Deep Learning based Classification of Motor Imagery Electroencephalography Signals

By

JIAYANG ZHANG

A thesis submitted for the degree of
Doctor of Philosophy

School of Electronic & Electrical Engineering
Faculty of Engineering
UNIVERSITY OF LEEDS

APRIL 2024

ABSTRACT

Brain-Computer Interface (BCI) is a technology that enables direct communication between the brain and external devices. BCI systems often use Electroencephalography (EEG) to measure the electrical fields produced by brain activities, serving as a prominent brain mapping and neuroimaging technique utilized extensively within and beyond clinical settings. Motor imagery (MI), a prevalent BCI paradigm, enables individuals, particularly those with disabilities, to regulate brain signals voluntarily, bypassing the need for external stimuli. By decoding MI-EEG signals, the gap between motor intention and sensory feedback in motor movements disrupted by brain disorders is bridged, thereby facilitating swift motor functional recovery. However, the non-linear and nonstationary nature of MI-EEG signals poses a challenge to MI intention recognition, leaving room for potential classification enhancement. Moreover, factors such as subject variability, experimental conditions, and EEG recording devices impact the adaptability and robustness of models, consequently constraining the practicality of MI-EEG applications. The primary objective of this thesis is to investigate efficient deep learning models for decoding EEG signals and classifying MI tasks. The specific contributions of the thesis are outlined as follows:

1) A multi-view convolutional neural network (CNN) encoding approach for MI-EEG signals is proposed in Chapter 3. First, multiple frequency sub-band MI-EEG signals are created as the CNN model inputs through bandpass filters based on brain rhythms. Then, temporal and spatial features are captured based on the whole frequency band and the filtered sub-band signals, respectively. Further, utilizing two dense blocks with multi-CNN layers enhances model learning capabilities and strengthens information propagation. The proposed method achieves an average accuracy of 75.16% on the public Korea University EEG dataset which consists of 54 healthy subjects for the two-class motor imagery tasks.

2) Chapter 4 introduces a local and global convolutional transformer-based MI-EEG classification model. To make up for the shortcomings of the CNN model, a local transformer encoder is employed to dynamically extract temporal features. The global transformer encoder and densely connected network

are combined to improve the information flow and reuse. The spatial features from all channels and the difference in hemispheres are obtained to improve the robustness of the model. In the experiment, three scenarios including within-session, cross-session, and two-session are designed. Results show that the proposed model achieves up to 1.46%, 7.49%, and 7.46% accuracy improvement respectively in the three scenarios for the public Korean dataset compared with Tensor-CSPNet. For the BCI-IV-2a dataset, the proposed model also achieves a 2.12% and 2.21% improvement for the cross-session and two-session scenarios respectively.

3) Chapter 5 presents a cross-subject MI-EEG decoding method with domain generalization. In this study, the domain-invariant features from source subjects are extracted. The knowledge distillation framework is adopted to obtain the internally invariant representations based on spectral features fusion. Then the correlation alignment approach aligns the mutually invariant representations between each pair of sub-source domains. In addition, we use distance regularization on two kinds of invariant features to enhance generalizable information. The results demonstrate that the proposed model achieves 8.93% and 4.4% accuracy improvements on the public Korean dataset and BCI-IV-2a dataset respectively compared with the ConvNet and Dyanmic EEGInception model.

4) Chapter 6 proposes a graph convolutional network (GCN) based on transfer learning for cross-device MI-EEG decoding. Leveraging multi-channel information, the GCN module is employed to aggregate topological features. The pre-trained model is guided with few-channel signals as inputs through a knowledge distillation framework and adapted to the few-channel dataset using a transfer learning strategy with minimal data training. Experimental results show that the proposed model achieved an accuracy of 71.19% based on across-dataset, 7.04% higher than filter bank common spatial pattern (FBCSP) and EEG-ARNN model, demonstrating the effectiveness of our approach in cross-dataset MI-EEG decoding and enhancing the practicality of MI-BCI applications.

In summary, this study endeavors to decode MI-EEG signals using deep learning methods, improve the accuracy of motor intention recognition, and enhance the practicality of MI-based systems by enhancing model performance and robustness on cross-session, cross-subject, and cross-dataset scenarios.

NOMENCLATURE

Chapter 1. Abbreviations

BCI Brain-computer interface

CNN Convolutional neural network

EEG Electroencephalography

GCN Graph convolutional network

Chapter 2. Abbreviations

AMT Active Motor Training

AR Autoregressive

BN Bayesian network

CSP Common spatial pattern

CSSP Common spatiospectral pattern

CSSSP Common sparse spectral-spatial pattern

CWT Continuous Wavelet Transform

DFBCSP Discriminative filter bank Common spatial pattern

DFFS Dynamic frequency feature selection

DL Deep learning

EMD Empirical mode decomposition

ERD Event-related desynchronization

ERPs Event-related potentials

ERS Event-related synchronization

FBCSP Filter bank common spatial pattern

FFT Fast fourier transform

HHT Hilbert-Huang transform

ICA Independent component analysis

IMF Intrinsic mode functions

K Key

LDA linear discriminant analysis

LSTM Long short-term memory

NN Neural networks

PCA Principal component analysis

PSD Power Spectral Density

Q Query

RBF Radial Basis Function

RF random forest

RLDA Regularized linear discriminant analysis

RNN Recurrent neural network

SPD Symmetric positive definite

SSVEP Steady-state visually evoked potentials

STFT Short-time Fourier transform

SVM Support vector machine

V Value

VL Variance layer

VMD Variational mode decomposition

WT Wavelet decomposition

WT Wavelet transform

Chapter 3. Abbreviations

BCIC-IV-2a BCI Competition IV 2a

CV Cross-validation

DenseNet densely connected network

ELU Exponential linear unit

KU Korean University

RBF Radial bias function

Chapter 3. Parameters and Variables

C Channels

E EEG signal

$F_l(\cdot)$ Non-linear transformation

l The index of a layer

T Time Samples

X_i The $i-th$ EEG trial

x_l The output of each layer

Y_i The matching EEG label

y_p The prediction values

y_t The true labels

Chapter 4. Abbreviations

CV Computer Vision

NLP Natural Language Processing

PE Positional Encoding

SA Self-Attention

SBCSP Sub-band Common Spatial Pattern

SD Standard deviation

T-Densenet Block Transformer-based densenet block

t-SNE t-distributed Stochastic Neighbor Embedding

TSM tangent space mapping

w/o$_D iff-hemi$ Without the hemisphere difference

w/o$_T - dense$ Without T-dense units

w/o$_T rans$ Without transform encoders

Chapter 4. Parameters and Variables

μ The mean value

σ The standard deviation

d The dimension

d_k The dimension of keys

d_{model} The dimension of the outputs

hd_v The dimension of the values

pos The Position

x The raw data of each channel

Chapter 5. Abbreviations

CORAL Correlation alignment

DA Domain adaptation

DG Domain generalization

DICA Domain-invariant component analysis

GRL gradient reversal layer

KL Knowledge-leverage

MMD Maximum mean discrepancy

SCA Scatter component analysis

SMM Support matrix machine

TCA Transfer component analysis

TL Transfer learning

Chapter 5. Parameters and Variables

$\lambda_1, \lambda_2, \lambda_3$ The hyperparameters to limit the loss function

$\mathscr{L}_{div}$ The divergence

$\mathscr{L}_{mse}$ The Mean Squared Error

$\mathscr{L}_{cls}$ The cross-entropy loss

θ_S^c The parameters of feature classifier in the student network

θ_S^f The parameters of feature extractor in the student network

θ_T^c The parameters of feature classifier in the teacher network

θ_T^f The parameters of feature extractor in the teacher network

C Channels

C_i The covariance matrix

$d(.)$ The $L2$ distance

E The expectation

G_S^c The feature classifier in the student network

G_S^f The feature extractor in the student network

G_T^c The feature classifier in the teacher network

G_T^f The feature extractor in the teacher network

$h(n)$ The 3-order Butterworth filter

k_s, k_t The kernel size

N The number of subdommains

N_b The nubmer of sub-band

P^{tr}) The data distribution in the source domain

P_{XY} The joint distribution

S_{test} The test domain

S_{train} The source domain

T Timepoints

X The input space of EEG signals

X_{MB} The fused multi-band EEG data

Y The output space of EEG signals

$z1$ The internally-invariant features

$z2$ The mutually-invariant features

Chapter 6. Abbreviations

GCN Graph convolutional network

GNNs Graph neural networks

SENet Squeeze and Excitation Networks

Chapter 6. Parameters and Variables

$\widetilde{D}$ The degree matrix

$A_{initial}$ The initialized adjacency matrix

$A_{trainable}$ The mask matrix

E The edges

f_s The sampling rate

G The graph

k^i The kernel size

N The number of channels

P_{IJ} The Pearson's correlation coefficient

V The nodes

W The adjacency matrix

$x_student$ The feature maps of the student network

$x_teacher$ The feature maps of the teacher network

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my esteemed research supervisors, Prof. Kang Li and Prof. Shengquan Xie, for their patience, constructive feedback, and invaluable suggestions that have significantly broadened the scope of my research.

I am also profoundly thankful to my colleagues for their steadfast support, inspiration, and encouragement. Our collaborative efforts have enriched my research experience and broadened my perspectives. Their warmth and inclusivity have created a welcoming and supportive environment.

I would especially like to thank my family for their unconditional love during the challenging time of my Ph.D. journey. Without their unwavering support, I could not have achieved my goals.

Once again, I would like to extend my heartfelt thanks to all those who have contributed to my Ph.D. journey in one way or another. Their support, guidance, and encouragement have been invaluable and greatly appreciated.

DECLARATION

I declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Details of the publications which has been used in this thesis are as follows:

- In Chapter 3: Zhang J, Li K. A multi-view CNN encoding for motor imagery EEG signals[J]. Biomedical Signal Processing and Control, 2023, 85: 105063.
  As the lead author, the candidate contributed to the technical modelling and validation work as well as the paper drafting. Prof. Kang Li, as the co-author provided valuable professional advice and did proofreading.

- In Chapter 4: Zhang J, Li K, Yang B, et al. Local and global convolutional transformer-based motor imagery EEG classification[J]. Frontiers in Neuroscience, 2023, 17.
  As the lead author, the candidate contributed to the work including the modelling, algorithm development, validation as well as paper drafting. Prof. Kang Li provided valuable comments. Banghua Yang and Xiaofei Han provided the related model conceptions and ideas.

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Background

Electroencephalography (EEG) is a non-invasive technique used to record the brain's electrical activity. Detected by placing electrodes on the scalp, EEG signals are shown to reflect the macroscopic activity of the brain surface underlying various cognitive and motor functions [1], which makes it widely applicable across neuroscience and clinical medicine. In neuroscience, EEG is used to study brain function and dysfunction, including sleep patterns [2], seizures [3], and cognitive processes [4][5]. In clinical medicine, EEG is a valuable tool for diagnosing and monitoring neurological disorders such as stroke [6] and traumatic brain injury [7]. Additionally, EEG is increasingly used in brain-computer interface (BCI) systems.

The brain-computer interface (BCI) establishes a direct, and bidirectional, communication link between the brain and external devices without the need for muscular stimulation [8]. The EEG signals reflect patterns of brain activity, which are transmitted to external devices through BCI systems. By decoding EEG signals associated with specific mental tasks, such as imagining moving or focusing attention on a particular stimulus, people can control computer cursors, robotic arms, or other assistive devices

[9]. Motor imagery (MI), as one of the most important mental tasks in EEG experimental paradigms, refers to the mental simulation of body movements [10]. When a person is imagining or executing motor behavior, the related motor cortex on the brain scalp generates the corresponding MI responses with massive neuron activities [11]. Such a mechanism represents conscious access to the content of a movement, which is functionally analogous to unconscious motor planning. The framework of a MI-BCI system is shown in Fig 1.1. First, subjects will be guided to image motor movements. Huge amounts of MI-EEG signals are generated during the MI paradigm and collected for further analysis. Then, the MI-BCI system modeling for decoding MI-EEG signals and classifying MI-task by feature extraction and classification algorithms. The recognized task labels are then transferred to commands to control external devices, such as robotic arms and virtual reality (VR) equipment, which subsequently provide feedback to the subjects.

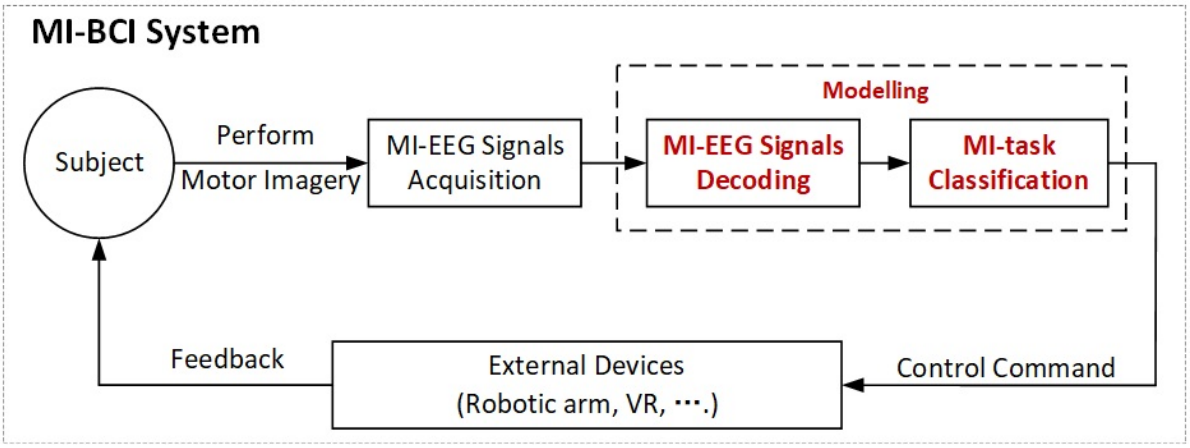


Figure 1.1: The framework of a MI-BCI system.

MI-BCI have a broad range of applications across various fields. In rehabilitation, MI-BCIs can help patients regain motor functions by enabling them to control external devices through mental imagery alone [12]. Beyond rehabilitation, MI-BCIs are also used in areas like gaming, where they allow players to interact with the game environment using their thoughts, creating more immersive experiences [13]. In the field of education, MI-BCIs offer potential for developing new learning tools that adapt to the mental

states of students, enhancing focus and engagement [14]. Additionally, MI-BCIs are being explored for their potential in communication for individuals with severe motor impairments, providing them with alternative means of expression [15]. These diverse applications highlight the versatility and potential of MI-BCI technology in improving lives and expanding human-computer interaction.

1.2 Challenges

MI-BCI has been proven to be a potential tool for rehabilitation, but its effectiveness and performance are limited by the capability of decoding EEG signals. The non-stationary, noisy, and non-linear nature of the MI-EEG signals makes informative feature extraction and classification highly difficult [16]. The specific challenges are as follows:

Low Classification Accuracy - The mainstream research content in MI-BCI involves binary classification tasks such as distinguishing between left and right-hand imagery movements, or multi-class classification tasks such as distinguishing between hands, feet, and tongue imagery movements. Achieving classification accuracy above 75% is often challenging in both binary and multi-class classification tasks. Additionally, the generalization performance of models is typically poor, influenced by factors such as individual variability, dataset characteristics, and preprocessing methods.

High Training and Calibration Cost - Significant individual differences often need to collect substantial data from each subject to build personalized models that achieve optimal classification performance. However, in real-world scenarios, it is challenging to gather large datasets, and there is limited data to train new models. Therefore, there is a growing need for plug-and-play functionality or the ability to train new models with minimal new data or without the need for extensive experiments. This poses a significant challenge to the generalization ability of the models, aiming to reduce the training and calibration cost.

Poor Generalization Ability and Practicality - The generalization ability and practicality of classification models are not only limited by individual differences among subjects but also significantly affected by the quality of data collected using different devices. The characteristics of MI-EEG signals are highly dependent on various individual differences. Even for the same individual and the same event, the MI-EEG may be different at different times [17]. Particularly in the experiments, no methodology has been employed to assess a subject's mental state, such as emotions and cognition. For instance, when an experiment lasts too long, the subject usually becomes less focused on the tasks because of fatigue. Signals collected during this period can be corrupted with more noise. Consequently, experiments conducted by the same individual at varying times may exhibit markedly disparate data quality. Moreover, the factors contributing to superior motor imagery performance in some subjects over others remain unclear. Variations among patients can be magnified by individual pathology. These challenges pose hurdles when applying the pre-trained model to new tasks or subjects. Different EEG-collected devices also need to be considered. For example, in laboratory settings, EEG data is often collected using wet electrodes with a resistance of about 10 kΩ, which requires extensive preparation time but results in better data quality. In practical applications, to reduce preparation time, semi-dry or dry electrodes are often used, but the high impedance of these electrodes can lead to poor data quality. Additionally, the varying number of electrodes across different devices also limits the feasibility of dataset transfer learning.

1.3 Research Motivations and Objectives

The decoding algorithm in MI-BCI rehabilitation strategies plays a crucial role. The accuracy and generalization ability of the model directly impacts the effectiveness and practicality of rehabilitation. Targeting the challenges raised in the last section, the following objectives will be addressed in this thesis:

- To improve the MI-EEG classification accuracy, feature extraction approaches need

to extract discriminative features from signals. Since the temporal, spatial, and spectral features are proven to be informative, effective and highly efficient model structures and processes need to be developed.

- To reduce the high training and calibration costs, the pre-trained model based on the existing historic dataset needs to be developed. Huge amounts of data in the existing dataset can help the model learn more feature representations. The techniques of domain generalization will be investigated to reduce the need for new data collection and new model training. To improve the model generalization ability, besides enhancing the training dataset, exploring invariant features across subjects that represent the common information and essence in MI-EEG signals should be considered.

- To enhance the model's practicality, it is important to develop models that can handle various scenarios, including cross-subject, cross-session, and cross-device challenges. Cross-subject scenarios involve using a model trained on one subject to recognize another subject's intentions. Cross-session scenarios address variations within the same subject across different experiment times. Cross-device scenarios deal with situations where practical devices may collect data of lower quality and with fewer electrodes, requiring the model to effectively utilize information transferred from high-quality, lab-collected data with more electrodes. These scenario-based MI-BCI systems face practical challenges, such as individual differences between subjects and variations in channels and data quality, which can lead to discrepancies in data distribution and impact the model's classification performance. Developing suitable models for different applications and ensuring high classification accuracy are key measures of the model's practical effectiveness.

1.4 Main Contributions

The thesis focuses on using deep learning methods to decode MI-EEG signals. To achieve the objectives above, different model structures are designed and validated on different

datasets and scenarios. The specific contents are outlined as follows:

- A Multi-View convolutional neural network (CNN) encoding approach for MI-EEG signals is proposed. First, multiple frequency sub-band MI-EEG signals are created as the CNN model inputs through bandpass filters based on brain rhythms. Then, temporal and spatial features are captured based on the whole frequency band and the filtered sub-band signals, respectively. Further, utilizing two dense blocks with multi-CNN layers enhances model learning capabilities and strengthens information propagation. The proposed method achieves an average accuracy of 75.16% on the public Korea University EEG dataset which consists of 54 healthy subjects for the two-class motor imagery tasks.

- A local and global convolutional Transformer-based MI-EEG classification model is proposed. To make up for the shortcomings of the CNN model, a local transformer encoder is employed to dynamically extract temporal features. The global transformer encoder and Densely Connected Network are combined to improve the information flow and reuse. The spatial features from all channels and the difference in hemispheres are obtained to improve the robustness of the model. In the experiment, three scenarios including within-session, cross-session, and two-session are designed. Results show that the proposed model achieves up to 1.46%, 7.49%, and 7.46% accuracy improvement respectively in the three scenarios for the public Korean dataset compared with current state-of-the-art models. For the BCI-IV-2a dataset, the proposed model also achieves a 2.12% and 2.21% improvement for the cross-session and two-session scenarios respectively.

- A cross-subject MI-EEG decoding method with domain generalization is proposed. In this study, the domain-invariant features from source subjects are extracted. The knowledge distillation framework is adopted to obtain the internally invariant representations based on spectral features fusion. Then the correlation alignment approach aligns the mutually invariant representations between each pair of sub-source domains. In addition, we use distance regularization on two kinds of

invariant features to enhance generalizable information. The results demonstrate that the proposed model achieves 8.93% and 4.4% accuracy improvements on the public Korean dataset and BCI-IV-2a dataset respectively compared with current state-of-the-art models.

- A graph convolutional network (GCN) based on transfer learning for cross-device MI-EEG Decoding. Leveraging multi-channel information, the GCN module is employed to aggregate topological features. The pre-trained model is guided with few-channel signals as inputs through a knowledge distillation framework and adapted to the few-channel dataset using a transfer learning strategy with minimal data training. Experimental results show up to 7.04% accuracy improvement compared to state-of-the-art models, demonstrating the effectiveness of our approach in cross-dataset MI-EEG decoding and enhancing the practicality of MI-BCI applications.

1.5 Thesis Structure

The thesis is organised in seven chapters and the relationship between each chapter is shown in Figure 1.2.

- Chapter 1 introduces the background of the MI-BCI and EEG signals. The main research purpose of this thesis is to use deep learning methods to decode MI-EEG signals for better rehabilitation.
- Chapter 2 reviews the literature related to MI-BCI for rehabilitation and the current classification methods for MI-EEG decoding. The following chapters include four main contributions of the study to address the issues mentioned in Chapter 1.
- Chapters 3 and 4 target the within-subject modelling based on deep learning methods. Chapter 3 divides MI-EEG signals into multi-subbands based on different brain rhythms. The CNN structure combines with the DenseNet Block to enhance the information flow which improves the classification accuracy. On this basis,

Chapter 4 adopts the Transformer structure with global and local schemes to make up the defects of CNN layers. Three practical scenario experiments including within-session, cross-session and two-session are conducted to validate the model's robustness and performance.

- Chapter 5 focuses on the cross-subject modelling. To realize the plug-and-play function, the model is trained to explore the invariant features across different subjects. Without retraining on new data, a pre-trained model can achieve excellent classification performance.

- Chapter 6 concentrates on the cross-device modelling. EEG signals collected in practical applications may have poor data quality with fewer electrodes due to the high impedance of the EEG-collected device. To harness the dataset in the lab with higher quality and more electrodes, a novel deep learning model is proposed to transfer useful MI information across different datasets or devices.

- Chapter 7 concludes the thesis. The main findings and contributions of the research conducted in the previous chapters are summarised. Reflections and suggestions for future work are given.

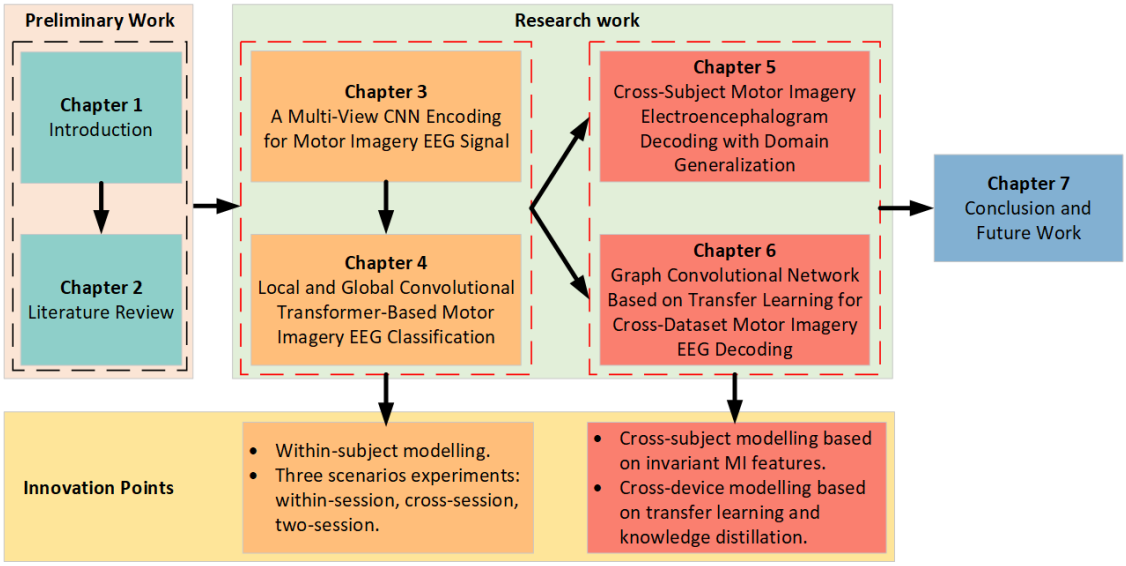


Figure 1.2: Thesis structure.

CHAPTER 2

LITERATURE REVIEW

2.1 An Overview of Motor Imagery Brain-Computer Interface

The MI-BCI system is gradually emerging as an effective tool and applied in multiple fields, particularly in rehabilitation. For instance, stroke is the third leading cause of adult disability worldwide [18]. Recovery of the motor function after a stroke is crucial to performing activities of daily living, but this recovery is often variable and incomplete [19]. Traditional rehabilitation strategies mainly include non-invasive stimulation and robotics assistance. Non-invasive procedures are elegant and powerful neuromodulatory techniques that create electric currents in the brain to change cortical excitability [20]. However, there is currently conflicting evidence regarding the efficacy due to unclear stimulation site, frequency, and intensity. Robotics provide intensity by increasing the number of repetitions that a therapist could impose. However, it is less effective for acute stroke patients due to passive mode and is limited since physical movements of stroke patients are often impossible [21]. In comparison, MI-BCI can potentially engage the same neural circuitry as actual movement, highlighting the prospects of BCI for

rehabilitation. Neurofeedback training and operant conditioning in MI-BCI get patients involved in the experiment which actively restores the nerves between the brain and the muscles [6].

