no.2. First two rows show the positive and negative of valence and Next two rows show the positive and negative of valence arousal.

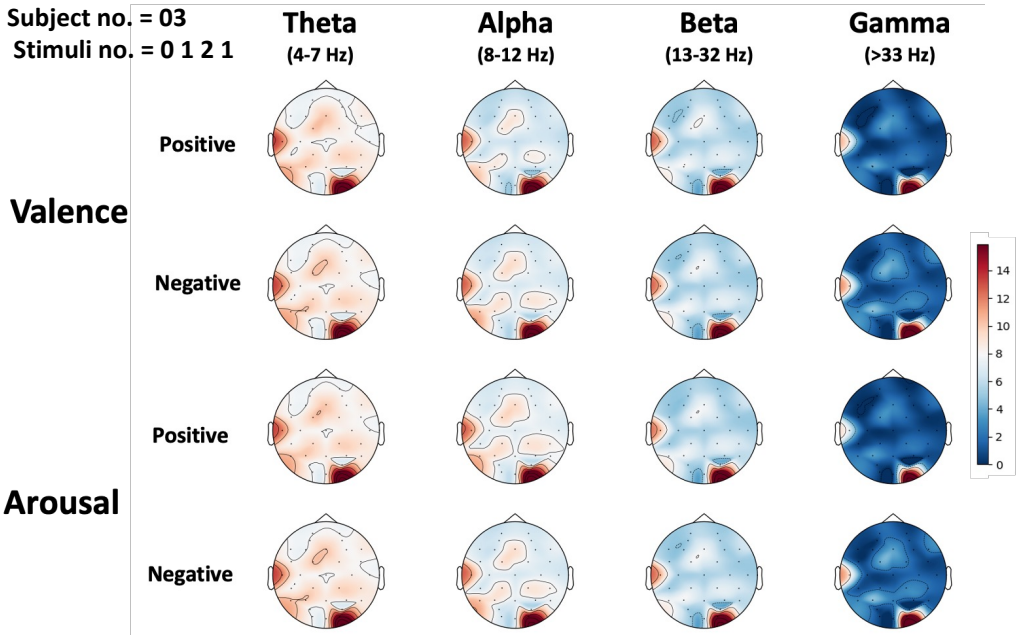

Figure A.4: Visualization of power spectrum features derived from subject no.3. First two rows show the positive and negative of valence and Next two rows show the positive and negative of valence arousal.