In the medical application field, MI-BCI is intended for the replacement or restoration of central nervous system (CNS) functionality [22]. Elstob et al [23] propose a low-cost BCI prosthetic arm based on MI, which has 5 degrees of freedom of movement. An accuracy of between 56% and 100% depending on the movements were carried out, demonstrating that MI-based systems might be preferable to steady-state visually evoked potentials (SSVEP) systems because they are more intuitive and eliminate the fatigue caused by viewing flickering stimuli. Müller-Putz et al [24] also developed an MI-based robotic arm system utilizing a novel 64-electrode sleeve that can be worn by the user. This sleeve provides feedback on the movements performed through electrical pulses in a process called functional electrical stimulation (FES), which can offer feedback and assist in restoring certain aspects of CNS functionality in some patients.

Besides these biomedical applications, neurogames have also become increasingly more advanced by incorporating MI-BCI systems with other devices such as Virtual Reality (VR) and Augmented Reality (AR) environments [13]. Regular exposure to video games has improved their visual and spatial attention, memory, and mental rotation abilities over time [25] and sensorimotor learning, leading to better performance in tasks with consistent and predictable structures. Li et al [26] conducted the experiment based on a MI-BCI system in a 3D Tetris and an analogous 2D game-playing environment to enhance the player's BCI control ability. Certain contemporary technologies rely on evoked potentials, exemplified by the basic SSVEP-based implementation described in [27] and a system in [28] that integrates SSVEP and MI data to control a version of Tetris.

To sum up, MI-BCI plays an important role in various fields. For the entire BCI system, effective decoding of EEG signals is a crucial step, as it not only improves the accuracy of MI task classification but also advances the application of MI-BCI in real-world scenarios.

2.1.1 The Principle of EEG Decoding Algorithms

Event-related potentials (ERPs) are brain electrical activities recorded by EEG following specific stimuli or events, reflecting the neural activity response of the brain to particular stimuli or events [29]. ERPs arise from synchronous neuronal firing following specific events, and they manifest as distinct waveforms observable in EEG recordings. The event-related phenomena reflect frequency-specific alterations in ongoing EEG activity, generally characterized by decreases or increases in power within specific frequency bands. These changes are attributed to either a decrease or an increase in synchrony among underlying neuronal populations, respectively [30]. The former is termed event-related desynchronization (ERD), while the latter is known as event-related synchronization (ERS). The paradigm in MI-BCI is a classical event-related task. Taking the example of motor imagery tasks involving left and right-hand movements, ERD/ERS phenomena can be observed based on EEG decoding. Specifically, during the execution of the task, an increase in EEG energy is observed in the motor cortex corresponding to the same side as the limb involved in the task, while a decrease in EEG energy is observed in the motor cortex corresponding to the opposite side limb [31].

EEG signals contain information about various brain neural activities, which are typically divided into five frequency bands: δ (1-4Hz), θ (4-7Hz), α (8-13Hz), β (14-30Hz), and γ (31-45Hz). The δ band appears during deep sleep, extreme fatigue [32], drowsiness [33], or anesthesia [34]; the θ band is associated with hypnosis [35], and correlates with personality traits, anxiety, and working memory; the α band is present during eyes-closed resting states and represents normal physiological signals, remaining constant in the absence of external stimuli and disappearing momentarily upon exposure to external stimuli such as light; the β band appears during focused attention, tension, or excitement [36]; and the γ band is associated with cognitive processes [37]. As MI represents a normal physiological signal of the brain, it is typically decoded using the α and β bands in research [38]. With the advancement of EEG decoding techniques, the entire frequency spectrum of EEG signals is increasingly utilized in decoding.

2.1.2 The Preprocessing Approaches of EEG signals

Overall, the preprocessing methods for EEG signals can be categorized into three steps: channel selection, band-pass filtering, and artifact removal [39]. The choice of channels is closely related to the brain regions of interest and the conditions of the EEG acquisition equipment. For MI tasks, feature extraction primarily focuses on the channels related to motor areas or utilizes all brain channels to ensure sufficient spatial information is captured. Some studies also employ machine learning and deep learning approaches to refine channel selection. For instance, Tong et al [40] incorporated an efficient channel attention module into a neural network, allowing it to evaluate and assign weights to each channel based on their relative importance to BCI classification accuracy. Gaur et al [41] computed the correlation between EEG signals and selected highly correlated EEG channels for a specific subject without including classification accuracy by using the Pearson correlation coefficient. Jin et al [42] used the sum of logarithmic amplitudes and the first-order spectral moment features captured from bispectrum analysis to choose suitable EEG channels. In conclusion, the number of EEG channels can be reduced without significantly affecting accuracy, decreasing computational time and memory requirements.

In the step of signal frequency filtering, previous studies have explored the impact of different frequency bands on Mi classification results [43–45]. Avilov et al [44] analyzed the performance of deep learning models for MI classification using (4–38 Hz) and (0–38 Hz) frequency bands. In [45], the inclusion of the delta band (0–4 Hz) was proven to give better performance. According to [46] showed that the accuracy of MI classification using the neural network was higher using a wide frequency band (1–100 Hz) compared to 8–30 Hz. [43] showed that all three bands (0.5–4 Hz, 0.5–38 Hz, and 0.5–100 Hz gave better performance than the commonly used frequency range (4–38 Hz), and the best performance was achieved using the raw full-band without frequency filtering. In summary, for MI, the alpha and beta EEG rhythms have been proven effective, but there is no universally accepted standard for selecting other rhythms or frequency bands. Additionally, the optimal brain rhythms vary from person to person, sometimes

significantly impacting the final classification results.

Artifact removal approaches are also mainstream to preprocess the EEG signals. Among them, independent component analysis (ICA) [47] and common average reference (CAR) [48] are most common methods. ICA is a computational method that separates a multivariate signal into additive, independent components. When applied to EEG data, ICA can isolate and remove artifacts such as eye blinks, muscle activity, and electrical noise, thus enhancing the clarity of the neural signals [49, 50]. CAR, on the other hand, is a referencing technique that improves the signal-to-noise ratio by averaging the signals from all EEG electrodes and subtracting this average from each electrode's signal. This approach reduces common noise shared across channels, such as electrical interference, and highlights the differences between individual channels [51, 52]. However, the noise removal process is not always necessary, according to [39], the automatic and manual removal only occupied 24% among previous studies while 40% studies do not have any removal process.

2.2 Machine Learning-Based Decoding Algorithms for MI-EEG

2.2.1 Feature Extraction Approaches

Traditional machine learning-based processes for MI-EEG decoding can be divided into two steps: feature extraction and feature classification. In the initial research on EEG decoding, frequency-domain features, and temporal-domain features are two common types [53]. Band power features can be computed using diverse methods and are widely employed in BCIs that leverage oscillatory activity. Fast Fourier transform (FFT) is a classical frequency-domain method used to analyze EEG signals facilitating the transformation of signals from the time domain to the frequency domain for spectral analysis [54]. With FFT, features are extracted by utilizing mathematical tools to compute the Power Spectral Density (PSD) which converts the amplitude of EEG signal over time

into the spectrum of EEG signal power, thus visually observing the distribution and changes of EEG rhythm [55]. The spectral features of EEG signals can be captured by various methods such as Welch's periodogram [56], though its main contribution lies in visualization and interpretability. Besides FFT, other similar methods such as the Fourier decomposition method [57], the variational mode decomposition (VMD) method [58], and the Hilbert-Huang transform (HHT) method [59] have been developed for analyzing EEG signals.

Among the temporal-domain analysis approaches, independent component analysis (ICA) [60], principal component analysis (PCA) [61] and autoregressive (AR) [62] models are usually employed. ICA is a computational technique used to separate a multivariate signal into additive, statistically independent components. It assumes that the observed data is a linear combination of underlying independent components, each of which has a distinct statistical distribution. The goal of ICA is to find a set of basis vectors, known as independent components, such that the observed data can be represented as a linear combination of these components with maximally independent coefficients. PCA has a similar core idea used for dimensionality reduction and data compression. Both of them are more suitable for noise removal from EEG signals. AR assumes that real EEG signals can be predicted with the order and parameters of the approximation model. However, the effectiveness of analyzing nonlinear and non-stationary EEG signals based on AR models is not satisfactory, and they often require huge computational costs [63]. Besides that, combining time-domain and frequency-domain analysis is also one of the mainstream approaches. For instance, wavelet transform (WT) analyzes the features of EEG signals in the frequency domain while maintaining precise localization in the time domain [64]. This approach demonstrates strong performance in analyzing irregular and nonstationary signals across various window sizes [65]. The key advantages of the WT is its ability to offer precise frequency and time information for low and high frequencies, respectively.

Spatial features are also highly discriminative features within EEG signals. Common spatial pattern (CSP), as one of the most successful algorithms, can be used to extract

common spatial patterns underlying the EEGs from two MI tasks [66]. These spatial patterns explain the maximum variance in EEGs from one class and the minimum variance in the other, making them optimal for quantitatively discriminating between individual EEGs in the two tasks. [66]. The covariance of the mixed space can be expressed as:

$$C_c = C_l + C_r \qquad C = \frac{EE^T}{trace(EE^T)} \tag{2.1}$$

where C_l and C_r represent normalized covariance matrix from two kinds of signals. In MI signals, it usually represents the signal from two tasks (imagine moving right/left hand)

$$C_c = U_c \Lambda_c U_c^T \tag{2.2}$$

U_c represents eigenvector matrix and Λ_c represents eigenvalue diagonal matrix. Then whitening:

$$P = \Lambda_c^{-\frac{1}{2}} U_c^T \tag{2.3}$$

$$S_l = PC_lP^T \qquad S_r = PC_rP^T \tag{2.4}$$

$$S_l = B_l \Lambda_l {B_l}^T \qquad S_r = B_r \Lambda_r {B_r}^T \tag{2.5}$$

The sum of the eigenvalues of the two kinds of matrices is always one. The maximum eigenvalue of S_l corresponds to the minimum eigenvalue of S_r. The eigenvectors of the two matrices B_l and B_r are equivalent. The eigenvector corresponding to the maximum eigenvalue of S_l causes S_r to have the minimum eigenvalue and vice versa. The eigenvalues in Λ_l are arranged in descending order, then the corresponding eigenvalues in Λ_r are arranged in ascending order. m eigenvalues are selected for each of the maximum and minimum eigenvalues inΛ_r, and the corresponding eigenvectors are integrated as B. The space filter can be represented as:

$$W = (B^T P)^T \tag{2.6}$$

Multiply the raw EEG data $E_{\mathrm{n}*d}$ with the space filter, d represents the number of data and n represents the number of channels:

$$Z_{n*d} = W_{n*n} E_{\mathrm{n}*d} \tag{2.7}$$

the eigenvector f_p is computed as:

$$f_p = log\left(\frac{var\left(Z_p\right)}{\sum_{i=1}^{2m} var\left(Z_i\right)}\right) \qquad p = 1:2m \tag{2.8}$$

Many studies also developed other variants based on CSP. In [67], the common spatio-spectral pattern (CSSP) is proposed which adopts the technology of delay embedding to extend the CSP algorithm to the state space. Dornhege et al proposed a new approach namely the common sparse spectral-spatial pattern (CSSSP) that optimizes both the spatial filter and the spectral filter together to enhance the difference between multi-channel EEG signals [68]. In 2007, Novi et al decomposed the EEG signals into sub-bands using a filter bank and got a score from each sub-band after using CSP [69]. The final decision is based on the scores based on different sub-bands. Based on the fusion of different sub-band scores, Ang et al [70] proposed a method namely filter bank common spatial pattern (FBCSP) to choose the optimal features based on CSP by several band-pass filters with different band ranges. Higashi et al proposed a discriminative filter bank CSP (DFBCSP) that considers the combination of finite impulse response filters and spatial weights. This method optimized the corresponding weights by a function which can be regarded as another variation of the CSP algorithm. [71].

2.2.2 Feature Classification Approaches

Conventional machine learning methods like random forest (RF), linear discriminant analysis (LDA) [72], support vector machine (SVM) [73], and neural networks (NN) [74] are adopted as various classifiers for the MI-EEG decoding tasks.

RF is a versatile ensemble learning method widely used for classification tasks in machine learning. It operates by constructing a multitude of decision trees during training and outputting the mode of the classes of the individual trees. Each decision tree is trained on a random subset of the training data and a random subset of the features, ensuring diversity among the trees. During prediction, the output of multiple trees is aggregated to provide a more robust and accurate prediction compared to individual trees. In 2014, Bentlemsan et al used RF to combine bagging for bootstrap aggregation

and features that are selected randomly [75]. In research [76], Luo et al proposed a feature-selected approach namely the dynamic frequency feature selection (DFFS) with an RF classifier to decode MI-EEG signals. Nonetheless, the performance of Random Forest models is impacted by overfitting and instability, especially when dealing with trees of varying sizes.

LDA is a supervised learning algorithm utilized for dimensionality reduction and classification tasks. By maximizing the separation between classes while minimizing the variation within each class, LDA constructs linear combinations of features that effectively discriminate between classes. The calculation cost of the LDA classifier is low which is beneficial for use in BCI applications based on MI-EEG [77]. Chen et al used LDA to obtain the classification results using multiple frequency band signals and vote through probability summation [78]. Fu et al adopted regularized linear discriminant analysis (RLDA) to enhance the dimension of the diagonal elements of the scatter matrices to improve classification accuracy [79]. However, the noisy and non-linear nature of EEG signals makes it difficult for the LDA classifier to get excellent results.

SVM is another common classifier in the BCI field. It works by finding the hyperplane that best separates the classes in the input space. This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors. Islam et al used SVM on features with reduced dimensions obtained by adopting multi-band PCA for four-class classification problems [80]. Although SVM produces better classification results, it cannot deal with the multiclass problem and decode complex EEG signals effectively [77].

The NN has been extensively used in the BCI field for providing a reasonable balance between accuracy and training speed. Sagee et al used a Bayesian network (BN) for maximum probable channel selection and NN for feature classification [81]. Hamedi et al employed Radial Basis Function (RBF) neural networks to reduce training time while ensuring high accuracy [82].

2.3 Deep Learning-Based Decoding Algorithms for MI-EEG

Deep learning (DL) has emerged as a prevalent methodology in machine learning in recent years, leading to significant breakthroughs in computer vision and speech recognition [83]. The learning capacity of deep neural networks stems in part from their ability to discover intricate feature representations from raw data. This has inspired a growing interest among neuro-engineering researchers to apply deep learning to the development of BCI systems because it largely alleviates the need for manual feature extraction as seen in conventional BCI, which requires domain-specific expertise in the signal [84]. Deep learning is an end-to-end process, researchers can focus on the input formulation, structure, and parameter optimization factors. However, there are no clear advantages or disadvantages to the influence of these factors. Despite more and more examples have shown impressive progress based on deep learning, there is still room for considerable improvement with respect to several important aspects of information extraction from the EEG, including its accuracy, interpretability, and usability for offline or online applications [85].

2.3.1 Input Formulation

EEG signals inherently exhibit noise and are susceptible to channel crosstalk. In conventional scalp EEG recording setups, each electrode captures signals from its surrounding area, resulting in coarse spatial resolution (typically several centimeters) [86]. The decomposition of these signals is complex due to the conduction properties of human brain tissues, skull, scalp, and hair. Consequently, a prominent challenge in EEG data analysis lies in formulating suitable inputs. As outlined in the survey by Craik et al. [86], many neural networks, particularly CNNs, utilize various inputs such as images generated from EEG data, raw signals, and computed features.

2.3.1.1 Images Input

The CNN's unprecedented ability to learn images encouraged researchers to transform raw EEG signals into images as input to the classifier. Among them, PSD, wavelet decomposition (WD), and short-time Fourier transform (STFT) are the three most common methods used in the reviewed studies. For instance, Xu et al used WT to convert multichannel EEG signals into two-dimensional time-frequency images, to obtain its comprehensive information, including both the time-frequency features and the relative position of the electrodes [51]. Li et al used Continuous Wavelet Transform (CWT) to map MI-EEG signals into two-dimensional image signals and extract the mu and beta rhythms from these image signals [87]. The expression of the continuous WT is given

$$W_s(\alpha,\tau) = \frac{1}{\sqrt{\alpha}} \int s(t)\phi^* \left(\frac{t-\tau}{\alpha}\right) dt \tag{2.9}$$

where s(t) is the input signal, α is the scale of WT, ϕ is the wavelet basis function, and$\tau$is the time shift. The wavelet function is the Morlet wavelet. Its expression is as follows:

$$\phi(t) = \left(\frac{2}{\pi T^2}\right)^{\frac{1}{4}} \exp\left(-\frac{t^2}{T^2} + jw_c t\right) \tag{2.10}$$

The expression of frequency is:

$$\Phi(w) = \left(\frac{T^2}{2\pi}\right)^{\frac{1}{4}} \exp\left(-\frac{(w-w_c)^2}{4T^2}\right) \tag{2.11}$$

where $\phi(t)$ is the time domain expression after CWT and $\Phi(t)$ is the frequency domain expression after CWT. Kant et al used a similar method to deal with EEG signals [88]. In the experiment, the wavelet used for CWT is the analytic Morse wavelet as it has better time-frequency localization. For Morse wavelet symmetry parameter (gamma) and time-bandwidth product were kept at 3 and 60 respectively. Data from electrodes C3 and C4 are stacked together, C4 after C3 to represent all the data into a single representation of one event (Left or Right) hand imagery.

Besides using WT or the variant from WT, STFT is also an easy way to get time-frequency images. Dai et al extracts EEG signals with a length of 2s from each MI EEG recording [89]. The STFT was conducted with time lapses = 14 and window size

= 64. Among all 500 samples, the STFT was computed for 32 windows on the first 498 samples. Therefore, a 257 × 32 image is produced, where the numbers 32 and 257 represent the samples on the axes of time and frequency. Subsequently, the beta and mu frequency bands were extracted from the spectrum of the output. Frequency bands of 6 ~ 13 and 17 ~ 30 are taken as the mu and beta bands, respectively. Fig 2.1 shows the transformed images using different mother wavelets. Table 2.1 shows the classification results using different channels and mother wavelets [51]. Besides directly using images

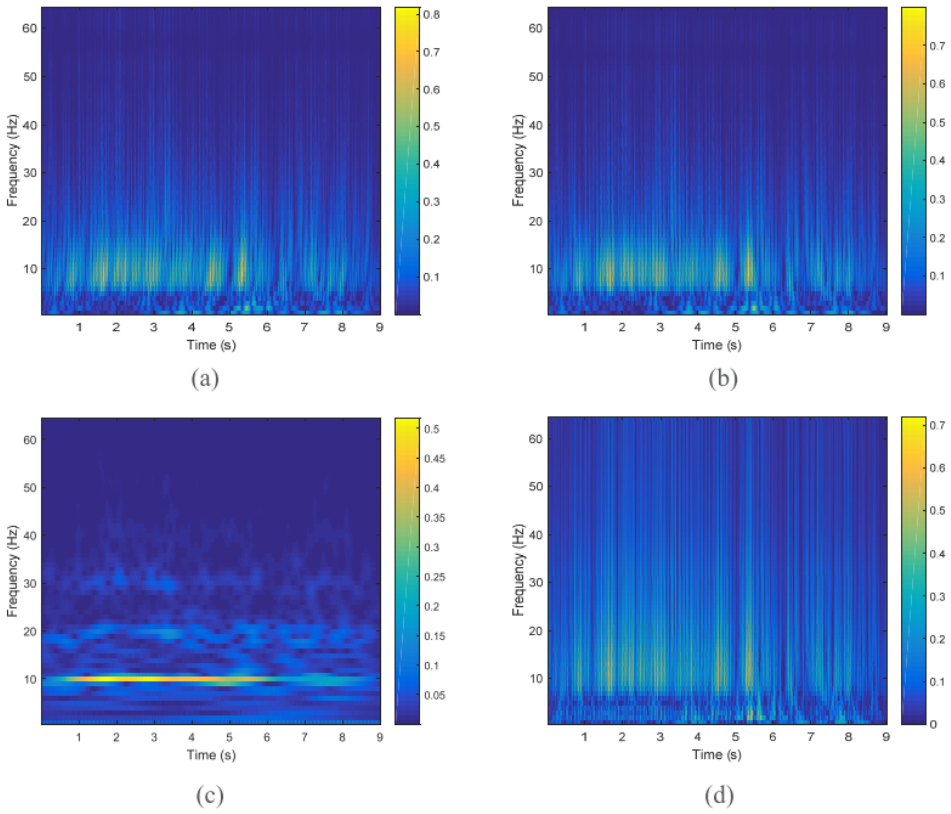


Figure 2.1: Transformed images using mother wavelet: db4(a), sym(b), cmor3-3(c) and haar(d).

transferred from WT or CWT, some studies also combine several images as one input. Piyush et al [88] used the analytic Morse wavelet as the wavelet for CWT to have better time-frequency localization. For Morse wavelet symmetry parameter (gamma) and time-bandwidth product were kept at 3 and 60 respectively. Voices per octave were kept at 10. Data from electrodes C3 and C4 are stacked together, C4 after C3 to represent all

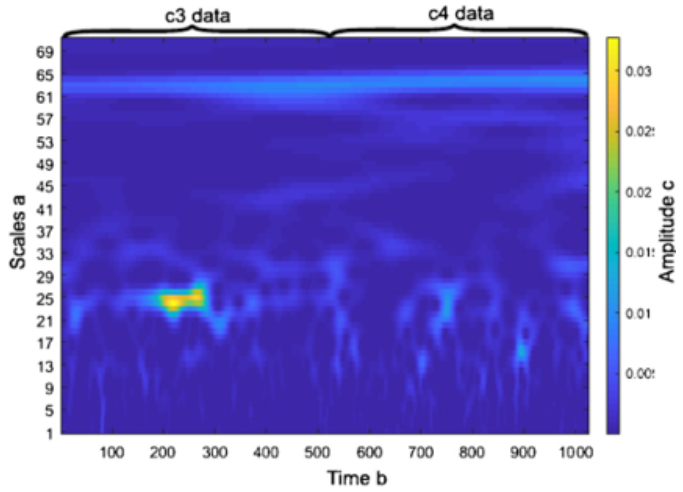
Table 2.1: Classification results using different channels and mother wavelets.

Wavelet name	Accuracy(%)					
	2 channels			3 channels		
	Worst	Best	Mean	Worst	Best	Mean
db4	77.63	81.25	79.25	71.43	80.6	75.6
sym4	80.14	85.5	81.398	71.43	82.14	73.21
cmor3-3	87.5	92.75	89.56	78.25	83.5	82.37
haar	68.25	72.5	70.31	64.3	69.6	67.13

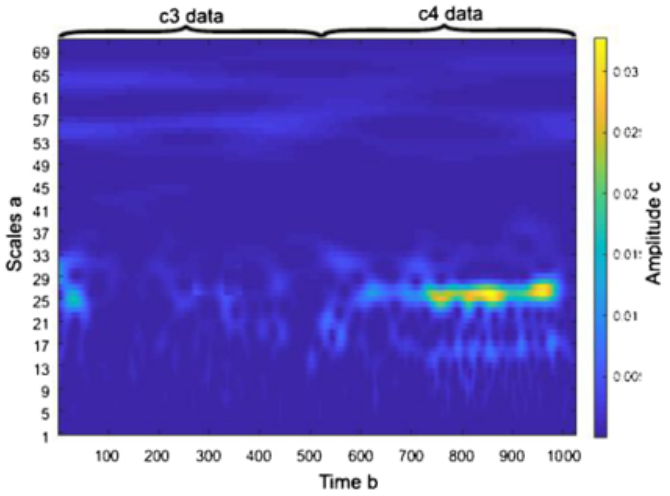
the data into a single representation of one event (Left or Right) hand imagery as shown in Fig 2.2 (a) and (b) [88]. Processed data is further used to train the model using deep neural networks. STFT is suitable for processing linear nonstationary signals, while WT and CWT can process nonlinear nonstationary signals in theory, but they can only process linear nonstationary signals in the actual algorithm implementation. Therefore, Huang et al used Hilbert–Huang transform (HHT) to transform EEG signals into time-frequency representation [90]. Fig 2.3 shows the decomposed EEG signals. HHT [91] is an adaptive signal processing method that is suitable for processing nonlinear and nonstationary signals. It mainly consists of two parts: The first part is empirical mode decomposition (EMD); the second is Hilbert transform (HT). In the first part, EMD adaptively decomposes any complex signal into a series of intrinsic mode functions (IMFs) according to the signal characteristics. This satisfies the two following conditions: (1) the average value of the mean value tends to be 0, and (2) the difference between the number of extreme points of the original signal (including the number of maximum points + the number of minimum points) and the number of intersections of the original signal cannot be greater than 1 (less than or equal to 1). For the original signal $x(t)$ EMD can be used to decompose it into

$$x(t) = \sum_{i=1}^{K} IMF_{(i)}(t) + r_K(t) \tag{2.12}$$

where $x(t)$ is the original signal, and $IMF(i)$ is K intrinsic mode functions $r_{(K)}$ is the negligible residue of the signal, which is the remainder of the subtraction of the original signal and $IMF(i)$. The decomposition process of EMD is divided into four steps. Step 1: calculate the mean value $m(t)$ according to the envelope line of the original signal $x(t)$.



(a) Left hand



(b) Right Hand

Figure 2.2: Scalogram representation for both hand motor imagery.

Meanwhile, $h(t)$ can be obtained:

$$h(t) = x(t) - m(t) \tag{2.13}$$

Step 2: Judge whether the $h(t)$ meets the two conditions of the IMF. If not, take $h(t)$ as the input signal and go back to step 1. If the conditions are met, get an IMF and go to the next step. Step 3: set the kth IMF as $hk(t)$, assign it to $ck(t)$, obtained as follows:

$$c_k(t) = h_k(t) \tag{2.14}$$

$ck(t)$ is separated from the original sequence and a new residual term is obtained:

$$r_k(t) = x_k(t) - c_k(t) \tag{2.15}$$

Step 4: judge whether the new remaining item meets the end condition of EMD, if not, bring the remaining item back to step 1; if it meets, end the EMD. After decomposition, the original signal can be expressed in the form of (nIMF+1 residual item):

$$x(t) = \sum_{i=1}^{n} c_i + r_n \tag{2.16}$$

Finally, HT is used to calculate the instantaneous frequency and amplitude to transform signals into Hilbert spectrum as inputs sent into deep learning structure.

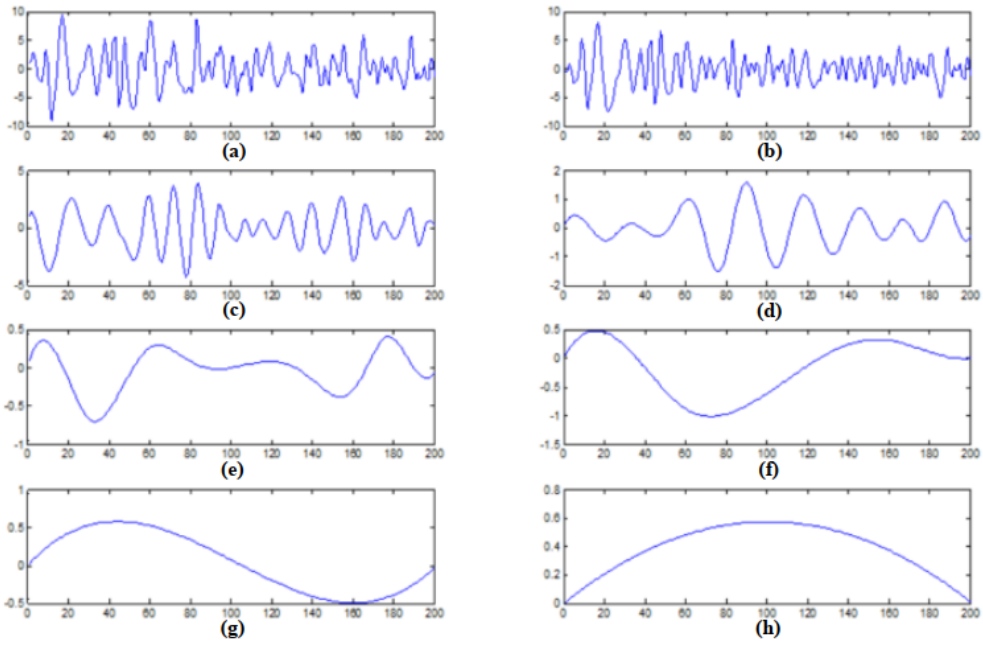


Figure 2.3: Schematic diagram of processed EEG signal, (a) is the original signal of EEG, and then (b) ~ (h) is IMF.

2.3.1.2 Signal Values Input

Signal values are also used directly as inputs to the neural networks. In contrast to two-dimensional static images, the EEG signal is a dynamic time series from electrode measurements obtained on the three-dimensional scalp surface. Also, the EEG signal

has a comparatively low signal-to-noise ratio, that is, sources that have no task-relevant information often affect the EEG signal more strongly than the task-relevant sources. These properties could make learning features in an end-to-end fashion fundamentally more difficult for EEG signals than for images. Thus, the existing structures from the field of computer vision need to be adapted for EEG input and the resulting decoding accuracies rigorously evaluated against more traditional feature extraction methods. For that purpose, in 2017, Robin et al created three ConvNets with different structures, with the number of convolutional layers ranging from 2 layers in a "shallow" ConvNet over a 5-layer deep ConvNet up to a 31-layer residual network (ResNet) [85]. Among them, the "Shallow" and "Deep" ConvNet structures have a great influence on the following new models. The most two important steps proposed by the team are that one dimension filter is first used to get the time feature from each channel and then uses depthwise convolution to get the space feature from signals. The detail is shown in Fig 2.4. These

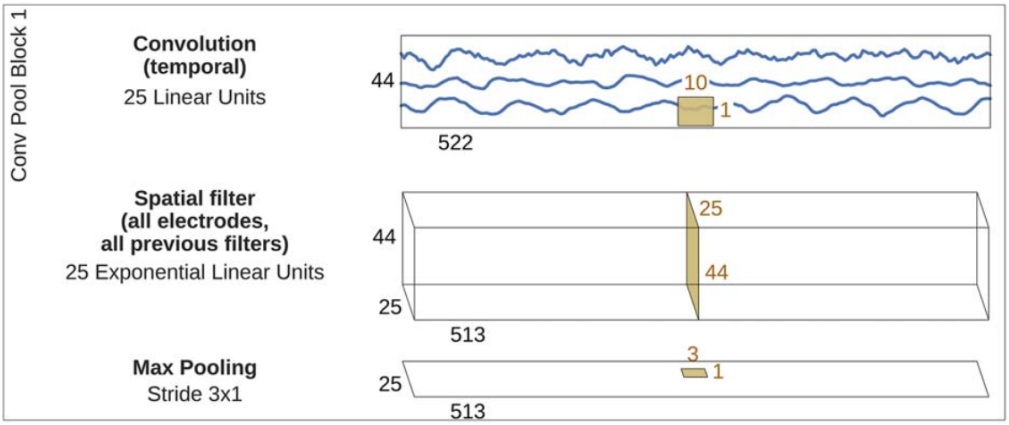


Figure 2.4: The first block structure in "Shallow" and "Deep" ConvNet.

two models are tested in the BCI Competition IV 2a&2b datasets. The accuracy of the four categories can reach around 70%, surpassing the champion method filter bank common spatial patterns (FBCSP) [70] by 2% – 5%. This is a huge improvement based on deep learning and the team also uses visualization to add more interpretability to deep learning.

However, the number of trainable parameters per model is around 152219 which is time-consuming to compute. To reduce parameters, Lawhern in 2018 proposed a new

structure called “EEGnet” [92]. Fig 2.5 shows the details of its structure. The network

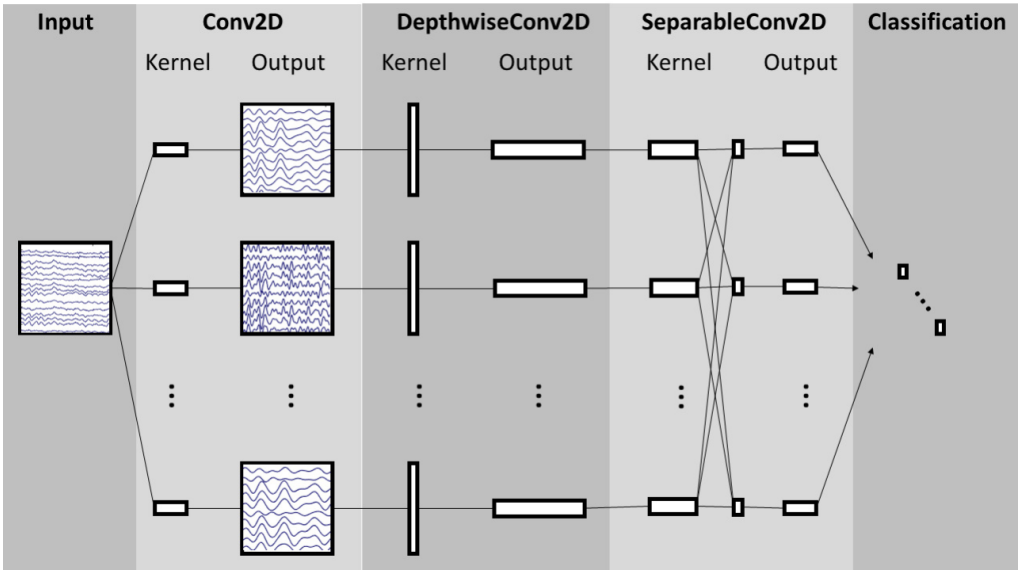


Figure 2.5: The structure of EEGNet.

starts with a temporal convolution to learn frequency filters and then uses a depthwise convolution, connected to each feature map individually, to learn frequency-specific spatial filters. The separable convolution is a combination of a depthwise convolution, which learns a temporal summary for each feature map individually, followed by a pointwise convolution, which learns how to optimally mix the feature maps. The number of parameters in the EEGnet is only 796, less than the one used in Deep ConvNet. Table 2.2 is the result of parameters used in different deep learning models and datasets. Meanwhile, Deep ConvNet uses the cropping method to generate more data but EEGnet does not use any data augmentation methods. It reduces much time to train and test while only sacrificing a small accuracy rate. Fig 2.6 shows the accuracy result compared with deep and shallow ConvNet. The team also tried the model on other BCI paradigms such as P300 besides MI and got success.

Besides these three classical models’ structures, researchers have made other changes to these models. For instance, Syed Umar Amin [92] proposed a multi-layer CNNs method for fusing CNNs with different characteristics and structures to improve EEG MI classification accuracy. There are four CNN structures with different depths, and then

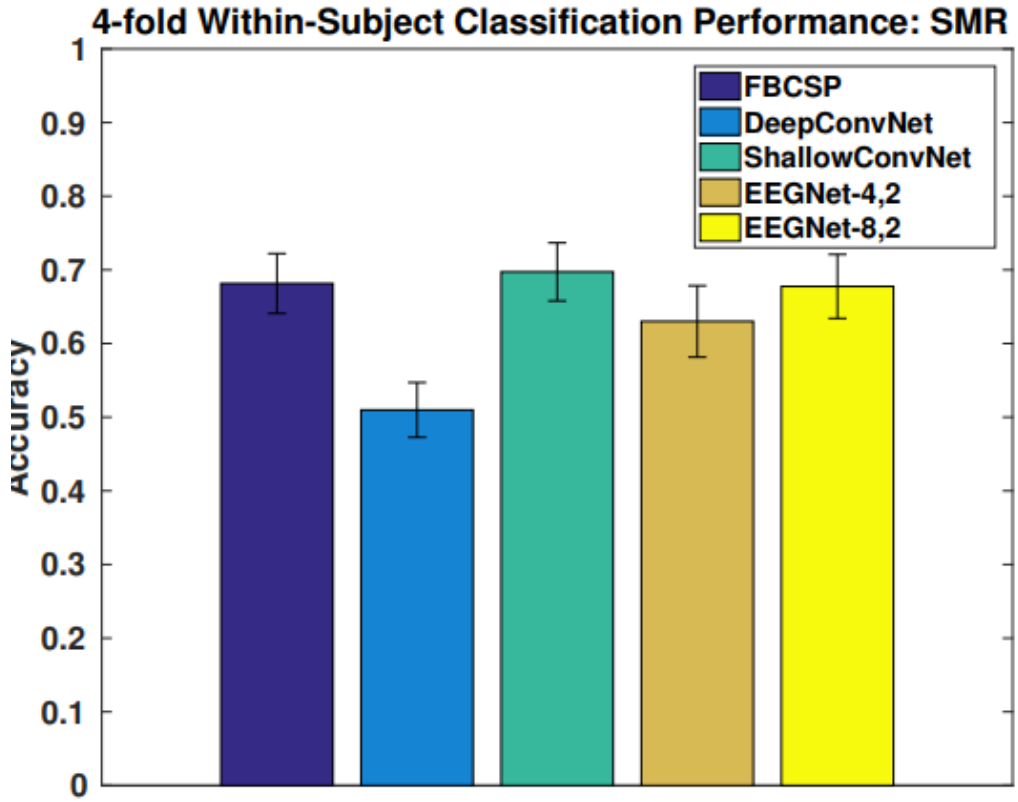


Figure 2.6: 4-fold within-subject classification performance for the SMR dataset for each model, averaged over all folds and all subjects.

Table 2.2: Number of trainable parameters per model and per dataset for all CNN-based models.

	Trial Length(sec)	DeepConvNet	ShallowConvNet	**EEGNet-4,2**	EEGNet-8,2
P300	1	174,127	1,004,002	**1,006**	2,258
ERN	1.25	169,927	91,602	**1,082**	2,290
MRCP	1.5	175,725	104,722	**1,098**	2,322
SMR	2	152,219	40,644	**796**	1,716

a connected layer combines the outputs of each CNN structure. More parameters need to be trained, but the accuracy rises from 72% to 75.7% in dataset BCI Competition IV 2a. In 2020, Mane proposed the FBCnet [93] which uses a Variance layer (VL) that computes the temporal variance of the individual time series. Following the VL features from all parallel branches are concatenated and fed to a FC layer with linear activation. The output of the FC is then passed through the softmax layer to get the output probabilities of each class. The binary classification accuracy on Korean Dataset [94] can reach 74%,

9% higher than EEGnet and Shallow CovNet. Furthermore, many studies show the advantages of using raw signals as input.

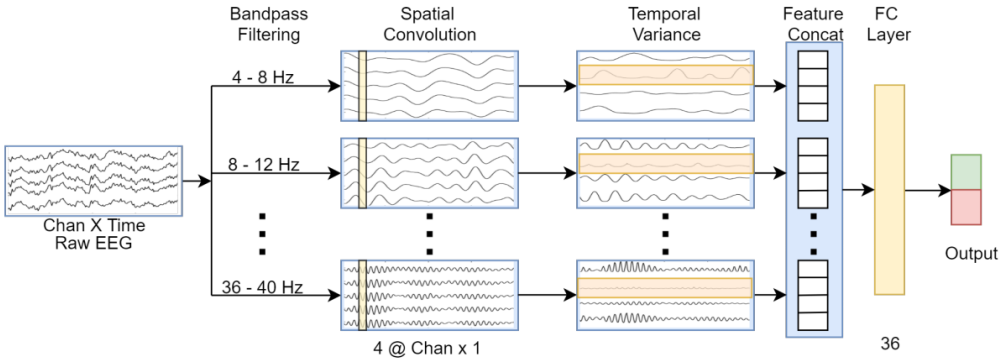


Figure 2.7: The structure of FBCNet.

2.3.1.3 Calculated Features Input

Compared with the other two types of input, using calculated features as input based on deep learning is not popular. The main reason is that deep learning is an end-to-end process. As a black box, researchers do not know if the traditional features can be learned or given useful information to a deep learning model. Therefore, more studies prefer to let the DL models learn everything by themselves. However, some studies still use traditional artificial features and have achieved good results. For instance, Ce et al [95] used covariance matrices of EEG signals as symmetric positive definite (SPD) matrices and sent them into the DL models. SPD is compatible with calculations in Riemannian space which can better handle high-dimensional data and capture the nonlinear structures and complex features. Ma et al [52] calculated the Pearson correlation coefficient and captured coherence frequency band features. Calculated features often contain prior knowledge, which can enhance the interpretability of the model and reveal potential neural mechanisms.

2.3.2 Model Structure

A crucial choice in the DL-based EEG processing pipeline is the neural network architecture. In the systematic review [96], until the end of 2014, DBNs and FC networks comprised the majority of the studies. However, since 2015, CNNs have been the architecture type of choice in most studies. This can be attributed to its capabilities of end-to-end learning and of exploiting hierarchical structure on the data, as well as their success and subsequent popularity on computer vision tasks, such as the ILSVRC 2012 challenge [97]. The proportion of studies using CNNs and combinations of recurrent and convolutional layers has been growing steadily.

2.3.2.1 Convolutional Neural Networks

In deep learning, a CNN is a class of artificial neural networks, most commonly applied to analyze visual imagery. CNNs are regularized versions of multilayer perceptrons. Multilayer perceptron usually means fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks makes them prone to overfitting data. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters.

A CNN consists of an input layer, hidden layers, and an output layer. In a feed-forward neural network, the middle layers are called hidden because their inputs and outputs are masked by the activation function and final convolution. The hidden layers include layers that perform convolutions. As the convolution kernel slides along the input matrix for the layer, the convolution operation generates a feature map, which in turn contributes to the input of the next layer. This is followed by other layers such as pooling layers, fully connected layers, and normalization layers. Fig 2.8 shows the process of the feed-forward in CNN structure.

Compared to other deep learning models, CNN is more flexible, especially in dealing with EEG signals [86]. When the EEG signals are transformed into images, researchers

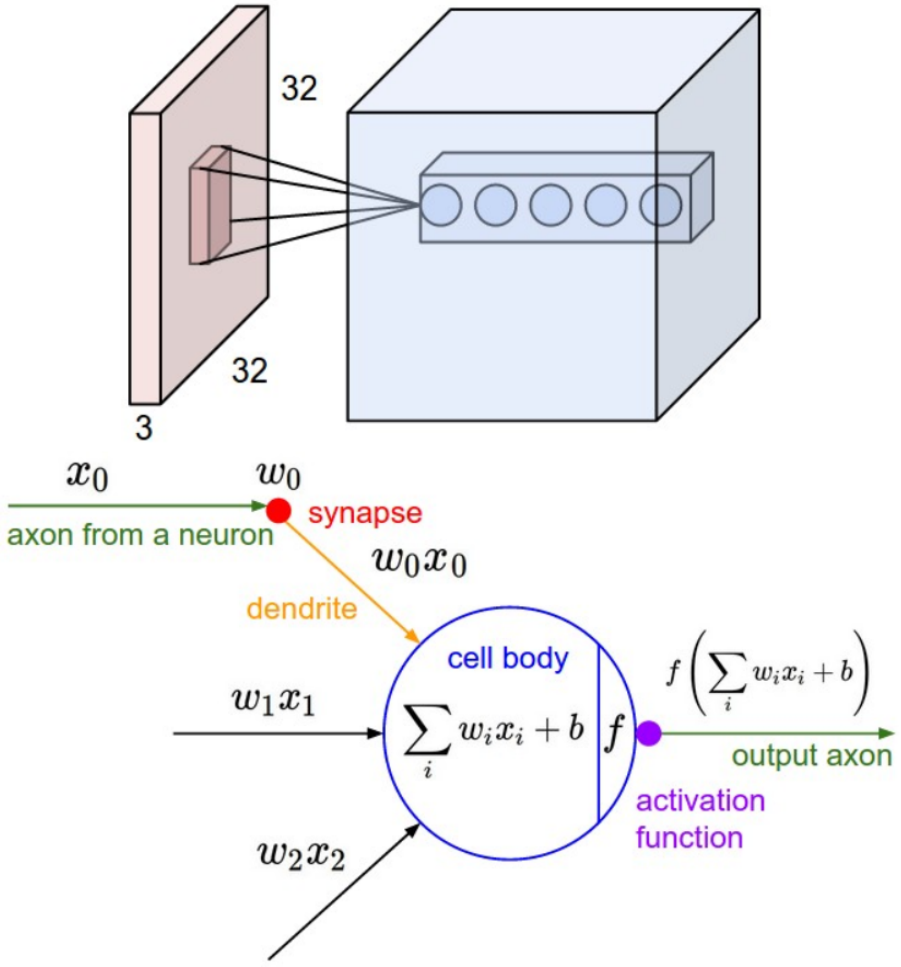


Figure 2.8: Process of the feed-forward in CNN structure.

analyze them as traditional image processing. When the input type is a raw signal, the CNN can also extract different kinds of features. 1D convolution layer creates a convolution kernel that is convolved with the layer input over a single spatial (or temporal) dimension to produce a tensor of outputs. Because EEG signals are dynamic series, a 1D CNN layer can be used to get temporal information from each channel. Depthwise 2D convolution is a type of convolution in which a single convolutional filter is applied to each input channel. It is implemented via the following steps: 1) Split the input into individual channels; 2) Convolve each input with the layer's kernel (called a Depthwise kernel); 3) Stack the convolved outputs together (along the channels axis). Unlike a regular 2D convolution, depthwise convolution does not mix information across

different input channels. Fig 2.9 is the process of using depthwise convolution to get the feature map from each channel from the input. After using a 1D convolution layer to get each channel's temporal feature, depthwise convolution is usually used to learn information from all channels. For instance, there are 20 channels used in the EEG

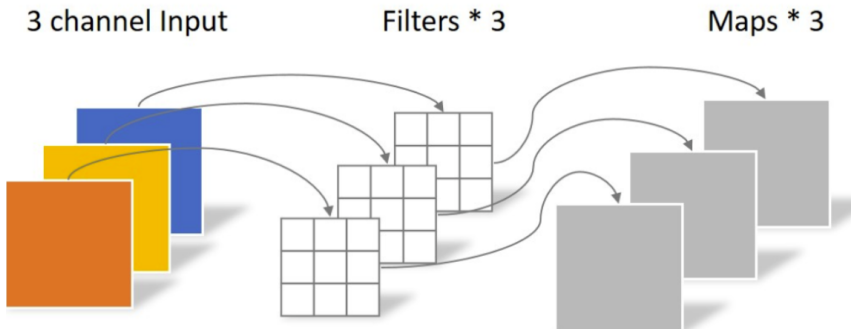


Figure 2.9: Process of the feed-forward in CNN structure.

signals. Each instant time point has 20 feature samples. Then each instant timepoint map is regarded as a channel. If there are 500 time points, it will be 500 channels and each channel has 20 feature samples. The depthwise convolution can get the information from these 20 feature samples which are regarded as the spatial feature. Therefore, 1D convolution and depthwise convolution are always used to learn temporal and spatial features from EEG signals in a CNN model. The flexibility of the CNN layer brings different styles to different kinds of features.

There are three classical variants of CNN models namely Inception (GoogLeNet) [98], VGGNet, and ResNet. GoogLeNet is notable for its "Inception modules," which allow the network to perform convolutions with multiple filter sizes in parallel. This design (shown in fig 2.10) enables the network to capture features at various scales, making it highly efficient in terms of both computation and memory usage. Riyad et al [45] developed a ConvNet based on Inception and Xception modules to extract temporal and spatial features. Amin et al [99] combine the attention mechanism with the inception module to capture features based on importance from MI data. Zhang et al [100] adopted an inception-time network to analyze EEG signals, which showed to be highly efficient and accurate for time-series classification.

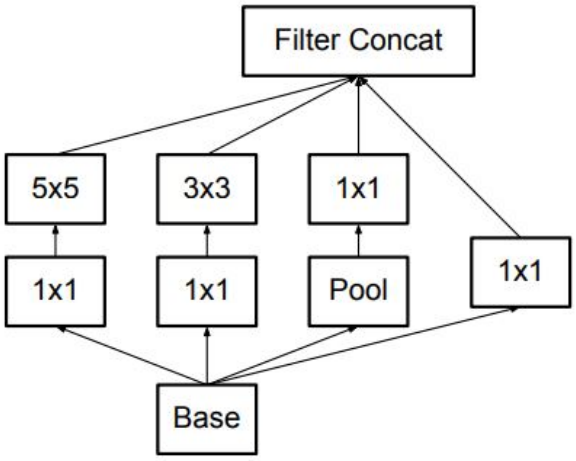


Figure 2.10: Inception modules where each 5 × 5 convolution is replaced by two 3 × 3 convolution.

VGGNet is characterized by its use of very small (3x3) convolutional filters and a deep architecture, with networks typically having 16 or 19 layers (shown in fig 2.11). Despite its simplicity, VGGNet is highly effective and has been widely used in image classification tasks [101]. Owing to its deeper architecture, the variants of VGGNet are frequently employed in the construction of pre-trained models using extensive historical data. Through the application of transfer learning strategies, specific parameters are fine-tuned, thereby extending the model's generalization capabilities. For instance, Li et al [102] adopted pre-trained VGG-16 CNN model for MI classification and fin-tuned parameters based on the target domain data. Xu et al [103] introduced a framework that utilizes a VGG-16 CNN model pre-trained on ImageNet, paired with a target CNN model that mirrors the VGG-16 architecture, except for the softmax output layer. The parameters from the pre-trained VGG-16 model are directly transferred to the target CNN model, which is then employed for the classification of MI EEG signals.

ResNet is also is known for its deep architecture, which can go hundreds of layers deep. The key innovation is the introduction of "residual blocks," where the input to a layer is added directly to the output of a few layers ahead (shown in fig 2.12), helping to address the vanishing gradient problem and allowing for the training of very deep networks [104].This module is highly flexible, allowing it to be used in combination

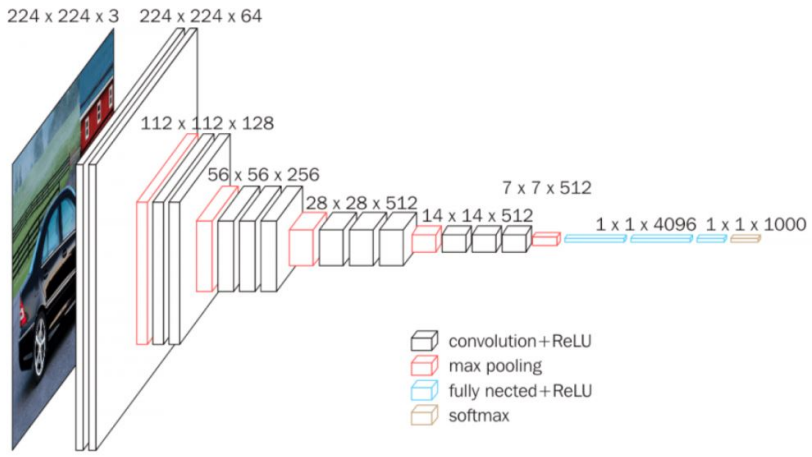


Figure 2.11: The framework of the VGGNet.

with many other modules. For example, Khademi et al [105] utilized pre-trained CNN networks, specifically ResNet-50 and Inception-v3, within a hybrid network to help address the challenge of limited MI-EEG dataset size. Jia et al [106] employed ResNet in the graph convolutional neural network to address the degradation problem led by deeper networks.

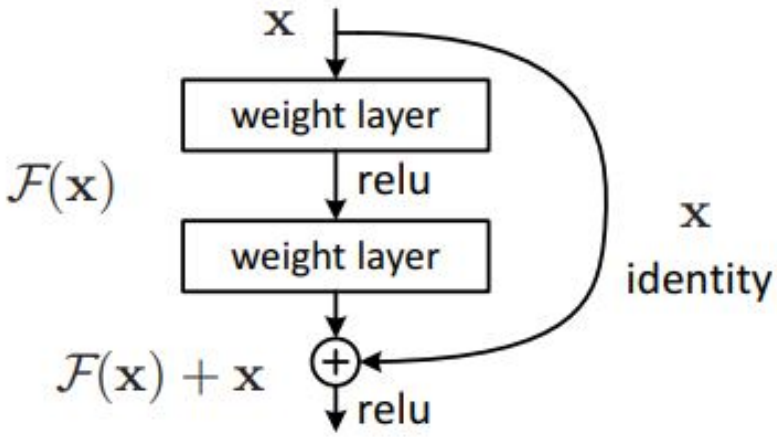


Figure 2.12: The framework of the residual block.

2.3.2.2 Long Short-Term Memory

Besides CNN, recurrent neural network (RNN) is also a popular model in the field of BCI. 1D convolution layer is used to get temporal features. In theory, RNN is better

for learning temporal features. Long short-term memory (LSTM) is an artificial RNN architecture [107] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images) but also entire sequences of data (such as speech or video). For example, LSTM applies to tasks such as unsegmented, connected handwriting recognition[108], speech recognition [109], and anomaly detection in network traffic. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models, and other sequence learning methods in numerous applications. The forget gate is used for discarding useless information of the prior LSTM cell, Considering the EEG signal is a dynamic time series from electrode measurements, LSTM may get temporal features from the EEG signal better than CNN. However, Tayeb et al compared the classification accuracies achieved using the developed neural classifiers (RCNN, LSTM, pCNN) and two other models dCNN and sCNN proposed by Schirrmeister et al [110]. The LSTM-based raw EEG data approach did not outperform any of the other developed models and the results remained slightly inferior to those obtained by state-of-the-art methods. There may be several reasons. Firstly, the EEG signals have low SNR which causes the LSTM model not to gain useful temporal features. Secondly, the relationship between different electrodes and brain regions or other space features is hard to learn through LSTM. LSTM is not as flexible as CNN. Therefore, more studies prefer using CNN. To sum up, the CNN model is widely used in EEG signal processing. Although RNN and LSTM do not perform as well as CNN, it still deserves to study the hybrid of CNN with other types of models such as LSTM, VAE et al.

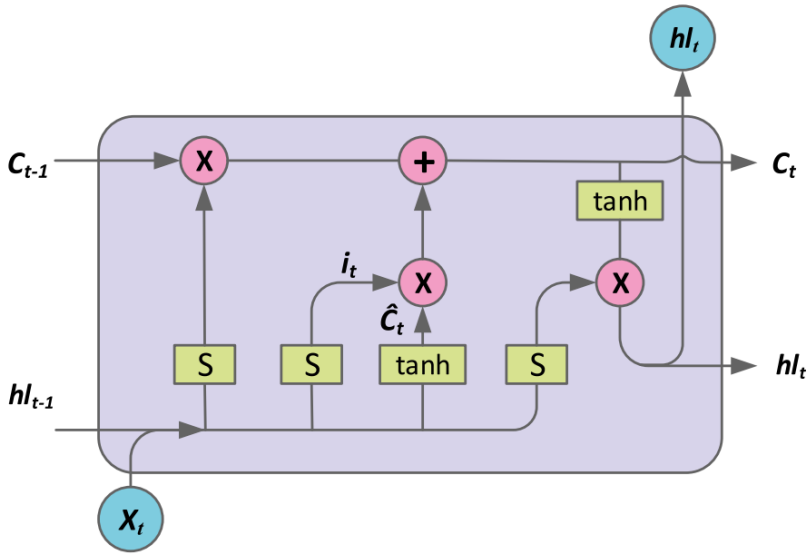


Figure 2.13: The framework of LSTM cell, 'S' denotes sigmoid activation function, 'tanh' denotes hyperbolic tangent activation function, '+' is plus, and '×' is multiplication. The 'C_t' represents the state of the LSTM cell at the current moment. The 'C_{t-1}' represents the state of the LSTM cell at the last moment. The 'hl_t' represents the output of the LSTM cell at the current moment. The 'hl_{t-1}' represents the output of the LSTM cell at the last moment.

2.3.2.3 Transformer

The Transformer model is a deep learning architecture that has gained popularity for its effectiveness in various tasks. Unlike traditional RNNs or CNNs, the Transformer architecture relies entirely on self-attention mechanisms to weigh the importance of different input tokens when generating output tokens [111]. This self-attention mechanism allows the model to capture long-range dependencies in sequences more effectively, making it particularly suitable for tasks involving sequential data. The Transformer model consists of an encoder-decoder structure, where the encoder processes the input sequence and generates a representation that can be sent to the classifiers or decoders (shown in Fig 2.14).

The multi-head attention mechanism in the Transformer model enables the model to focus on different parts of the input sequence simultaneously. It achieves this by projecting the input embeddings into multiple subspaces namely Query (Q), Key (K), and Value (V) (shown in Fig 2.15), and computing attention scores independently for each

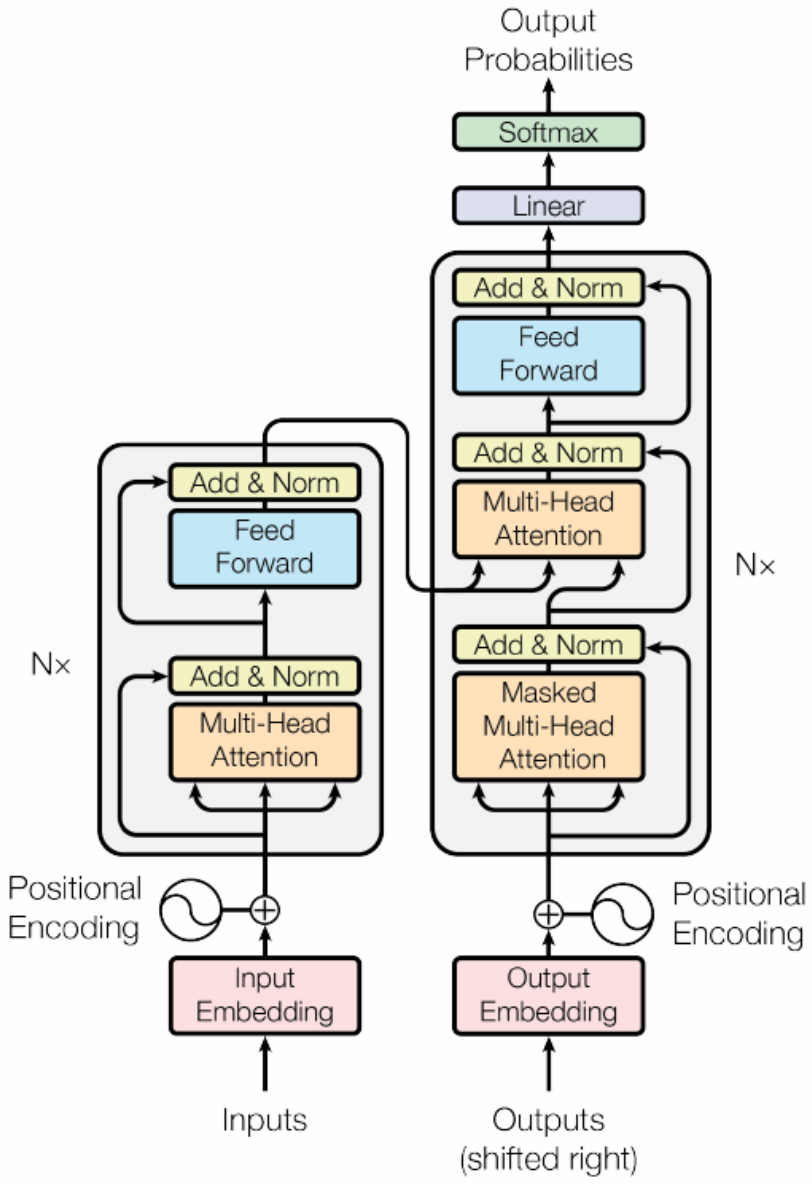


Figure 2.14: The Transformer model architecture.

subspace. These attention scores are then combined across all heads, allowing the model to attend to different aspects of the input sequence in parallel. This parallelization (Fig 2.16) enhances the model's ability to capture diverse relationships and dependencies within the input sequence, leading to more effective representation learning. Additionally, the use of multiple attention heads provides the model with the flexibility to attend to different parts of the input sequence with varying levels of granularity, enabling it to capture both local and global dependencies effectively.

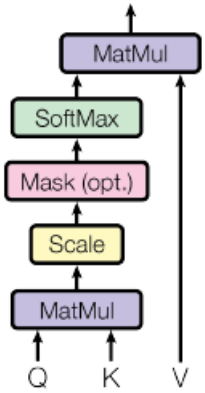


Figure 2.15: Scaled Dot-Product Attention framework.

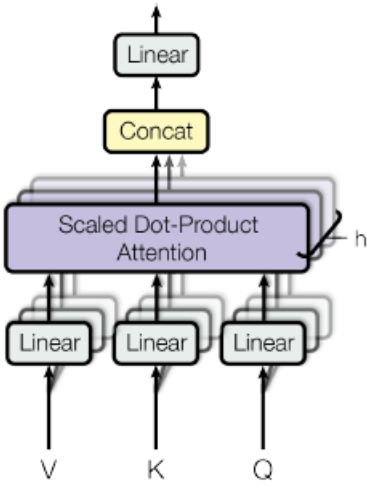


Figure 2.16: Multi-Head Attention framework.

In the BCI field, the transformer is also adapted to handle signals in the applications such as person identification [112], emotion recognition [113], visual stimulus classification [114] and signal denoising [115]. For MI-EEG decoding, Ma et al [116] proposed a hybrid CNN-Transformer model to weigh spatial features and frequency signals by employing the attention mechanism. Song et al [117] proposed a hybrid model with six transformer encoders after extracting features from MI-EEG by CNN layers. Tao et al[118] employed the gating mechanism on the transformer to improve the model performance. Xie [119] designed five hybrid models with different layers in the CNN and transformer. Although research on using transformers for EEG decoding is still limited overall, transformers remain a promising tool for exploring long-time-series data like

EEG signals.

2.4 Transfer Learning-Based Decoding Algorithms for MI-EEG

Transfer learning solves insufficient data problems by adjusting the model via prior knowledge to make it adaptable to new tasks. The transfer learning model keeps learning ability without a large amount of data based on prior knowledge learned in related tasks. Unlike traditional machine learning, the focus of transfer learning is on the target task with one or multiple source tasks which are trained to provide priori knowledge for the target task [120]. Transfer learning is mainly used for data augmentation and saving time for building a new model by transfer parameters. Santana et al. proposed a cross-subject classifier to predict the stimulus presented to a subject from the analysis of the brain activity [121]. In the previous study, there are three transfer learning approaches based on EEG [122]:

- Feature representation transfer: Encode the transferred information into a new feature representation, which enables later classification models to generalize well on the target test set. Discriminative information across subjects or sessions may be transferred to reduce calibration time, and stationary information may be transferred to avoid repeated calibrations before each use of BCIs. So far, many transfer learning methods in this setting have learned a new feature representation by spatial filtering with data collected and organized in matrixes.

- Instance-transfer: Transfer information by reusing certain parts of data from the source domains. Different from the feature-representation-transfer approach, the instance-transfer approach transfers discriminative information by weighting data from source domains instead of changing their feature representations. Approaches in this case weight data from source domains to reuse certain parts of data. How-

ever, few instance transfer approaches have been proposed to transfer stationary information across domains.

- Classifier-transfer: Reuse classifiers learned from other domains to aid the target task. Domain adaption of classifiers and ensemble learning of classifiers are two major techniques. This case can be further organized into two rough subcategories: domain adaption of classifiers and ensemble learning of classifiers. Domain adaption of classifier is a common method, which reuses learned classifiers from source domains and adjusts classifier parameters according to the target domain. This type of approach generally requires that there are enough parallels between source domains and target domains. Therefore, a common scenario using domain adaption of the classifier is to transfer the discriminative and stationary information from session to session. Different from the former, ensemble learning of classifiers is a promising method to combine base classifiers learned from multiple domains into a single one. Generally, diversity among the base models is deemed to be a primary cause for the final classifier to obtain an accurate performance and a good generalization ability.

2.5 Summary

This section has performed a comprehensive review of MI-BCI background and corresponding EEG decoding algorithms. In Section 2.1, the principles of MI-BCI systems are introduced and prove the effectiveness for rehabilitation. By decoding EEG signals, the recognized motor intentions can be employed as commands to control external devices or bring neuro-feedbacks which help patients get involved in the rehabilitation training for a better recovery effect. The ERS/ERD phenomenons generated by the correlated brain area based on MI stimulations provide theoretical foundations for classification algorithms. In Section 2.2, conventional machine learning approaches including feature extraction and classification are presented. Although previous studies have gained progress on MI-EEG decoding, the mismatch between feature extraction and classifi-

cation may limit the capability of machine learning models. Deep learning has been proven a potential tool for analyzing signals in diverse fields and Section 2.3 shows two important factors namely input formulation and model structure in DL methods. The main types of input include images, signal values, and calculated features while the mainstream structures involve CNNs, LSTM, and transformers. Selecting appropriate input signals and designing efficient model structures will be discussed in the next four chapters. Section 2.4 investigates the transfer learning strategy for improving DL models' performance and practicability which will be used in Chapter 6.

CHAPTER 3

A Multi-View CNN Decoding for MI-EEG Signals

Since CNN has gained lots of attention in decoding MI-EEG for improving stroke rehabilitation strategies, the extremely non-linear, nonstationary nature of the EEG signals and diversity among individual subjects results in the overfitting of a CNN model and limits its learning ability. In this chapter, a densely connected convolutional network with multi-view inputs is proposed. First, different data subsets from the original EEG signals are created as the CNN model inputs through bandpass filters applied to the EEG signals to generate multiple frequency sub-band signals based on brain rhythms. Then, temporal and spatial features are captured based on the whole frequency band and the filtered sub-band signals, respectively. Further, two dense blocks with multi-CNN layers, which connect each layer to every other layer in the feed-forward path, are used to enhance the model learning capabilities and strengthen information propagation. Finally, a concatenation fusion method is used to integrate the extracted features and a fully connected layer for finalizing the classification. The proposed method achieves an average accuracy of 75.16% on the public Korea University EEG dataset which consists of the EEG signals of 54 healthy subjects for the two-class MI tasks, demonstrating that the proposed method effectively extracts much richer MI information from the EEG signals and improves classification accuracy.

3.1 Introduction

The DL-based algorithms have been demonstrated to be effective in decoding MI-EEG signals. However, few models considered the influence of frequency band range which varies from subject to subject, and the reuse of features in a deep learning structure. There exist two key challenges in frequency band selection in the data preprocessing step. The first is the difficulty to determine brain rhythms as they vary with time and with different subjects, the other is the diversity among different subjects. The brain rhythms reflect the functional states of different neuronal cortical networks. The most common frequency band used in the MI-EEG field is α rhythm [123] which is about 10 Hz and β rhythm which is around 20 Hz [38]. In the research [67], 7 - 30 Hz was selected as the α and β rhythm. In another study [12], it is shown that 26 Hz is the upper limit of the β rhythm. Besides that, θ rhythm (4-7 Hz) was also proved useful in decoding MI-EEG signals [124, 125]. It is clear that, the frequency band range for special brain rhythms is found to be different in these studies. Meanwhile, Novi et al [69] proved that the variation among different subjects is enormous. That is also the reason that FBCSP and other later methods focused on selecting the appropriate operational frequency band for extracting discrimination. In DL methods, if no band selection is performed before building the classifier model, redundant information and noise will lead to the overfitting problem and make it difficult to learn useful features. However, on the other hand, if the band selection is performed, the differences among different subjects may have a great impact on the final classification result. All these have to be taken into account.

The other challenge is the lack of reusable feature maps in the deep learning model. In the computer vision field, many works have demonstrated that the CNN model with shorter connections between layers close to the input and the ones close to the output can be trained more efficiently and accurately, and these shorter connections can help reuse features in the previous layers [126]. ResNets [104] redesigned the layers that learn residual functions concerning the layer inputs and side signals from one layer to the next through connections. Larsson et al [127] proposed FractalNets that combine several parallel layer sequences with different numbers of CNN and maintain many

short connections in the network. The densely connected network (DenseNet) [126] uses direct connections between any two layers with the same feature-map size. Short paths lead to feature map reuse which can ensure efficient information flow between layers in the network and improve model learning capabilities. Liu et al [128] connected the DenseNet with 3D CNN to decode MI-EEG. However, the 22 input channels were forced to be the size of 7×7 by padding with the mean of all the EEG signals which introduced lots of redundant information. Yu et al [129] proposed a model combining the attention mechanism with DenseNet. However, the model only used three channels (C3, C4, Cz) to generate the input images by Continuous Wavelet Transform (CWT), which ignored massive spatial information provided by the other channels in the cortex.

To address the two outstanding issues in deep learning for EEG signal decoding in BCI systems, this chapter proposes a novel end-to-end CNN architecture with multi-views of EEG signals based on the densely connected convolutional network. First, the MI-EEG signals are fed to a 3-order Butterworth filter with different frequency bands (1-5 Hz, 4-8 Hz, 8-13Hz, and 13-32 Hz) according to the brain rhythms. Then each of the four EEG sub-band signals and the raw data covering the whole frequency band is fed into a CNN model respectively. Then two CNN layers are used to capture temporal and spatial features. Next, these features pass two dense blocks with multi-CNN layers, which helps to connect each layer to every other in a feed-forward mode. In the dense block, the feature maps extracted by the preceding layers are used as inputs, and their feature maps are fed into the succeeding layers, which help reuse feature maps and reduce overfitting on tasks. The final extracted features from each kind of input signal are fused in a fully-connected layer and end with the softmax classifier.

The remainder of the chapter is organized as follows. The details of the proposed method including data description, preprocessing steps, the detailed structure and parametrization of the proposed CNN model are given in Section 3.2. Experimental results and discussions are presented in Section 3.3 and Section 3.4 respectively. Finally, Section 3.5 concludes Chapter 3.

3.2 Methods

This chapter first introduces the public dataset used in the experiment. Then the preprocessing steps and the proposed method are given in detail. The structure and parameters are shown at the end.

3.2.1 Data Description

1) The public Korean University (KU) dataset [94] which includes fifty-four healthy subjects (ages 24-35; 25 females) is the largest binary dataset so far available in the public domain. Every subject had 200 trials of data (100 trials for imaging the left hand and right hand respectively). The EEG signals were collected with 62 electrodes based on the standard international 10–20 system placement. The sampling rate was 1,000 Hz. To ensure a fair comparison with other methods, the raw signals was downsampled to 250 Hz. 20 electrodes from the region related to motor function were selected (shown in Fig 3.1). The channel selection is based on the previous model [93, 130] which performed well on this dataset. The black fixation cross on the center of the monitor lasted for 3 seconds for subjects to prepare for the MI task. Next, the subject was asked to image the MI task for 4s according to the left or right arrow that appeared on the monitor and relaxed after a blank screen. We only intercept 4 s data for the MI tasks for the subsequent processing. The 10-fold cross-validation (CV) method was adopted to check the performance and robustness of our proposed model. 8 folds are used for training, 1 fold for testing, and the remainder for validation. Hence, for each subject's model, a total of 160 training trials and 20 validation and testing trials were administered.

2) BCI Competition IV 2a (BCIC-IV-2a) dataset [131] consists of recordings from 9 healthy subjects performing 4 different MI tasks: left-hand, right-hand, both-foot, and tongue. The signals were acquired using 22 EEG electrodes with a sampling frequency of 250 Hz and were bandpass filtered between 0.5 Hz and 100 Hz, as well as notch filtered at 50 Hz. Two sessions were recorded on different days for each subject, with each session comprising 288 trials. So there is an inherent drift in the statistical distributions between

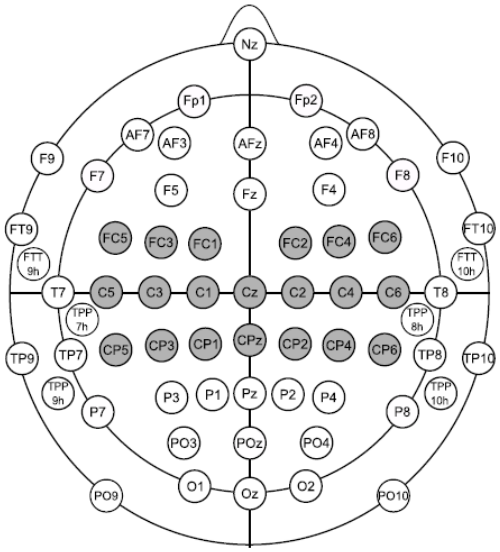


Figure 3.1: EEG electrodes position (KU dataset). The EEG electrodes shown in gray were used in the proposed model.

the two sessions. We used the first session for training and the second session for testing to validate the performance of the proposed model. Each trial lasts 4 seconds for the MI task, and we used the entire length of the data for decoding.

3.2.2 Preprocessing

We use a 3-order Butterworth filter to obtain multi-views of EEG signals based on different brain rhythms. Compared with the Chebyshev filted used in the [93], Butterworth filter provides a flat passband without ripples, making it ideal for applications requiring consistent signal quality. It also has a more linear phase response, reducing phase distortion and preserving signal integrity. We select four sub-bands with 1 Hz overlap, which are δ rhythm (1-5 Hz), θ rhythm (4-8 Hz), α rhythm (7-13 Hz) and β rhythm (12-32 Hz). As aforementioned, earlier work has already shown that the signals from these sub-bands can better decode MI-EEG signals and identify the intentions of the subjects. However, suitable boundaries of sub-bands for MI-EEG signals vary from person to person. To ensure that the feature diversity of each individual is properly learned, raw signals with the overall band covering the full frequency spectrum are also fed into the proposed model. Therefore, altogether five types of EEG signals are used as

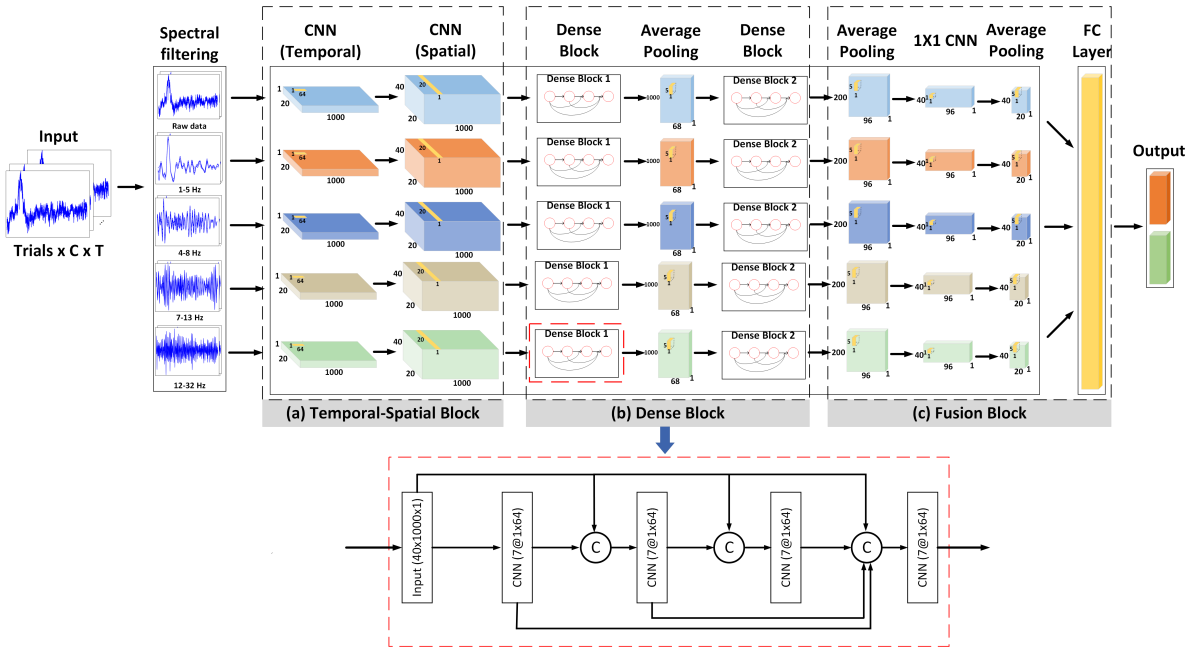


Figure 3.2: The overview of the proposed model structure. The structure is divided into three blocks: (a) Temporal-Spatial Block; (b) Dense Block; (c) Fusion Block. In the Dense Block, $7@1 \times 64$ means each CNN layer in the dense block has 7 filters with the size of (1×64). C means the concatenation procedure.

inputs including specific information related to the MI and the diversity among subjects.

3.2.3 The Proposed Model

The schematic of the proposed model is shown in Fig 3.2. Each view of specific MI-EEG signals is processed using the same deep-learning structure. Therefore, the final fully connected layer receives the same size of features through these five parallel structures.

3.2.3.1 Definitions

Define the raw EEG signals as $E = (X_i, Y_i)|i = 1,2,...,N$, where $X_i \in R^{C \times T}$ represents $i-th$ EEG trial with C channels and T samples. N is the total number of EEG signal trials. In our experiment, C equals 20 and T is 1000 because each trial has 4 s data with a 250 Hz sampling rate. Y_i is the matching label of X_i, which comes from the label set $M = \{m1 : right, m2 : left\}$.

3.2.3.2 Temporal and spatial block

As a time series, EEG signals contain abundant temporal features. Meanwhile, the different electrode positions on the scalp allow for spatial characterization of brain activities. Therefore, we use two CNN layers to extract temporal and spatial features which are the most important and commonly used in decoding MI-EEG tasks. First, a CNN layer with the kernel size of $1 \times k$ is adopted on each channel to perform a convolution over time. According to the experience of previous studies [132, 133] and ablation experiments on the kernel size selection, we allow $k = 64$ in the proposed model. Then, a depthwise CNN layer is applied across all channels. The number of the filter is set to one so that signals from all channels at each time instant were compressed into one feature map. This approach facilitates a reduction in the number of trainable parameters and enables the efficient extraction of features [132]. No activation function intervenes between the two layers [85]. Then, the exponential linear unit (ELU) [134] and batch normalization [135] techniques are employed to mitigate the overfitting problem.

3.2.3.3 Dense block

Assume that the network has L layers, each of which implements a non-linear transformation $F_l(\cdot)$ where l is the index of the layer and the output of each layer is x_l. Traditional transition with a single connection between each CNN layer is:

$$x_l = F_l(x_{l-1}). \tag{3.1}$$

As the network becomes deeper with more layers, some useful features may also be filtered. Meanwhile, more parameters need to be optimized which leads to overfitting, especially for MI-EEG signals with a limited number of subjects. To address this problem, we refer to [126] and build direct connections from any layer to all subsequent layers. The detailed structure is shown in Fig 3.2. In a dense block, the lth layer receives the feature maps of all preceding layers, and the activation function is:

$$x_l = F_l([x_0, x_1, ..., x_{l-1}]). \tag{3.2}$$

where $[x_0, x_1, ..., x_{l-1}]$ are the feature maps from the preceding CNN layers $[l0, l1, ..., l-1]$. Each CNN layer adopts ELU as activation, followed by batch normalization and dropout techniques. If a CNN layer produces k new feature maps, the lth layer has $k_0 + k \times (l-1)$ inputs, where k_0 is the number of feature maps from the input. The k is called the growth rate of the network which reflects how much new information a CNN learned and contributed to the proceeding layer. The short paths from preceding layers to proceeding layers are greatly increased to $\frac{L(L+1)}{2}$ whereas traditional architecture only has L ones in a L layers network. Through the dense block, each layer has access to all the preceding feature maps which enhances the flow of information and feature reuse.

In summary, in the proposed model, the dense block receives an input consisting of 40 feature maps. To mitigate the risk of overfitting, we set k=7, and use 4 CNN layers in each block. Following the first CNN layer, we concatenated the 7 extracted feature maps combined with the original 40 feature maps to form a new input that is then fed into the subsequent CNN layer to learn 7 additional feature maps. This process is repeated before the third CNN layer, thereby allowing for the extraction of more informative and discriminative features. Each subsequent layer of CNN is not only connected to the previous layer but also has connections to the outputs of all previous layers, enabling the establishment of shorter paths that help the flow of the information.

3.2.3.4 Fusion block

As mentioned earlier, in addition to the complete raw EEG data, four different inputs represent different views of the MI-EEG signals. These sub-band datasets are based on brain rhythms related to MI tasks. While the raw signals without filtering reflect the diversity of features that varies from person to person. After further learning by two dense blocks, the extracted features are sent into a 1×1 CNN layer to reduce the size of feature maps before fusion. Fewer maps help to improve computational efficiency and reduce fluctuations in the loss function trajectory along the training. Finally, features obtained through all parallel branches are fed to a fully connected layer with a softmax classifier.

Using the single-band input can reduce a number of computational parameters, as demonstrated in classical models such as ConvNet [85] and EEGNet [132]. However, this approach often ignores the impact of differences across different frequency bands. When using multi-branch input that divides EEG into different frequency bands, the learned features need to be integrated before being fed into the classifier. The previous models such as HS-CNN [133] used a simple fully connected layer to fuse the information. In contrast, our model has multiple layers for each branch, leading to numerous calculated feature maps. As a result, we need to reduce the dimensionality through one-dimensional CNN and pooling procedures to prevent overfitting and reduce the model complexity.

3.2.3.5 Training

The cross-entropy function is selected as a loss function which calculates the distance between the probability distribution of the neural network prediction values y_p and the true labels y_t [136]:

$$L\left(y_p, y_t\right) = -\sum_m y_{p,m} \log y_{t,m}. \tag{3.3}$$

where m is the index of y. The optimizer is Adam [137] and learning rate $lr = 0.0001$. The training takes 1000 epochs for each fold in the CV with 16 batches per epoch. The early stopping technology was used to save the best weights during each fold. The training step ended after checking if the validation loss value decreased for the last 150 epochs. After reaching the threshold, the model with the best weights produces the classification results of the test fold.

3.2.3.6 Baseline models

We use two traditional machine learning methods (CSP [138] and FBCSP [70]) and three state-of-the-art deep learning architectures (Shallow ConvNet [85], Deep ConvNet [85], and EEGNet 8-2 [132]) as baseline models to demonstrate the effectiveness of our proposed method. Since the baseline models use different datasets in the initial research, we have used the best parameters of these models and ensure a fair comparison. The details of the baseline models are described as follows:

1. CSP: The basic principle of the CSP algorithm is to find the optimal set of spatial filters for mapping data by using the diagonalization of the matrix. In this way, the difference between the variance values of the two tasks is maximized, thus gaining a feature vector with high discrimination. The methods CSSP, CSSSP, and DFBCSP mentioned earlier all use CSP as the kernel algorithm.

2. FBCSP: FBCSP is also a successful algorithm commonly used in the BCI field. After extracting the features through CSP, FBCSP used a feature selection method to automatically select discriminative pairs of frequency bands. According to [70], we decompose EEG signals into nine frequency bands with a bandwidth of 4 Hz from 4 to 40 Hz through a Chebyshev filter. The classifier is the SVM with the default kernel radial bias function (RBF).

3. ConvNet: ConvNet included the shallow and deep two structures. Shallow ConvNet is a DL model with only two CNN layers and an average pooling layer, which achieved a better performance than FBCSP on the public dataset BCI competition IV 2a [139] and high-gamma dataset [85]. Deep ConvNet includes a temporal and spatial filter which are similar to the head of the Shallow ConvNet. Then the layer was followed by several convolution-max-pooling blocks and a fully-connected layer with a softmax classifier. It extended the choice of learning parameters and optimization plans and has achieved excellent results [140].

4. EEGNet 8-2: Based on the Shallow ConvNet, this model adopted a separable CNN layer after extracting temporal-spatial features, which ensured the quality of the classification results with reduced calculation cost.

5. FBCNet: FBCNet divides the raw EEG signals into several frequency bands. Then the depthwise CNN layer was used to extract spatial features. After that, the model employed a variance layer to compute the temporal variance of the time series. Finally, the features are fused in a fully-connected layer and ended with the softmax. The model referred to the principle of FBCSP and achieved the best classification result on the Korean public MI dataset[93]

3.3 Results

The computer used in this experiment had 8 Intel cores Intel processors and 16 GB RAM. GTX 2080 GPU with 8 GB memory was used for training and testing EEG data. Keras was used for building the proposed model and the baseline models. The results and statistical analysis of the proposed model are reported in this section, including the performance comparison with other baseline models.

3.3.1 Overall Performance

The averaged classification accuracy of different methods is shown in table 3.1. The results from different methods are 64.69% (±15.25), 63.50% (±19.09), 67.81% (±17.55), 61.83% (±16.89), 68.96% (±17.17), 73.44% (±13.28), 75.16% (±14.01) for CSP, FBCSP, EEGNet 8-2, Deep ConvNet, Shallow ConvNet, FBCNet and our proposed method in KU dataset, respectively. The proposed method achieves 1.72% higher than the best result among the baseline models. To better validate the results, we use statistical significance tests including an Analysis of Variance (ANOVA) test for the multiple comparison tests and paired t-tests between each baseline method and the proposed method. In an ANOVA test, the proposed method significantly exceeds the others [$F = 5.054$, $p < 0.001$]. The results of paired t-test are CSP [$t_{(53)} = -6.074$, $p < 0.001$], FBCSP [$t_{(53)} = -6.220$, $p < 0.001$], EEGNet 8-2 [$t_{(53)} = -5.931$, $p < 0.001$], Deep ConvNet [$t_{(53)} = -8.226$, $p < 0.001$], Shallow ConvNet [$t_{(53)} = -5.788$, $p < 0.001$] and FBCNet [$t_{(53)} = -1.875, p < 0.01$]. The performance comparison for individual subjects based on the scatter plots is presented in Fig 3.3. There is a significant improvement in classification accuracy for most of the subjects. The proportions of subjects who had classification accuracy over 80% were in the proposed model than in the baseline models are 81.4% (44 of 54), 85.2% (46 of 54), 79.6% (43 of 54), 87.03% (47 of 54), 77.8% (42 of 54), and 55.5% (30 of 54) for CSP, FBCSP, EEGNet 8-2, Deep ConvNet, Shallow ConvNet and FBCNet, respectively. In the BCI IV 2a dataset, while FBCNet exhibited 0.26% higher than the proposed model, our model demonstrated notable classification performance when compared with other baseline

Table 3.1: Comparison of average classification accuracy (%) with standard deviation (SD) for different methods.

	KU Dataset	BCI IV 2a Dataset
CSP [138]	64.69 (15.25)	54.01 (12.77)
FBCSP [70]	63.50 (19.09)	65.79 (14.21)
EEGnet 8-2 [132]	67.81 (17.55)	67.81 (17.55)
Deep ConvNet [85]	61.83 (16.89)	65.34 (13.54)
Shallow ConvNet [85]	68.96 (17.17)	68.96 (14.28)
FBCNet [93]	73.44 (13.28)	**72.71 (14.67)**
Proposed model	**75.16 (15.03)**	72.45 (14.10)

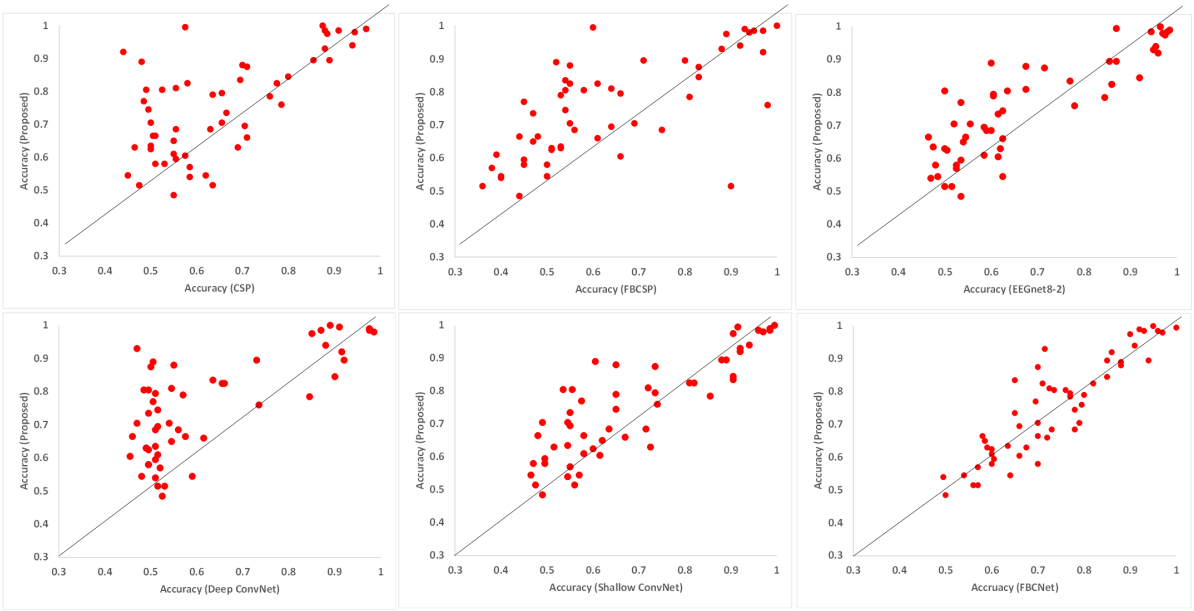


Figure 3.3: Scatter plot of individual classification performance. The horizontal axis represents the classification accuracy from baseline methods (CSP, FBCSP, EEGNet 8-2, Deep ConvNet, Shallow ConvNet, and FBCNet), and the vertical axis represents the classification accuracy from our proposed method.

models.

3.3.2 Model Performance on Different Sub-bands

Features extracted from different sub-bands have different impacts on the final classification accuracy. We tested different methods based on both the four commonly used sub-bands related to brain rhythms in MI-EEG decoding tasks and the raw data covering

Table 3.2: COMPARISON OF AVERAGE CLASSIFICATION ACCURACY(%) FOR METHODS BASED ON DIFFERENT RHYTHMS.

Algorithm	δ	θ	α	β	Overall
CSP[138]	52.93	53.81	64.50	63.29	54.39
EEGNet 8-2[132]	55.71	**54.09**	**66.63**	67.34	69.69
Deep ConvNet[85]	54.19	51.07	62.56	61.32	65.68
Shallow ConvNet[85]	53.68	52.53	64.38	**68.57**	64.16
Proposed method	**57.43**	54.07	66.52	67.47	**73.52**

the whole frequency spectrum. Our proposed model fuses extracted features through five paralleled structures with the same parameters. Therefore, to validate the effect of different sub-bands on the proposed model, only one paralleled structure without the fusion step is used in this experiment. The comparison results based on the KU dataset are shown in Table 3.2. On the θ and α sub-bands, EEGNet 8-2 achieved 54.09% and 66.63% respectively which are the best results. The Shallow ConvNet method performed well on the β sub-band. The proposed methods also produced good classification results based on different sub-bands, especially having much higher accuracy in decoding the raw MI-EEG signals over all frequency bands.

Fig 3.4 shows the accuracy of the proposed method based on different brain rhythms for each subject. The classification results based on α and β rhythms are higher than other cases for most subjects, which further confirms the findings reported in the previous research [123][38]. θ and δ rhythms also contain useful MI information which can significantly improve the accuracy of some subjects. Furthermore, the proposed model extracts the most comprehensive information from the raw MI-EEG data covering the whole frequency spectrum without applying any filtering and has achieved the best classification results for 33 out of 54 subjects.

3.3.3 Effect of Hyper-parameters

The kernel structure of the proposed model is the dense block. To improve the model learning capacity, we test the effect of different numbers of feature maps on the classification performance based on all subjects in the KU dataset. The results are shown in

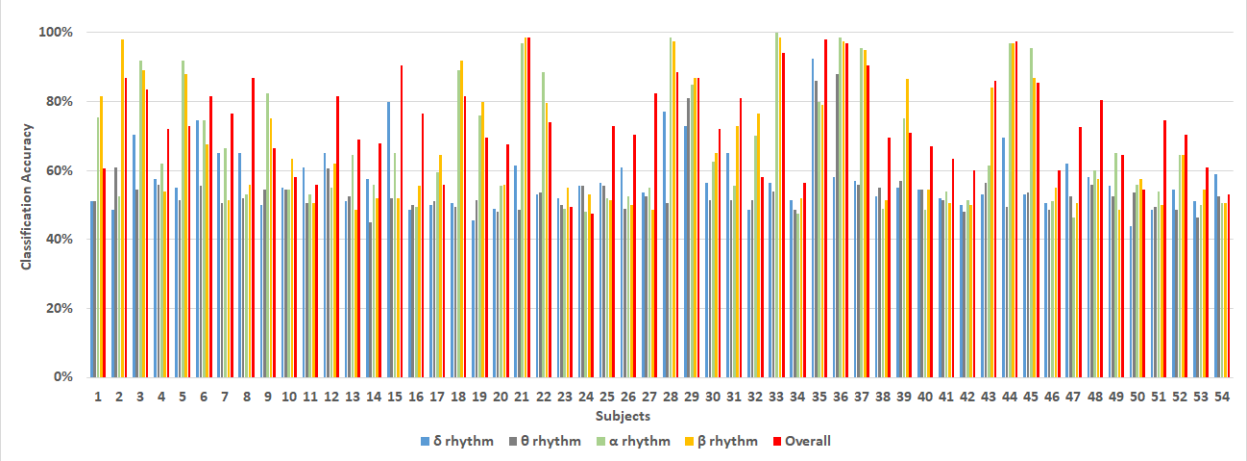
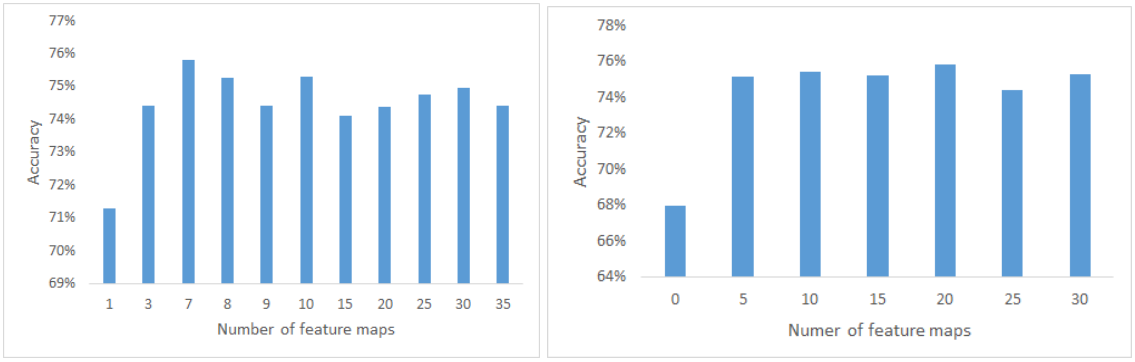


Figure 3.4: The average accuracy of the proposed method using a 10-fold CV based on different brain rhythms.

Fig 3.5(a). The highest accuracy is achieved when the CNN filter in a dense block learns 7 feature maps each time. Fewer feature maps limit the information learned while too many ones lead to the overfitting problem. Besides that, the number of feature maps in the final 1×1 CNN layer also plays an important role. The influence on the model is shown in Fig 3.5(b). When the dimension of feature maps is compressed to 20, the model has the best performance. From the results, we can find that the difference between different numbers of feature maps is not huge. However, if there are no 1×1 CNN layer in the fusion block, the classification rate only reaches about 68% which is similar to other excellent methods. One possible reason is that there are many extracted features from each sub-band signal and the raw data comprising all bands. Without decreasing feature maps by 1×1 CNN layer, a large number of parameters with redundant information will be calculated in the fully connected layer and in the final softmax layer which will have a negative impact on the final classification results. Besides that, we also test the influence of different activation functions (Fig 3.6). The ELU function performs best with the highest accuracy and runs fast.



(a) (b)

Figure 3.5: The effect of numbers of feature maps of the proposed model. (a) The feature maps come from the CNN filters in the dense block. (b) The feature maps come from the 1×1 CNN filters in the fusion block.

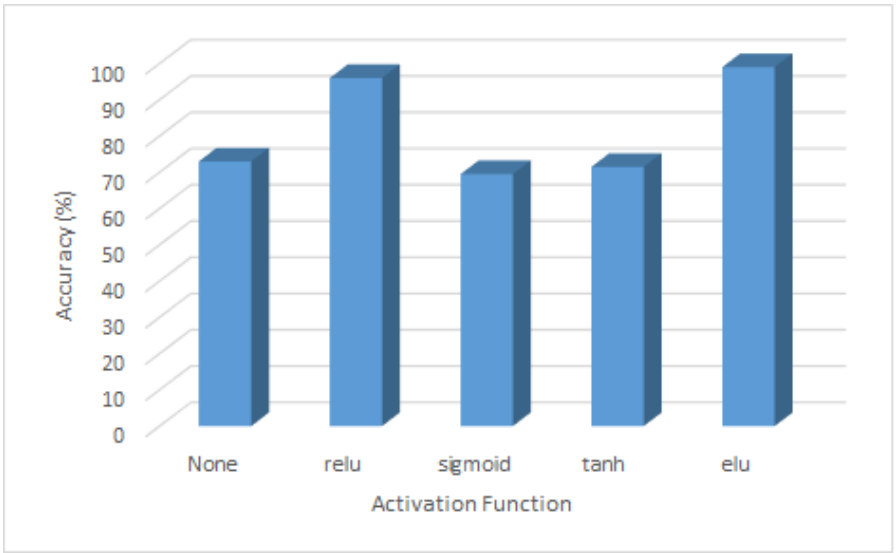


Figure 3.6: The effect of activation functions on example subject

3.3.4 Effect of Fusion of Sub-bands and The Overall Band

The classification results for methods based on different rhythms are shown in Table 3.2. Although the accuracy of the proposed model on the raw EEG signals covering the full frequency spectrum (overall band) is good enough, the combination with subset signals of specific brain rhythms can achieve better performance. As shown in Fig 3.7, without combining with the features from the raw data covering the full spectrum (whole bands), the results of the combination of only sub-bands covering brain rhythms δ, θ, α and β are no more than 68%. When we introduced the features from the overall band, the accuracy significantly improved. The proposed model with all sub-bands and the overall band together achieves 75.16% while the combination of the overall band with (α, β) and (θ, α, β) reaches 74.42% and 74.94% respectively.

3.4 Discussions

3.4.1 Comparison of Different Methods

Traditional machine learning methods included feature extraction and classification steps. Inappropriate combination of the feature extraction methods and classifiers leads to poor classification results. Deep learning, which has an end-to-end projection, shows great performance in decoding MI-EEG tasks [86]. Our proposed model uses different MI-EEG representations based on various sub-bands and raw data covering the overall band as inputs. The proposed CNN structure improves the model performance through feature reuse and fusion technology. In Fig 3.3, we show the comparison results with other methods on each subject. For most subjects, the classification accuracy has been significantly improved based on our proposed method, even more than 50% than some benchmark models. The number of subjects whose accuracy is over 80% is 11 for CSP, 15 for FBCSP, 16 for EEGNet 8-2, 12 for Deep ConvNet, 17 for Shallow ConvNet, 16 for FBCNet and 22 for our model, which further verifies the superiority and robustness of the proposed method. Additionally, we utilize the t-SNE [141] to achieve full visualization

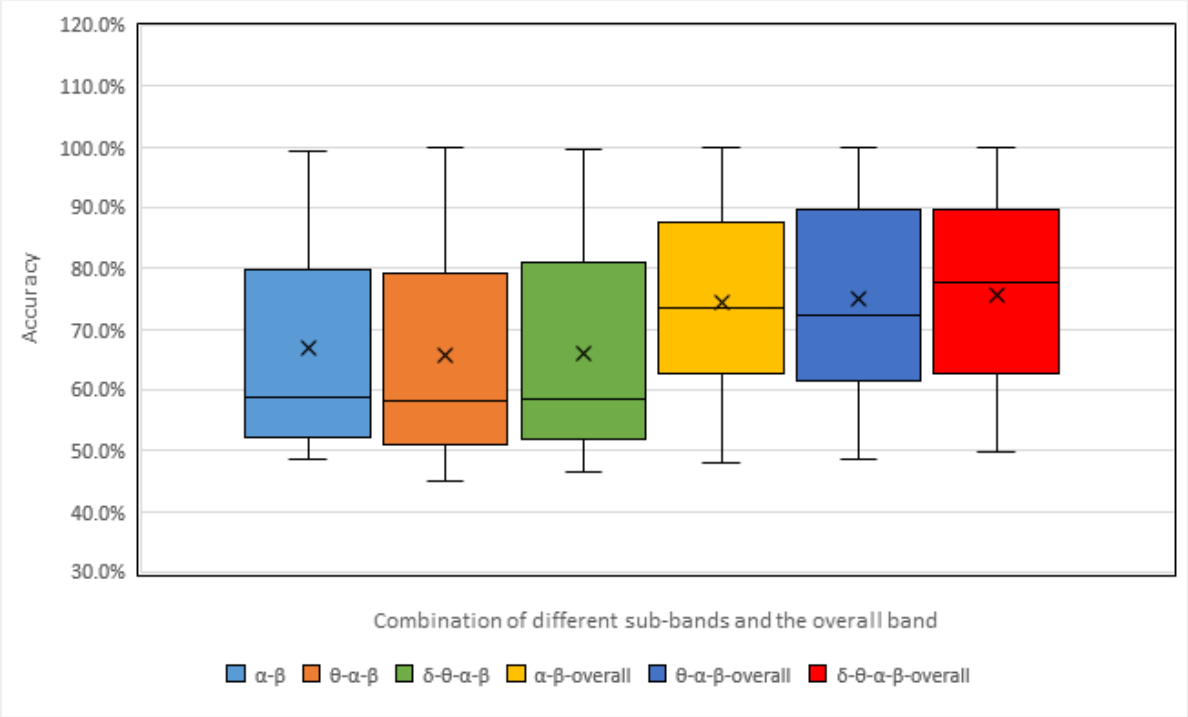


Figure 3.7: The classification rate of different combinations of sub-bands and the overall band based on the proposed model.

of the learned features from different methods. The t-SNE algorithm was used on the last fully connected layer. All inputs of the t-SNE are reshaped to trials × features to show the feature distribution in a two-dimension space. For a fair comparison, the extracted features of all methods are taken from the same subject (Fig 3.8). Compared with the other three baseline CNN models, our proposed model was able to extract more discriminative features through the dense blocks after learning temporal-spatial information from MI-EEG signals.

3.4.2 Analysis of Fusion of Features from Sub-bands

To learn useful information related to MI, the signals will pass a filter whose frequency bands are associated with the brain rhythms (commonly use α and β rhythm). The Previous study has shown the significant impact of frequency selection on the final classification results [67]. However, unclear EEG rhythm boundaries and differences in optimal frequency bands for each individual make it difficult to build an effective

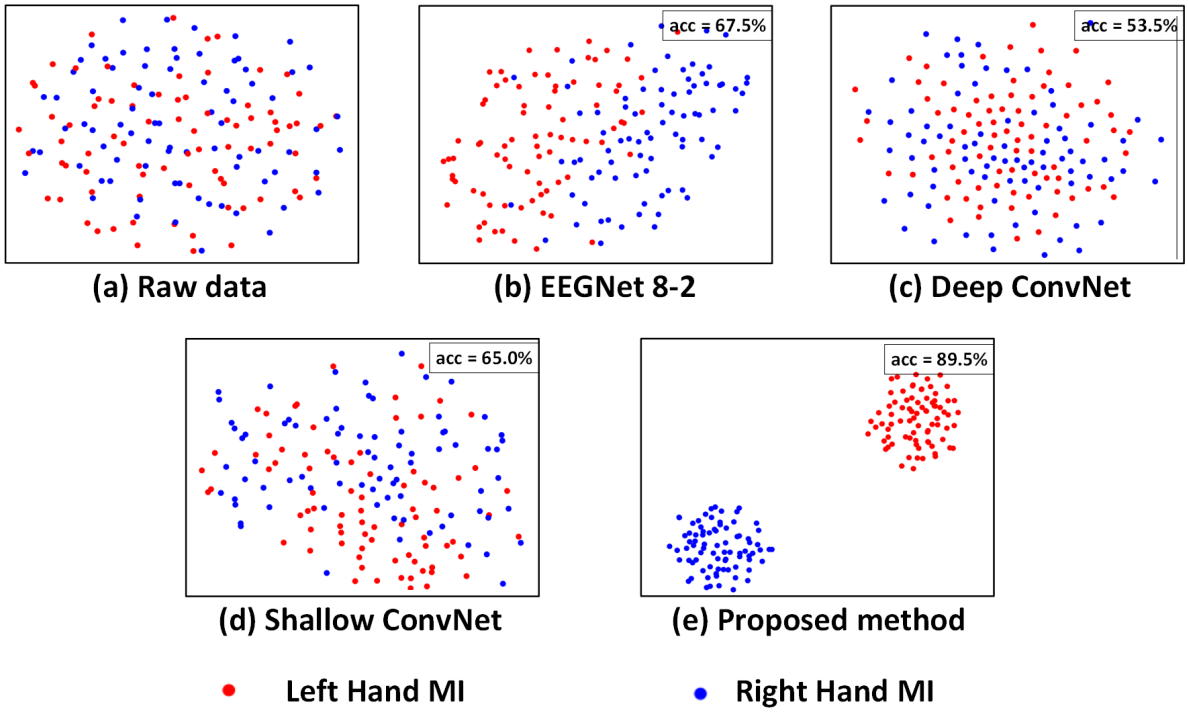


Figure 3.8: The feature map of the sixth subject obtained by various methods in 2-D embedding based on t-SNE. Part (a) is the distribution of the raw EEG data. Parts (b), (c), (d) and (e) show the distribution of extracted features in trained EEGNet 8-2, Deep ConvNet, Shallow ConvNet and the proposed method. The proposed method achieved 89.5% classification results, whereas EEGNet 8-2, Deep ConvNet and Shallow ConvNet resulted in 67.5%, 53.5% and 65.0% respectively.

model. In Table 3.2, the same method gives completely different results in decoding MI-EEG signals on different rhythms. These methods tend to give about 10% higher classification results on the α and β rhythm than on the θ and δ rhythm. Our proposed method does not show better performance than other baseline methods on these four commonly used rhythms. However, our method produces much better results on the overall frequency band which reaches 73.52%. Although using signals covering the whole frequency spectrum may introduce more redundant information and noise, it ensures that all MI information and individual diversity among different subjects. The baseline models such as the EEGNet 8-2 and Deep ConvNet also proved that the classification accuracy on the overall bands is higher than on the α or β rhythm. We again use t-SNE before the final fully connected layer to show the distribution of the extracted features

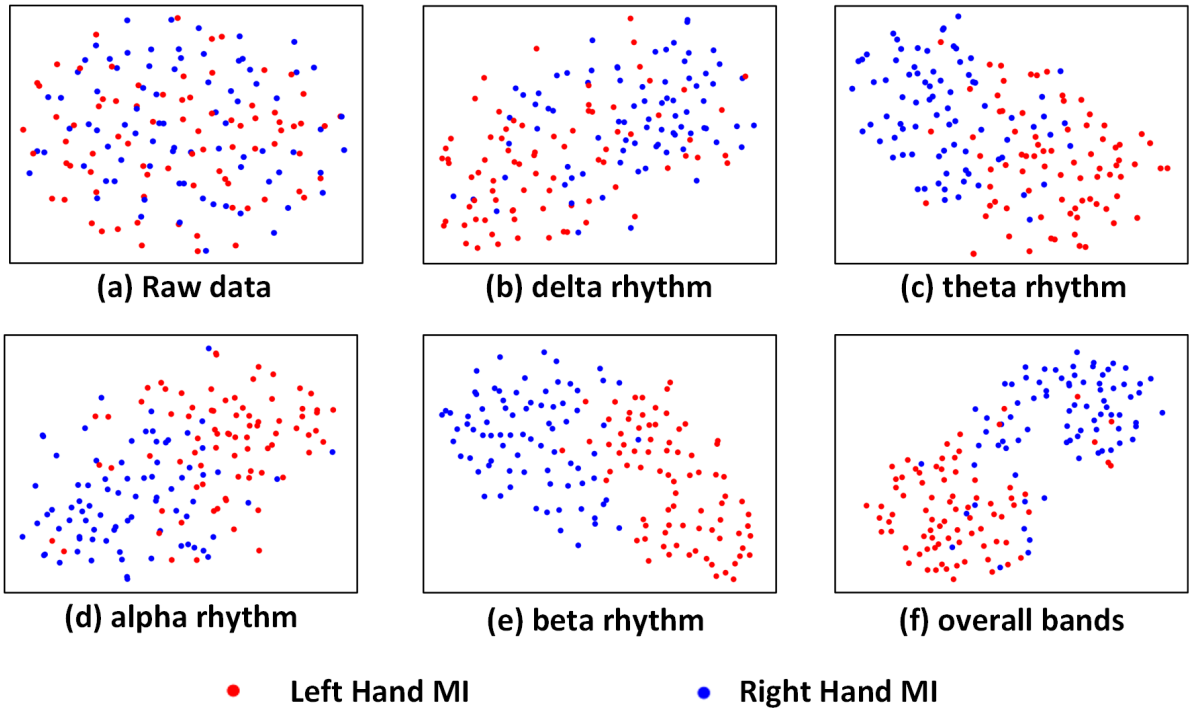


Figure 3.9: The feature map of the sixth subject with various inputs in 2-D embedding based on t-SNE. Part (a) is the distribution of the raw EEG data. Parts (b), (c), (d), (e) and (f) show the distribution of extracted features from δ rhythm, θ rhythm, α rhythm, β rhythm and the overall bands. The proposed method achieved 89.5% classification on this subject.

from different inputs (shown in Fig 3.9). Except for the features extracted from δ rhythm, other ones can be distinguished clearly through the t-SNE visualization on the sixth subject.

3.4.3 Analysis of Dense Block

Compared with EEGNet 8-2, Deep and Shallow ConvNet, our proposed model has the same structure in the first and second CNN layers to extract temporal-spatial features from EEG signals. The difference is that we have further introduced two dense blocks to make the whole structure deeper. The deeper CNN layers can usually learn more abstract and high-level features which help to improve the model performance. However, a complex model also leads to the overfitting problem, especially on limited data such as MI-EEG signals. Further, the non-linear and nonstationary characteristics of EEG may

let deeper CNN layers learn more noise and redundant information rather than useful information embedded in the signals. Therefore, previous studies in the literature usually adopted a structure with only one or two CNN layers [86]. From the results listed in Table 3.1, EEGNet 8-2 and Shallow ConvNet with fewer CNN layers than Deep ConvNet perform much better. However, fewer CNN layers may limit the learning capability of a deep learning model. In [85], Schirrmeister et al combined ResNet structure to decode MI-EEG tasks, but the performance was worse than traditional methods like FBCSP. One possible reason is that ResNet evaluates the difference between the output of one layer and the input in the preceding layer. However such information is not suitable to be fed into the following layers because simple addition and subtraction can lose useful EEG information, since the EEG signals are nonstationary and include a mass of noise signals. Our proposed model learns all features in the preceding layers instead of the difference between the inputs and outputs. In the dense blocks, the features learned by any of the CNN layers are connected by all subsequent layers which encourage information flow and feature reuse over the whole model. In the softmax classifier, the outputs are not only influenced by the features fed from the latest layer which may include redundant information due to the overfitting problem but also affected by the feature maps fed from all other preceding layers, which avoid unnecessary information loss as signals pass across through different layers. From table 3.2, it is evident our proposed model can extract more valuable features from the overall frequency bands, while reducing the effects of noise, achieving a better trade-off between model complexity and model learning capability.

3.5 Conclusions

In this chapter, a novel DL architecture based on the densely connected CNN is proposed for the recognition of MI tasks. The model employs both filtered MI-EEG signals based on four commonly used brain rhythms and the overall frequency band as inputs. The network first extracts temporal and spatial features using the first two CNN layers.

Then, two dense blocks connect each CNN layer to all the rest layers in a feed-forward mode to further learn discriminative MI-EEG information. The dense block encourages feature reuse and strengthens information propagation. Next, average pooling layers and 1×1 CNN layer help to reduce computation and avoid the overfitting problem. Finally, the fully connected layer fuses the extracted feature from different inputs and ends with a classifier. The fused features include special MI information based on different brain rhythms and also consider individual diversity among subjects without finding the optimal sub-bands. Both the classification accuracy and the distribution of extracted feature maps have demonstrated the superiority of the proposed method in the decoding MI-EEG tasks when compared with benchmark models, achieving an average accuracy of 75.16% on the public Korea University EEG datasets, higher than other state-of-the-art deep learning methods.

CHAPTER 4

Local and Global Convolutional Transformer-Based MI-EEG Classification

Transformer, a deep learning model with the self-attention mechanism, combined with the CNN has been successfully applied for decoding EEG signals in MI-BCI. In this chapter, a local and global convolutional transformer-based approach for MI-EEG classification is proposed. The local transformer encoder is combined to dynamically extract temporal features and make up for the shortcomings of the CNN model. The spatial features from all channels and the difference in hemispheres are obtained to improve the robustness of the model. To acquire adequate temporal-spatial feature representations, the global transformer encoder and Densely Connected Network are combined to improve the information flow and reuse. To validate the performance of the proposed model, three scenarios including within-session, cross-session, and two-session are designed. In the experiments, the proposed method achieves up to 1.46%, 7.49%, and 7.46% accuracy improvement respectively in the three scenarios for the public Korean dataset compared with current state-of-the-art models. For the BCI competition IV 2a dataset, the proposed model also achieves a 2.12% and 2.21% improvement for the cross-session and two-session scenarios respectively. The results confirm that the proposed approach can

effectively extract a much richer set of MI features from the EEG signals and improve the performance in the BCI applications.

4.1 Introduction

In the BCI field, the transformer based on a self-attention mechanism is adopted for MI-EEG decoding. For instance, Ma et al [116] proposed a hybrid CNN-Transformer model to weigh spatial features and frequency signals by employing the attention mechanism. However, the model uses the CSP features as inputs which loses the advantage of the end-to-end process in the DL model. Song et al [117] also proposed a hybrid model with six transformer encoders after extracting features from MI-EEG by CNN layers. The model performs well in the hold-out tests, but the huge computational costs caused by encoders limit the actual use. Tao et al [118] employed the gating mechanism on the transformer to improve the model performance but missed the extraction of EEG spatial information. Xie [119] designed five hybrid models with different layers in the CNN and transformer. This study adapted the model in the cross-subject scenario with much more training data than small data scenarios like within-subject and within-session applications, limiting the model's robustness. Besides that, the shortfall of these studies is that they only extract the spatial features from the fusion of all channels, neglecting the possible information learned from the differences between the hemispheres.

To address the above issues, a novel approach with the local and global transformer combined with CNNs for MI-EEG classification is proposed in this study. First, the local transformer and 1-dimension CNN filter with the same kernel size is adopted to extract temporal features from each channel. Although the respective fields from the local transformer and CNN are the same in the beginning, the different mechanisms allow the model to learn a comprehensive set of useful and subtle features from multi-views. The local transformer also avoids the overfitting problem compared with the global transformer which extracts more subtle features from raw EEG signals in the first layer. Then, two parallel branches use different depthwise CNNs to extract and

fuse different spatial information. One branch focuses on all channels in the motor cortex and the other one extracts the features of channels from the left and right motor regions respectively. Next, for better mining the temporal-spatial features, the Densely Connected CNN (DenseNet) [126] is used on both CNN and global transformer layers by connecting each layer to every other in a feed-forward way. The short path helps the information reuse and flow which improves the model's adaptability and robustness. Finally, the proposed model is validated and compared with other baseline models in different scenarios including within-session and cross-session to verify its performance. The remainder of the chapter is organized as follows. In Section 4.2, the materials and methods including dataset, preprocessing, scenario descriptions, and detailed model are developed. Section 4.3 shows the results of experiments and visualization. Section 4.4 is discussion and Section 4.5 concludes this chapter.

4.2 Materials and Methods

In this chapter, the dataset and preprocessing approach used in the experiment are briefly introduced. The different scenarios are given in detail. Then, the proposed model including the mechanism, structure, and hyper-parameters is presented.

4.2.1 Dataset and Preprocessing

We used the Korea University dataset [94] and the BCI Competition IV 2a [131] dataset to evaluate the proposed model performance on the two-class and four-class MI tasks classification.

1) Korea University (KU) Dataset: We used the Korea University Dataset containing 54 subjects with binary MI tasks of the left hand and right hand. Two sessions were conducted on different days in the dataset, each with 200 trials for every subject. The MI-EEG signals were collected by 62 Ag/AgCl electrodes with impedances of less than 10 $k\Omega$. To better decode MI information, 20 electrodes in the motor cortex region were

selected (C-z/1/2/3/4/5/6, CP-z/1/2/3/4/5/6, FC-1/2/3/4/5/6) according to previous studies [93, 95, 130]. The sampling rate was 1,000 Hz and we downsampled to 250 Hz.

2) BCI Competition IV 2a (BCIC-IV-2a) Dataset: The BCIC-IV-2a consists of recordings from 9 healthy subjects performing 4 different motor imagery tasks: left-hand, right-hand, both-foot, and tongue. The signals were acquired using 22 EEG electrodes with a sampling frequency of 250 Hz and were bandpass filtered between 0.5 Hz and 100 Hz, as well as notch filtered at 50 Hz. Two sessions were recorded on different days for each subject, with each session comprising 288 trials. The dataset only has 22 channels so we feed all channel signals into the proposed model.

The most common frequency band used in the MI-EEG field is α rhythm [123] which is about 10 Hz and β rhythm which is around 20 Hz [38]. The filter bands that include useful spectral MI information vary from person to person [69]. Therefore, some studies [93, 116, 130] divided the raw MI signals into several bands with a 4 Hz length ranging from 4 to 40 Hz by spectral filters. Considering the extra calculations caused by multi-inputs, we only feed three inputs including the raw signals and two filtered bands based on α (7-12 Hz) and β (13-32 Hz) rhythms. Each trial has 4 seconds with 1000 samples in total. We employed the Z-score normalization to handle the signals, as shown:

$$Z = \frac{x - \mu}{\sigma} \tag{4.1}$$

where x was the raw data of each channel. μ was the mean value of x and σ represents the standard deviation.

4.2.2 Scenarios Description

We design three scenarios of the within-subject analysis using data from the same subject for training, validation, and testing (Fig 4.1). Different scenarios help verify the models' adaptability and robustness for actual applications. The details of different scenarios are described as follows:

1) Within-Session Scenario: This scenario only uses one session with 200 trials for 10-fold cross-validation (CV). Although the training data is limited, within-session

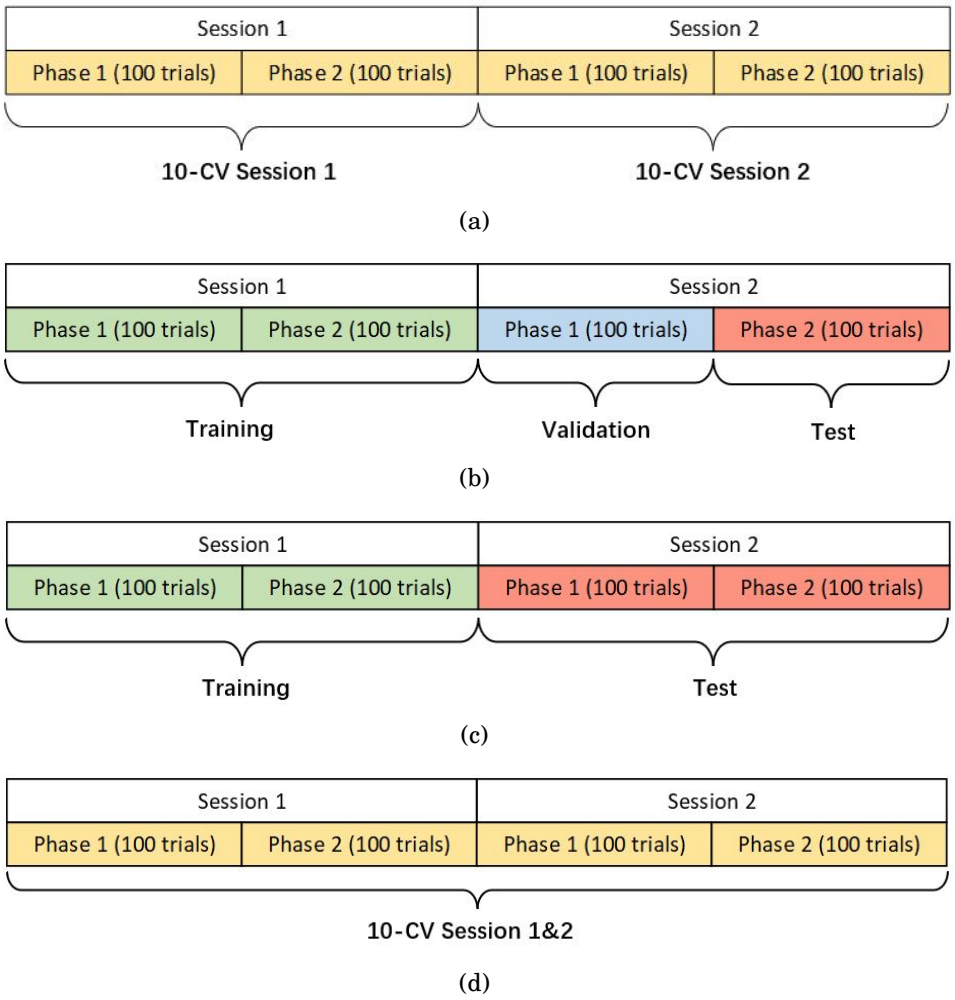


Figure 4.1: Descriptions of different scenarios (KU dataset). (a) Within-Session Scenario. (b) Cross-Session Scenario Case 1. (c) Cross-Session Scenario Case 2. (d) Two-Session Scenario.

ensures the stability of the data distribution as far as possible.

2) Cross-Session Scenario: The first session is used for training and the second one for testing. Two cases are presented in this scenario considering the different applications in reality. The first namely the hold-out scenario uses the part of the data in session two for validation and the rest for the test. The other case only uses the whole data in session two for the test, ensuring no data participates in validation at the modeling stage. In either case, the data from session two will not be used in training. Due to the circumstance that two sessions were conducted

on different days, the drift of statistical distributions brings the challenge for classification.

3) Two-Session Scenario: Two sessions of one subject are grouped for a 10-fold CV to show the performance of the models in big data.

In the BCIC-IV-2a dataset, each phase contains 144 trials because there are 288 trials for each session while one phase in the KU dataset has only 100 trials.

4.2.3 The Proposed Model

4.2.3.1 Architecture

The proposed model has three branches which were fed from filtered data and concentrated by a fully connected layer for fusing features from multi-bands. Each branch has the same structure consisting of the temporal block, spatial block, and transformer-based densenet block (T-Densenet Block) (Fig 4.2).

4.2.3.2 Temporal block

Considering that the MI-EEG signals are time series, the previous studies [85, 93, 132] preferred using a 1-D CNN filter to extract the temporal feature which is one of the most distinguished MI information. CNN filter has a strong inductive bias of weight sharing [142]. Such a characteristic reduces a huge amount of computation and makes a model more parameter-efficient, but it ignores the dynamic relationship among the input data in a kernel with the filter sliding because the weights learned by the CNN are fixed after training.

Self-attention The self-attention mechanism focuses more on the correlation between each value in the kernel and all other values. First, the transformer encoder divides the input into three representations namely Queries (Q), Keys (K), and Values (V) by the linear dense layers. Then the specific attention "Scaled Dot-Product Attention" (shown in Fig 4.3(d))) computed the dot products of the queries with all keys. The results were

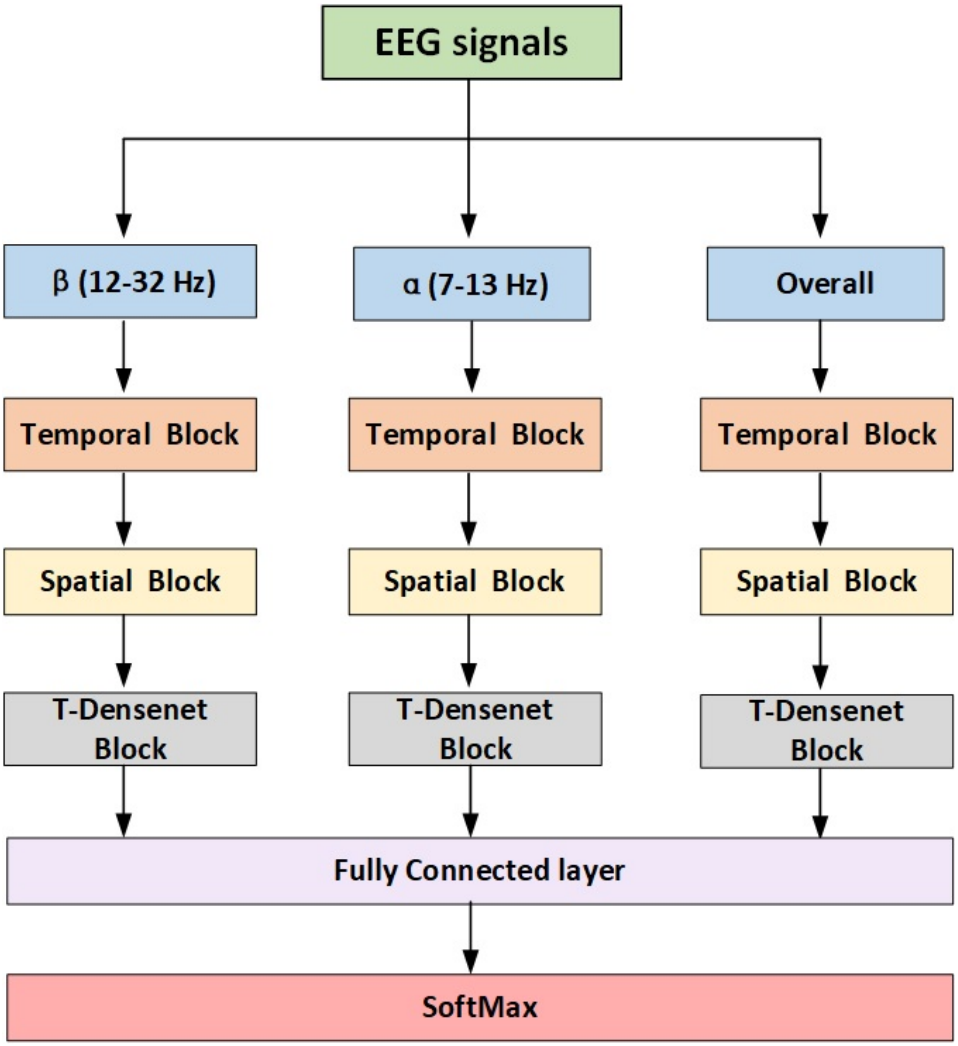


Figure 4.2: The proposed model structure

divided by $\sqrt{d_k}$ and ended with a softmax function to obtain the weights on the values. The formula is:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.2}$$

where d_k was the dimension of keys. To better jointly learn the information from different representation subspaces at different positions [111], the scaled dot-product attention

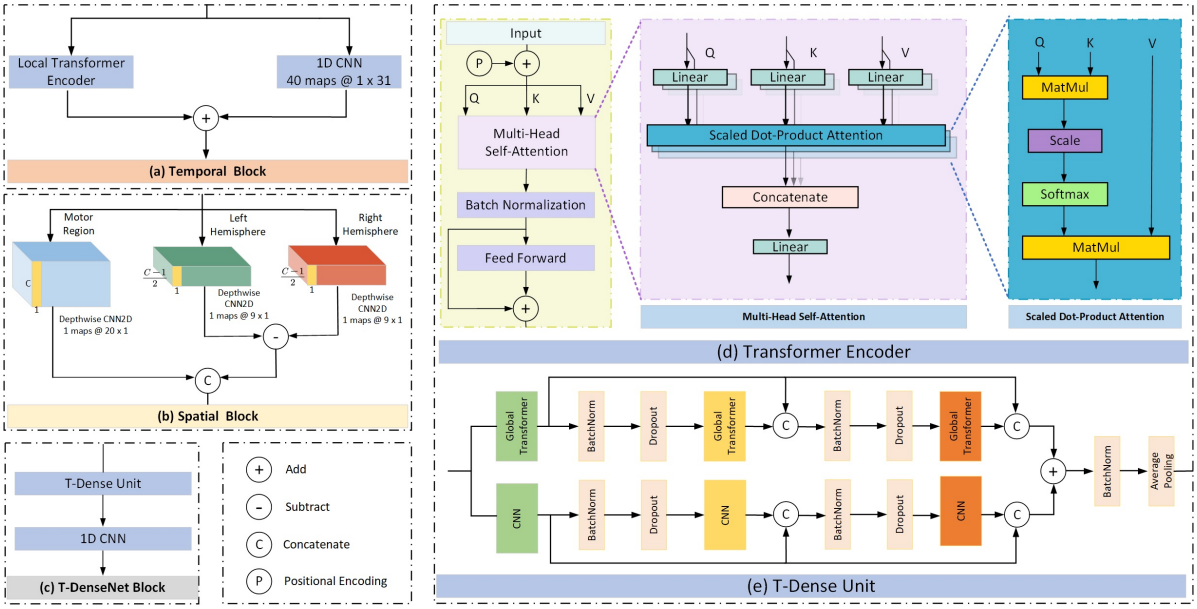


Figure 4.3: The proposed model structure. (a) Temporal Block; (b) Spatial Block; (c) T-DenseNet Block; (d) Transformer Encoder; (e) T-Dense Unit.

was embedded in the structure of the "Multi-head Self Attention" (Fig 4.3(d)):

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O$$
$$head_i = Attention(Q,K,V) \tag{4.3}$$

where $W^O \in R^{hd_v \times d_{model}}$, hd_v resprents the dimension of values and d_{model} is the dimension of the outputs. We employ $h = 2$ parallel attention layers in the proposed model. From equations (2) and (3), the calculation of the output is determined by a weighted total of the values, and the weight for each value is determined by a function that assesses the compatibility between the query and its corresponding key. Therefore, the weights are dynamic rather than fixed like CNN filters.

Local transformer encoder In this work, to take full advantage of the characteristics of the modes of CNN and transformer, we add the outputs from the local transformer encoder and those from the CNN filter together as the final temporal features (Fig 4.3(a)). Compared with the global transformer that obtains the attention score of a query based on all keys (Fig 4.4(b)), the local transformer encoder reduces the number of keys to ensure that the queries are multiplied by the limited keys every time (Fig 4.4(a)). Such

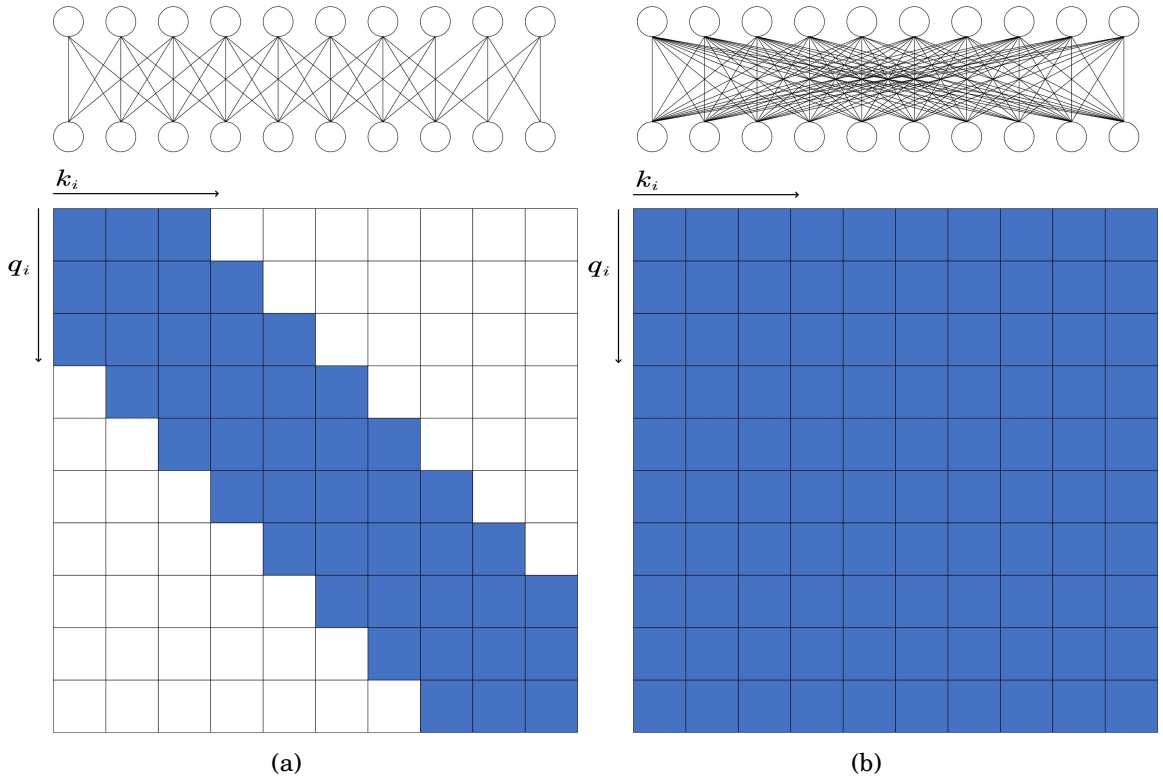


Figure 4.4: Attention patterns in the transformer. The blue squares represent corresponding attention scores are calculated and the blank ones mean the attention score is discarded. (a) Local pattern. (b) Global pattern.

a mode improves the temporal feature decoding by increasing the locality. Although the local mode cannot learn the global features, it selects local subtle features that otherwise are largely ignored in the global mode. It can further overcome the overfitting and underfitting problems for long raw EEG signals.

Positional encoding Considering that the MI-EEG signals are the sequence that has the order, the position information is injected by the sum of the Positional Encoding (PE) value and the raw signals. According to the successful PE application in the MI-EEG field [119], we used sine and cosine functions to represent the position as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{4.4}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{4.5}$$

where pos means the position and i is the dimension. d represents the dimension of the inputs.

4.2.3.3 Spatial block

Previous research has already demonstrated the feasibility of using the brain hemisphere to control both the left and right hands, but the degree of control for each hand differs due to lateralization [143, 144]. Therefore, the spatial feature differences between the two hemispheres may potentially be useful for motor imagery classification. After concatenating the temporal features learned from the CNN and the local transformer encoder, the depthwise CNNs are used to extract spatial information from the EEG channels. The proposed model sends the input into three parallel paths (Fig 4.3(b)). The first CNN filters extract the spatial features from all C channels in the motor region. The rest two CNN filters extract features from $\frac{C-1}{2}$ channels in the left hemisphere and the right hemisphere respectively. The extra channel Cz was deleted because it was set in the central position. Then the difference was obtained by subtracting the features of the two hemispheres. Finally, the spatial features based on the channels from the motor region and the difference caused by hemispheres are fed into the next block.

4.2.3.4 T-Dense block

This block comprises one T-Dense Unit and a 1-D CNN filter, as shown in Fig 4.3(c). The T-Dense Unit (Fig 4.3(e)) has two branches with the CNN filters and the global transformers (Fig 4.4(b)) respectively. Both branches in the T-dense unit has the similar structure and processing steps as shown in Fig 4.3(e). For instance, in the CNN filter branch, the features of the first CNN filter are concatenated with the ones from the second filter to feed into the third CNN filter. Each subsequent layer of CNN is not only connected to the immediate preceding layer, but also to all other preceding layers, enabling the establishment of shorter paths to help the flow and reuse of the information [126]. Meanwhile, batch normalization and dropout techniques are applied to address the overfitting issue. In the T-dense unit, the global mode is applied to the transformer

to retain its original advantages of extracting the global information using all neurons based on the SA mechanism while CNN filters work differently, as they learn the global information by sliding and pooling layers steps given the limited size of a filter. The branches in the T-Dense Unit are combined to learn a comprehensive set of features that otherwise can not be achieved using a single feature extraction mechanism. After the T-Dense Unit, the 1-D CNN layer is used to reduce the dimension thus reducing the calculation burden for the subsequent output neurons after three parallel branches.

4.2.3.5 Training Setup

The cross-entropy function is employed as a loss function which evaluates the distance between the probability distribution of the model prediction values y_p and the true labels y_t:

$$L\left(y_p, y_t\right) = -\sum_m y_{p,m} \log y_{t,m}. \tag{4.6}$$

where m is the index of y. The optimizer is Adam [137] and the learning rate is set to 0.0001. The training takes 800 epochs with 32 batches per epoch. The early stopping technology was used to save the best weights. The training step ended after checking if the validation loss value decreased for the last 100 epochs. After reaching the threshold, the model with the best weights produces the classification results of the test fold.

The computer used in this experiment had 15 Intel processors and 80 GB RAM. GTX 3090 GPU with 24 GB memory was used for training and testing MI-EEG signals. Keras based on TensorFlow was used for constructing the proposed model.

4.3 Results

4.3.1 Performance Comparison

We evaluate the proposed model and other models in the different scenarios. The average classification accuracies of all subjects of the KU dataset and BCIC-IV-2a with standard deviation (SD) are shown in Table 4.1 and Table 4.2 respectively.

In the KU dataset, our proposed model achieved the best performance in all scenarios, especially on the cross-session and two-session ones. Constrained by the limited data size of each subject, which only comprises 200 trials per session, achieving even slight improvements can be a challenge. In the within-session scenario, the proposed model achieved an accuracy of 75.94% and 77.38% in session 1 and session 2 respectively. which are 0.99% and 1.46% higher than the best public model namely tensor-CSPNet. When utilizing twice the amount of data in the two-session scenario, the proposed model achieved a classification rate exceeding 80%, 7.46% higher than the Shallow ConvNet. In two cases of cross-session scenarios, as Fig 4.1(b) and Fig 4.1(c) presented, both of them used the data from session 2 as the test and did not allow them to present in the training step. The difference was that case 1 used half of the data from session 2 to validate while case 2 did not use it. Considering the drift of data distributions caused by the different sessions conducted on different days, the results of cross-session are lower than the ones of within-session. The performances of most compared methods decrease including the proposed model. The existing high-performing models such as FBCNet and Tensor-CSPNet exhibited reduced performance to less than 70% while the proposed model only lost an average of 0.24% accuracy and still produced the best accuracy of 77.14% in case1. Given the BCI application that people often only used the data collected in one day to build the model without training or updating the following day to save patients' time, case 2 is more suitable in practical applications. The proposed model achieved 74.51%, a much higher accuracy than the benchmarks, which confirms the superiority of our proposed model on adaptability and robustness. The statistical test was also conducted to compare the performances of different models. We observed that the proposed model outperformed most baseline models ($p < 0.001$), FBCNet($p < 0.05$), and Tensor-CSPNet ($p < 0.05$) in different scenarios.

In the BCIC-IV-2a dataset, FBCNet performed best in two within-session scenarios while the proposed model showed an accuracy decrease of 3.07% and 0.33% respectively in session 1 and session 2. However, in the two cross-session scenarios, our proposed model improved significantly reaching 75.84% in case1, 1.24% higher than the Shallow ConvNet,

and 75.08% in case2, 2.12% higher than the Tensor-CSPNet. The result in the two-session scenario also reached 81.04% which improved the accuracy by 2% compared to the Shallow ConvNet. The statistical test showed that the proposed model outperformed all baseline models ($p < 0.05$) in both the cross-session and two-session scenarios.

We also checked the statistical significance of each scenario. The t-test result of within-session 1 with within-session 2 was [$correlation = 0.781, p < 0.001$; $t_{(53)} = -1.062, p = 0.293$] and [$correlation = 0.88, p < 0.01$; $t_{(53)} = -1.024, p = 0.336$] in KU and BCIC-IV-2a dataset separately, which shows that there is consistency between different sessions for each subject, but the difference between two sessions is not statistically significant. In the KU dataset, the t-tests of within-session1 with case 1 and case 2 in the cross-session scenario were [$t_{(53)} = -0.838, p = 0.406$] and [$t_{(53)} = 1.182, p = 0.242$]. The t-tests in BCIC-IV-2a were [$t_{(53)} = -0.627, p = 0.548$] and [$t_{(53)} = -0.475, p = 0.648$]. Both results of the t-test did not have statistically significant differences. Hence, the data quality of an individual varies on different days. Building a model for each day is time-consuming and impractical, but employing a cross-session model may result in a decrease in classification accuracy, making it a challenging task. The t-tests results of within-session 1 with two-session and within-session 2 with two-session are [$correlation = 0.882, p < 0.001$; $t_{(53)} = -4.514, p < 0.001$] and [$correlation = 0.901, p < 0.001$; $t_{(53)} = -3.102, p < 0.01$] separately. Evidently, an increase in the volume of data contributes to the enhancement of model performance even though the sessions were collected on different days.

In summary, for the KU dataset, the proposed model outperforms other models, achieving up to 0.99% and 1.46% for the session 1 and 2 respectively in the within-session scenario, up to 7.49% and 8.19% for cases 1 and 2 in the cross-session scenario and up to 7.46% for the two-session scenario. When testing on the BCIC-IV-2a dataset, the model can also improve the classification accuracy by 1.24% and 2.12% for cases 1 and 2 in the cross-session scenario and 2.21% for the two-session scenario, confirming the superiority of the proposed model in decoding MI-EEG information.

Table 4.1: Comparison of average classification accuracy (%) and standard deviation (SD) for different methods (KU dataset).

	Within-session		Cross-session		Two-session
	Session1 (SD)	Session2 (SD)	Case1 (SD)	Case2 (SD)	Session 1&2 (SD)
CSP [138]	56.53(13.10)	58.38(14.63)	61.70(16.14)	60.43(13.98)	55.80(11.07)
FBCSP [70]	64.41(16.28)	66.47(16.53)	59.67(14.32)	61.57(14.73)	65.62(14.75)
MDM [145]	50.47(8.63)	51.93(9.79)	52.33(6.74)	-	-
TSM [145]	54.59(8.94)	54.97(9.93)	51.65(6.11)	-	-
SPDNet [146]	57.88(8.68)	58.88(8.68)	60.41(12.13)	-	-
Shallow ConvNet [85]	67.73(17.58)	68.47(17.65)	67.79(19.16)	66.32(16.18)	72.74(15.82)
Deep ConvNet [85]	56.19(13.71)	57.38(15.27)	56.59(15.29)	56.75(13.03)	62.91(17.64)
EEGNet [132]	63.37(17.06)	64.73(17.97)	65.26(19.31)	63.28(15.69)	69.73(17.05)
FBCNet [93]	74.16(12.60)	73.81(13.99)	67.83(14.34)	-	-
Tensor-CSPNet [95]	74.95(15.27)	75.92(13.99)	69.65(14.97)	-	-
Proposed model	**75.94**(14.71)	**77.38**(15.29)	**77.14**(14.76)	**74.51**(13.93)	**80.20**(13.01)

Table 4.2: Comparison of average classification accuracy (%) and standard deviation (SD) for different methods (BCIC-IV-2a dataset).

	Within-session		Cross-session		Two-session
	Session1 (SD)	Session2 (SD)	Case1 (SD)	Case2 (SD)	Session 1&2 (SD)
CSP [138]	57.75(13.71)	60.60(14.29)	54.01(12.77)	54.07(12.13)	57.15(12.26)
FBCSP [70]	73.57(16.28)	72.46(16.53)	65.59(17.51)	65.79(14.21)	75.01(12.97)
MDM [145]	62.96(14.01)	59.49(16.63)	-	50.74(13.80)	-
TSM [145]	68.71(14.32)	63.32(12.68)	-	49.72(12.39)	-
SPDNet [146]	65.91(10.31)	61.16(10.50)	-	55.67(9.54)	-
Shallow ConvNet [85]	71.83(15.63)	72.64(19.62)	74.61(12.36)	68.96(14.28)	78.83(12.32)
EEGNet [132]	69.26(11.59)	66.93(11.31)	61.65(14.20)	60.31(10.52)	70.67(17.27)
FBCNet [93]	**77.26(14.82)**	**76.58(13.09)**	-	72.71(14.67)	-
Tensor-CSPNet [95]	75.98(14.26)	74.92(14.63)	-	72.96(14.98)	-
Proposed model	74.19(10.60)	76.25(12.67)	**75.85(14.11)**	**75.08(12.66)**	**81.04(8.54)**

4.3.2 Ablation Study

The purpose of an ablation study is to assess the impact of specific components on the overall performance of a model by removing them and analyzing their contribution. We conducted the ablation tests to evaluate the effectiveness of the transformer encoders, the hemisphere difference in the spatial block, and the T-Dense units in different scenarios. 1) The proposed model without transform encoders (w/o_Trans) removes both the local and global transformer encoders; 2) The proposed model without the hemisphere difference in the spatial block (w/o_Diff-hemi) removes the structures in Fig 4.3(b) which extract the spatial features from each hemisphere and calculate the difference. The previous models proposed in the literature were shown to achieve good classification results in the KU EEG dataset such as EEGNet [132], Shallow ConvNet [85] and FBCNet

Table 4.3: Ablation study of the proposed method on the different modules.

KU dataset					
	Within-session		Cross-session		Two-session
	Session1(SD)	Session2(SD)	Case1(SD)	Case2(SD)	Session 1&2(SD)
w/o_trans	75.81(14.22)	68.01(13..31)	75.02(14.89)	66.42(11.01)	72.88(12.63)
w/o_diff-hemi	75.20(14.92)	75.98(15.59)	73.89(16.01)	73.20(13.98)	78.89(13.60)
w/o_T-dense	67.88(12.36)	68.53(13.12)	68.54(12.88)	66.74(11.39)	73.61(12.81)
Proposed model	**75.94(14.71)**	**77.38(15.29)**	**77.14(14.76)**	**74.51(13.93)**	**80.20(13.01)**
BCIC-IV-2a dataset					
w/o_trans	**74.64(11.21)**	76.07(12.58)	73.14(13.64)	73.64(11.54)	**84.01(8.60)**
w/o_diff-hemi	72.56(11.09)	73.95(14.59)	70.67(15.15)	73.72(11.82)	78.91(10.08)
w/o_T-dense	63.31(8.43)	65.74(10.11)	68.82(13.73)	63.12(6.47)	77.21(8.66)
Proposed model	74.19(10.60)	**76.25(12.67)**	**75.85(14.11)**	**75.08(12.66)**	81.04(8.54)

[93] focus on the spatial features from all channels in the motor region and ignore the available information that might be learned from the hemispheric differences. 3) The proposed model without T-dense units (w/o_T-dense) replaces the two T-dense units with common CNN layers. The results of the ablation study in different scenarios are shown in Table 4.3. In the KU dataset, the T-test results indicated that there is no statistical significance ($p > 0.05$) to show the w/o_trans brings a negative impact on the classification accuracy in session 1 of the within-session scenario. Apart from this, the absence of any specific components will make the accuracy drop($p < 0.05$). Especially for session 2 in the within-session scenario, without the transformer encoders, the classification result decreased by 9.37%. In the BCIC-IV-2a dataset, although in within-session1 and two-session scenarios, the model without transformer encoders performed better, other cases still show the importance of the different modules. The statistical analysis showed that w/o_trans decreased the classification accuracy in the cross-session scenario ($p < 0.05$). The w/o_Diff-hemi had significance in within-session and cross-session scenarios ($p < 0.05$) while the T-test result in the two-session scenario was $p = 0.127$. The w/o_T-dense had statistical significance in all scenarios (($p < 0.01$)). The T-dense has significant contributions to the classification accuracy improvement, as it produces a comprehensive set of temporal-spatial features. Without this module, using a simple temporal block and spatial block is unable to capture sufficient and subtle useful features embedded in the highly corrupted and diffused EEG raw data. We also tested the selection of the activation functions on cross-session case 1 from the KU dataset (Fig

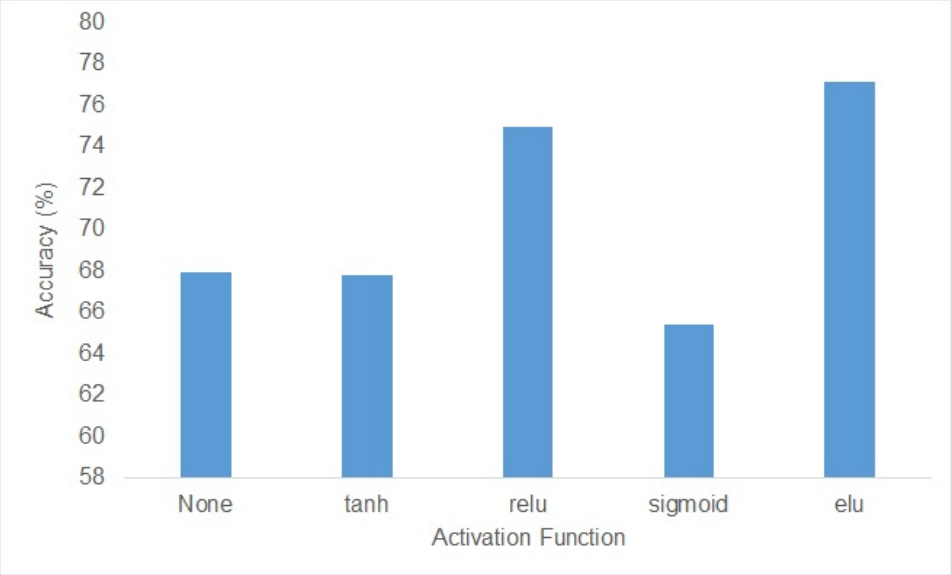


Figure 4.5: The effect of activation functions of all subjects in KU dataset

4.5). The ELU function performed best with the highest accuracy.

4.3.3 Complexity

Table 4.4 shows the model complexity based on the number of trainable parameters. The results show that there is no decisive relationship between the complexity of a model and its performance. Deep ConvNet has the most parameters because of more CNN layers used in the structure. However, regardless of the scenarios, the Deep ConvNet performs badly even worse than the traditional approach FBCSP. Among these compared models, the EEGNet only has no more than 2k trainable parameters because the depthwise separable convolution layer is employed to reduce the dimensions. However, EEGNet performs much better than the Deep ConNet in each scenario. The Tensor-CSPNet divides the raw signals into several frequency bands to learn subtle features within different frequency bands, thus encompassing the spectral differences among different subjects. This approach adds additional computational parameters but the model performance is the best as demonstrated in the previous studies. The proposed model includes 12K of trainable parameters that are only half of the Tensor-CSPNet but have better classification results, which demonstrates its efficacy and effectiveness.

Table 4.4: Model complexity based on the number of trainable parameters.

Models	Parameters
Shallow ConvNet	42884
Deep ConvNet	282004
EEGNet	1876
Tensor-CSPNet	232360
Proposed Model	118337

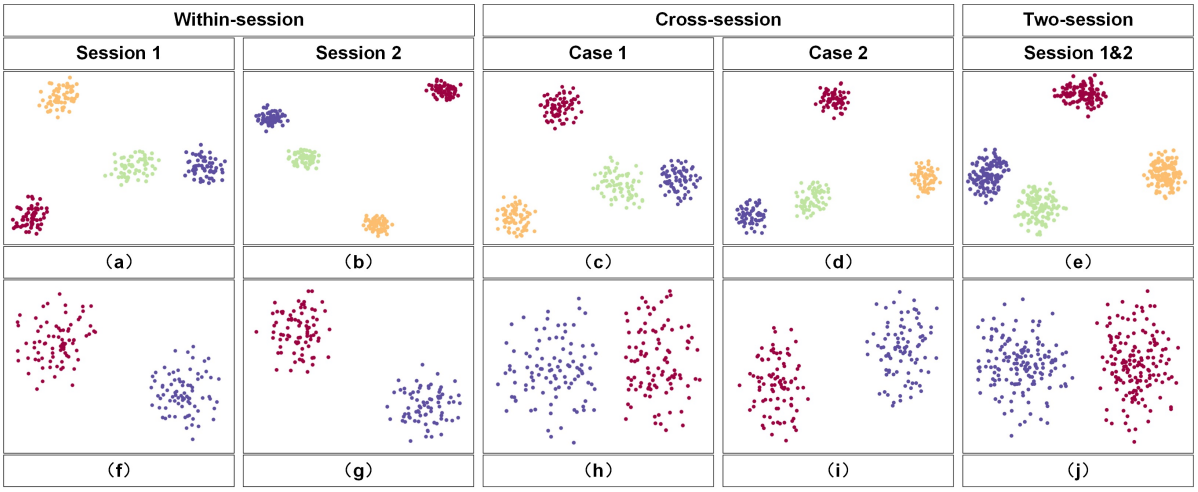


Figure 4.6: The feature map obtained by the proposed model in 2-D embedding based on t-SNE. Part (a) (e) is the distribution of the extracted features of the third subject from the BCIC-IV-2a dataset. Parts (f) (j) show the distribution of extracted features of the third subject from the KU dataset.

4.3.4 Feature Visualization

The t-distributed Stochastic Neighbor Embedding (t-SNE) approach was employed to visualize the feature distribution after the last fully connected layer of the proposed model. Fig 4.6 shows the comparison of the visualization based on the different scenarios. We used the data from subject 3 in the two datasets respectively. Fig 4.6 (a) (e) belongs to the BCIC-IV-2a while Fig 4.6 (f) (j) belongs to the KU dataset. Each color represents one label of MI-EEG tasks. According to the t-SNE result, the proposed model showed a great ability to classify EEG signals. In comparison to within-session, the feature distribution in cross-session and two-session scenarios appears to be more dispersed. However, there are still clear distinctions that can be observed, further showing the superior performance of the proposed model.

4.4 Discussions

In this chapter, we have proposed a local and global convolutional transformer-based model for MI-EEG classification. The transformer encoder with the self-attention mechanism is widely applied to the computer version and natural language processing. Compared with the CNN limited by the size of its filter, the transformer can capture all samples simultaneously, which is suitable for extracting global features. Meanwhile, the calculation step of the self-attention mechanism focuses on finding the relationship of different features while CNN extracts common mode from features. Once the CNN-based model is trained, the weights in the filters are fixed. However, in a transformer encoder, the weights depend on the inputs, so they are dynamically changed according to the data. Previous studies have shown that the EEG, as an intricate time series, varies from subject to subject which makes the transformer a suitable approach for processing EEG signals. Due to the distinct characteristics of CNN and transformer, combining and complementing each other makes for exploring more useful features of EEG signals and ensuring the robustness of the model.

In the proposed model, we employed two strategies for the transformer, specifically the local and the global modes. When extracting temporal features from raw EEG signals, such a long time series will significantly increase the cost of model computation and lead to severe overfitting problems. Using the local transformer encoder can limit the size of the filter like the learning mode of a CNN layer. Although this will cause the transformer to lose the chance of obtaining global features of long sequences at once, it can still leverage the advantage of dynamically extracting learning feature relationships, complementing the CNN. When the features are sent into the T-Dense block, the transformer encoder employs the global mode because the time series has been processed with the pooling layers. Meanwhile, the feed-forward fashion connecting each layer to every other layer in the CNN and transformer branch encourages feature reuse and information flow which improve the model performance. In the spatial block, compared with previous models the proposed model used the depthwise CNN layer to extract spatial features not only from all channels like ConvNet [85], EEGNet [132] and

FBCNet [93] which performed well in the KU and BCIC-IV-2a dataset but also from the difference of two hemispheres. The result of the ablation study has shown the efficiency of this module. After extracting features from the hemisphere differences, the proposed model got higher classification results in all scenarios, especially in the cross-session cases.

To better validate the superiority of the proposed model, we designed three scenarios including within-session, cross-session, and two-session in two famous public datasets. From Table 4.1 and Table 4.2, the results show that our proposed model achieved the highest classification result in the different scenarios. Compared with the other two scenarios, the cross-session scenario is closer to the real application which limits the model performance because of the number of data and the drift of statistical distributions. However, our proposed model still performed well and was less than only 3% than within-session results which further shows the good robustness and adaptability. Previous models based on the transformer for MI classification use the CNN layers [116, 117, 119] to extract temporal features while the transformer is used to refine features. While the proposed model adopted the local mode of the transformer to complement the functionality of CNN in time-series data analysis, rather than simply placing the transformer behind the CNN layer. Meanwhile, during the feature refinement stage, the proposed model not only employed the attention mechanism in the transformer but also combined with the DenseNet to improve the flow and reuse of information. Further, the spatial features learned from the differences in the hemispheres were also taken into consideration.

4.5 Conclusions

In this chapter, a novel and effective approach for MI-EEG classification using a local and global convolutional transformer-based model has been developed. The proposed model has been validated on the three scenarios and two public datasets. The combination of CNN filters and transformer encoders with local and global structures has the advantage

of extracting a comprehensive set of useful features from EEG signals. In the spatial module, we also consider the possible information from the differences between the hemispheres which helps improve the robustness of the model. Our results showed that the proposed model outperformed the state-of-the-art methods for MI-EEG classification on the KU dataset, achieving up to 0.99% and 1.46% for the session 1 and 2 respectively in the within-session scenario, up to 7.49% and 8.19% for the case 1 and 2 respectively in the cross-session scenario and up to 7.46% for the two-session scenario. For the BCIC-IV-2a dataset, the model can also improve the classification accuracy by 1.24% and 2.12% for cases 1 and 2 in the cross-session scenario and 2.21% for the two-session scenario.

Chapters 3 and 4 have explored the superior ability of deep learning in MI-EEG decoding. However, deep learning models require a large amount of data for training, which limits the practicality of BCI and cannot meet the demand for plug-and-play. Therefore, the next Chapter will discuss how to improve the generalization performance of models, so that the models can achieve excellent classification performance without the need of training on new data.

CHAPTER 5

CROSS-SUBJECT MI-EEG DECODING WITH DOMAIN GENERALIZATION

Decoding motor imagery (MI) electroencephalogram (EEG) signals in brain-computer interface (BCI) can assist patients in accelerating motor function recovery. To realize the implementation of plug-and-play functionality for MI-BCI applications, cross-subject models are employed to alleviate the time-consuming calibration and avoid additional model training for target subjects by utilizing the EEG data from source subjects. However, the diversity in data distribution among subjects limits the model robustness. In this study, we investigate a cross-subject MI-EEG decoding model with domain generalization based on a deep learning neural network that extracts the domain-invariant features from source subjects. Firstly, the knowledge distillation framework is adopted to obtain the internally invariant representations based on spectral features fusion. Then the correlation alignment approach aligns the mutually invariant representations between each pair of sub-source domains. In addition, we use distance regularization on two kinds of invariant features to enhance generalizable information. To assess the effectiveness of our approach, experiments are conducted on the BCI Competition IV 2a and the Korean University dataset. The results demonstrate that the proposed model

achieves 8.93% and 7.18% accuracy improvements on two datasets respectively compared with current state-of-the-art models. The results confirmed that the proposed approach can effectively extract invariant features from source subjects and generalize to the unseen target distribution, hence paving the way for effective implementation of the plug-and-play functionality in MI-BCI applications.

5.1 Introduction

DL-based models have shown excellent performance on MI-EEG decoding. However, DL applications are usually limited by the long training time, high resource consumption, and a heavy reliance on the number of labeled data [96]. In practical BCI applications, it is a challenge to collect sufficient data with good quality to build individualized models for each person. Meanwhile, achieving the goal of immediate usability with DL approaches is hard for patients because the models require a significant amount of time for training to achieve a high classification accuracy. Therefore, there is a strong desire to recognize patients' MI intentions without additional experimental data collection and modeling.

Domain generalization (DG) approaches only consider the data from the source domains and develop models that can generalize to unfamiliar distributions. Given the limited real training data, a simple way to enhance the generalization capability is to create more manual data. For instance, Tobin et al [147] added domain randomization for generalization in the real environment by changing the number, shape, texture and other characteristics of the objects. Zhang et al [148] proposed a data generation-based DG method namely Mixup to generate new training samples by linearly blending the features and labels of different data. Another group of methods is representation learning which adopts kernels, adversarial training or feature alignments to learn domain invariant representations [149]. Grubinger et al [150] employed transfer component analysis (TCA) [151] to learn a common subspace by reducing the disparities among domains. The approaches like domain-invariant component analysis (DICA) [152] and scatter component analysis (SCA) [153] are also classical kernel-based methods similar to the

idea of TCA. Li et al [154] extract the domain-invariant features through adversarial losses that consider the source-domain label information. In the BCI field, conventional data augmentation-based DG techniques including sliding windows [85], adding noise, over-sampling [155] and geometric transformation [156] have shown improvement in the classification accuracy. However, the inter and intra-subject variability constrain the models' generalization capacity so that previous studies primarily focused on constructing within-subject models and not fully harnessing cross-subject data within the source domain [157]. Therefore, DG-based models remain largely unexplored and have not yet reached the capability to provide a calibration-free BCI solution for real-world applications. Wang et al [158] utilized knowledge distillation to extract the invariant features from pictures in the computer vision field. Inspired by its framework, we apply knowledge distillation in our work to extract the cross-domain representations in MI-EEG signals.

In this chapter, we propose a cross-subject model with a DG approach. The dataset is divided into a source domain consisting of several subdomains and a target domain. The data in the target domain with the unseen distributions will not be involved in the model's training and validation. The proposed model improves the domain generalization ability by extracting the internally and mutually invariant features among different subjects. A knowledge distillation framework is employed to capture the spectral information of EEG signals as the internally invariant representations. For mutually invariant features, the correlation alignment (CORAL) [159] method is used to align the feature distributions between any two subdomains from the source data. To reduce the possible redundancy between the internal and mutual features, the proposed model utilizes a regularization technique to enhance their dissimilarity. In the model training phase, the early stopping (ES) technology and the two-stage training strategy are used to prevent model overfitting and fully utilize all source domain data. We conduct comprehensive experiments on two MI-EEG datasets to prove the excellent generalization capability of the proposed model.

The remainder of the chapter is outlined as follows. The data description, preprocessing steps and detailed model structure are presented in Section 5.2. The experiments

and results are detailed in Section 5.3. Then, the discussion is presented in Section 5.4. Finally, Section 5.5 concludes the chapter.

5.2 Methods

5.2.1 Definitions

In the domain generalization, X denotes an input space of EEG signals and Y is an output space. The domain is defined as $S = \{(x_i, y_i)\}_{i=1}^{n} \sim P_{XY}$, where P_{XY} denotes the joint distribution and $x \in X, y \in Y$. The source domain with labeled data is divided into multiple training subdomains, namely $S_{train} = \{S^i \mid i = 1, ..., N\}$, where N is the number of subdomains and the $S^i = \left\{\left(x_j^i, y_j^i\right)_{j=1}^{n_i}\right\}$ represents the i^{th} subdomain. In the real scenario of MI-EEG classification, the internal and external diversities among subjects make the joint distributions between each pair of sub-source domains different: $P_{XY}^i \neq P_{XY}^j, 1 \leqslant i \neq j \leqslant M$. According to [149], the domain generalization aims to acquire a resilient and broadly applicable predictive function $f : X \longrightarrow Y$ from the N subdomains to minimize errors when applied to an unseen test domain $S_{test}(i.e., P_{XY}^{test} \neq P_{XY}^i \; for \; i \in \{1, ..., N\})$:

$$\min_f E_{(x,y)\in S_{test}} [loss(f(x), y)] \tag{5.1}$$

where E is the expectation and $loss$ is the loss function. Differing from domain adaptation methods, data from S_{test} will not be involved in the training and validation processes.

5.2.2 Framework

The EEG dataset consists of the source domain and the target domain. The source domain is divided into multiple subdomains sent into the proposed model as shown in Fig 5.1. Then the internally and mutually invariant representations are captured through a feature extractor. To differentiate these two kinds of information, a regularization technique is adopted by maximizing the divergence. In the end, the invariant features are concatenated together for classification.

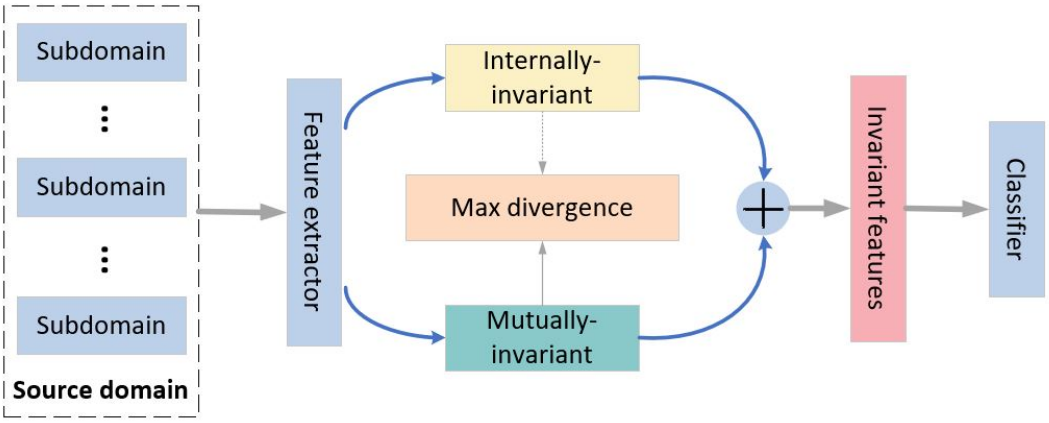


Figure 5.1: The framework of the proposed model.

5.2.3 Internally-invariant Features

Previous studies[123][38] have revealed that the most frequently utilized frequency bands in MI-EEG research are the α rhythm, typically around 10 Hz, and the β rhythm, typically around 20 Hz. In the study [124][125], the θ rhythm with a range of 4 to 7 Hz was incorporated and demonstrated its utility in decoding MI-EEG signals. Although the appropriate operational frequency bands vary from person to person [69], the information utilized for conducting the imagination classification task is primarily concentrated within these sub-bands. Hence, the spectral features based on multi-band EEG signals are employed as the internally-invariant representations in the source domain. Knowledge distillation is a straightforward framework for promoting specific characteristics within different networks [158]. The distillation framework consists of the teacher and the student network (Fig 5.2). The teacher network fuses the spectral information for MI classification and guides the student network to learn the invariant information. The structure of the teacher network, composed of three components is shown in Fig 5.3.

5.2.3.1 Spectral feature fusion

We select three sub-bands, which are θ (4-7 Hz), α (7-13 Hz), β (13-32 Hz) and the overall band as the inputs sent into the teacher model. The study [160] proved the robustness of spectral representation for MI tasks can be enhanced by adopting cross-frequency

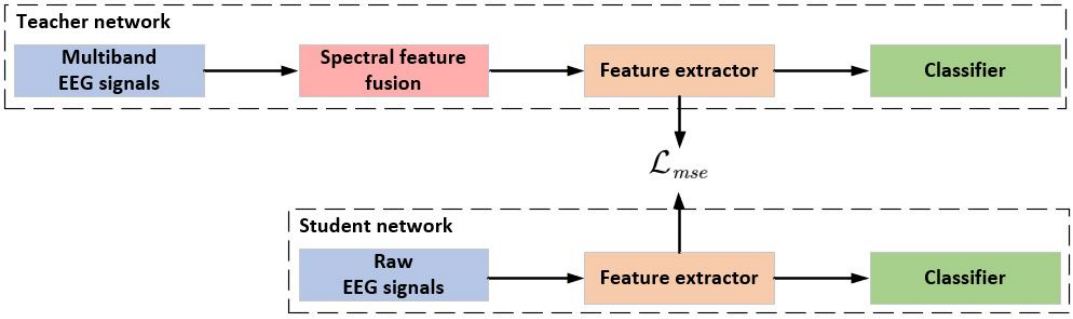


Figure 5.2: The framework of distillation to learn internally-invariant features.

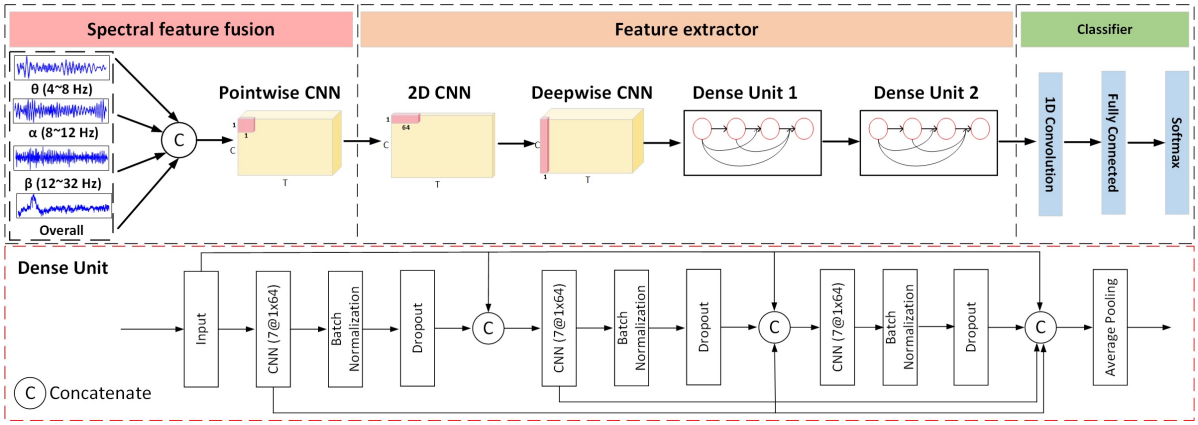


Figure 5.3: The model structure of the teacher network.

interactions. Therefore we concatenated the filtered data in the feature dimension to associate multiple frequency neural oscillations. The i_th single-trial EEG sample is defined as $X_i \in R^{C\times T}$, where C represents channels and T represents timepoints. The fused multi-band EEG data X_{MB}:

$$X_{MB} = X \times h(n) \in R^{N_b \times C \times T} \tag{5.2}$$

where $h(n)$ denotes the 3-order Butterworth filter corresponding to the n_{th} frequency sub-band and N_b is the number of sub-band. The pointwise CNN, subsequently utilized, performs convolution on each time point and channel of the EEG data. The output dimension is set to one so that the complementary information available in each frequency band is fused. Additionally, it assigns an adaptive weight to each frequency band, reducing noise in redundant frequency bands while enhancing valuable information in other frequency bands.

5.2.3.2 Feature extractor

Following the fusion of spectral features, we utilize two convolution layers to learn discriminative temporal-spatial information [85, 93, 132]. The first CNN layer using a $1 \times k_t$ kernel is employed for the EEG channel to extract temporal features. The value of k_t is equal to a fourth of the data sampling rate, enabling the capture of frequency information at $4Hz$ and beyond [132]. Then, we use a $k_s \times 1$ depthwise CNN to extract spatial features across all selected EEG channels. The kernel size k_s is configured to match the number of channels, allowing the compression of data collected at each time step into a single feature map. This strategy leads to a decrease in model parameters and enhances efficiency.

To further extract useful information from the temporal-spatial features, two dense units consisting of several CNN and pooling layers are applied subsequently (Fig 5.3). Suppose that the network comprises a total of L layers, with each layer utilizing a non-linear function $F_l(\cdot)$, where l represents the layer index, and the output of each layer is denoted as x_l. A common transformation involving a single path between each operation layer is:

$$x_l = F_l(x_{l-1}) \tag{5.3}$$

As the network becomes deeper and wider, parts of useful features are filtered. Additionally, an abundance of training parameters can result in significant overfitting issues, particularly when dealing with MI-EEG signals that contain a considerable amount of noise and redundant information. To tackle this issue, we establish short connections from any given layer to all subsequent ones. For instance, the lth layer obtains the feature maps from all preceding layers:

$$x_l = F_l([x_0, x_1, ..., x_{l-1}]) \tag{5.4}$$

where $[x_0, x_1, ..., x_{l-1}]$ are the feature maps from the preceding CNN layers $[l0, l1, ..., l-1]$. In the case where a CNN layer updates k feature matrices, the l-th layer encompasses a total of $k_0 + k \times (l-1)$ inputs. k_0 represents the raw dimension in the input layer, while k signifies the growth rate, indicating the extent to which further knowledge is acquired

and transmitted to the subsequent layer. As a result, the connections from preceding layers substantially increase to $\frac{L(L+1)}{2}$, in contrast to the traditional transition with only L connections in a network comprising L layers. Every layer obtains access to all the feature maps from preceding layers, facilitating improved information propagation and utilization of features. The ELU function is adopted as activation to reduce gradient explosion and increase model robustness. The following Batch normalization and dropout techniques help to reduce overfitting risks.

5.2.3.3 Classifier

The classifier includes a 1-D CNN, a fully connected layer and a dense layer with the softmax function for classifying MI tasks. The fused multi-band EEG signals $\tilde{x} \in X_{MB}$ and the corresponding label y are sent to the teacher network for training:

$$\min_{\theta_T^f, \theta_T^c} E_{(\tilde{x},y)\sim P^{tr}} \mathscr{L}_{cls}\left(G_T^c\left(G_T^f(\tilde{x})\right), y\right) \tag{5.5}$$

where θ_T^f and θ_T^c are the parameters of feature extractor G_T^f and the classifier G_T^c in the teacher network. E is the expectation while P^{tr} represents the data distribution in the source domain. The loss function $\mathscr{L}_{cls}$ is the cross-entropy loss, which quantifies the difference between the probability distribution of the model predictions represented as y_p and the real labels denoted as y_t:

$$\mathscr{L}_{cls}\left(y_p, y_t\right) = -\sum_m y_{p,m} \log y_{t,m}. \tag{5.6}$$

where m is the index of y. After training and optimizing the teacher network, we use the obtained features from the teacher network to guide the student network to learn the spectral invariant representations:

$$\begin{aligned} \min_{\theta_S^f, \theta_S^c} & E_{(\tilde{x},y)\sim P^{tr}} \mathscr{L}_{cls}\left(G_S^c\left(G_S^f(x)\right), y\right) \\ & + \lambda_1 \mathscr{L}_{mse}\left(G_S^f(x), G_T^f(\tilde{x})\right) \end{aligned} \tag{5.7}$$

where θ_S^f and θ_S^c are the parameters of feature extractor G_S^f and the classifier G_S^c in the student network. λ_1 is an adjustable hyperparameter to limit the Mean Squared Error

Table 5.1: The detailed architecture of the teacher network.

Block	Layer	# filters	size	Output	Activation	Options
Spectral features fusion	Input			(1, C, T)		
	Concatenate (filtered)			(N, C, T)		
	Pointwise Conv2D	1	(1, 1)	(1, C, T)	Linear	
Feature extractor	Conv2D	F1	(1, C1)	(F1, C, T)	Linear	padding = same
	Batch Normalization					
	Depthwise Conv2D	D * F1	(C, 1)	(F1, 1, T)	ELU	padding = same, depth = D
	Batch Normalization					
(Dense Unit 1)	Conv2D	F2	(1, C2)	(F1 + F2, 1, T)	ELU	padding = same
	Batch Normalization					
	Dropout					p = 0.5
	Conv2D	F2	(1, C2)	(F1 + 2 * F2 , 1, T)	ELU	padding = same
	Batch Normalization					
	Dropout					p = 0.5
	Conv2D	F2	(1, C2)	(F1 + 3 * F2, 1, T)	ELU	padding = same
	Batch Normalization					
	Dropout					p = 0.5
	Average Pooling		(1, 5)	(F1 + 3 * F2, 1, T // 5)		
(Dense Unit 2)		F2	(1, C3)	(F1 + 6 * F2, 1, T // 25)		
Classifier	Conv 1D	F3	(1, 1)	(F3, 1, T // 25)	ELU	
	Flatten					
	Dense	N*(F3 * T // 25)		N	Softmax	max norm = 0.25

(MSE) $\mathscr{L}_{mse}$ which brings the features of the student network into proximity with those of the teacher network:

$$\mathscr{L}_{mse} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{5.8}$$

where n is the index of y. Full details of the network structure are presented in Table 5.1. The parameters used in the Dense unit 1 are the same with the unit 2, hence specific details are not displayed. The difference between the student and teacher networks lies in the absence of spectral features fusion in the student network. Additionally, in the classifier block, the parameter F3 is twice the size of the one in the teacher network, in order to encompass two types of invariant features simultaneously.

5.2.4 Mutually-invariant Features

The student network learns the invariant spectral features from the teacher network by the knowledge distillation framework to classify MI tasks. However, it disregards the discrepancies in data distribution among subdomains, which means that internally invariant features alone are insufficient to guarantee excellent generalization capability. To learn the invariant representations from the source domain, the correlation alignment approach is employed to align the second-order statistics of the features from any two

domains:

$$\mathscr{L}_{align} = \frac{2}{N \times (N-1)} \sum_{i \neq j}^{N} \left\| C_i - C_j \right\|_F^2 \tag{5.9}$$

$$C_i = \frac{1}{n_i - 1} \left({X_i}^T X_i - \frac{1}{n_i} \left(1^T X_i \right)^T \left(1^T X_i \right) \right) \tag{5.10}$$

where C_i represents the covariance matrix. The internally-invariant features primarily highlight spectral information for MI task classification, while the mutually-invariant features center on cross-domain representations. To better represent these two kinds of features, the outputs of the 1-D CNN layer in the student network are divided into the internally-invariant features z_1 and mutually-invariant features z_2. Before feeding to the final classification layer, we expect to reduce the redundant information and make it more diversified between z_1 and z_2. Thus, we use the regularization tool to maximize their divergence:

$$\mathscr{L}_{div}(z_1, z_2) = -d(z_1, z_2) \tag{5.11}$$

where $d(.)$ denotes the $L2$ distance: $\mathscr{L}_{div} = -\left\| z_1 - z_2 \right\|_2^2$. In summary, the aim of the student network is established as:

$$\begin{aligned} \min_{\theta_S^f, \theta_S^c} \; & E_{(\tilde{x},y)\sim P^{tr}} \mathscr{L}_{cls} \left(G_S^c \left(G_S^f(x) \right), y \right) \\ & + \lambda_1 \mathscr{L}_{mse} \left(z_1, G_T^f(\tilde{x}) \right) \\ & + \lambda_2 \mathscr{L}_{align} + \lambda_3 \mathscr{L}_{div}(z_1, z_2) \end{aligned} \tag{5.12}$$

where λ_1, λ_2 and λ_3 are hyperparameters to limit the contribution of each loss function, which were manually tuned and finalized based on a series of experiments.

5.3 Results

5.3.1 Datasets

5.3.1.1 Dataset I

The BCI Competition IV 2a (BCIC-IV-2a) dataset, as described in [131], consists of 9 healthy subjects with 4 distinct MI tasks: left-hand, right-hand, both-foot, and tongue

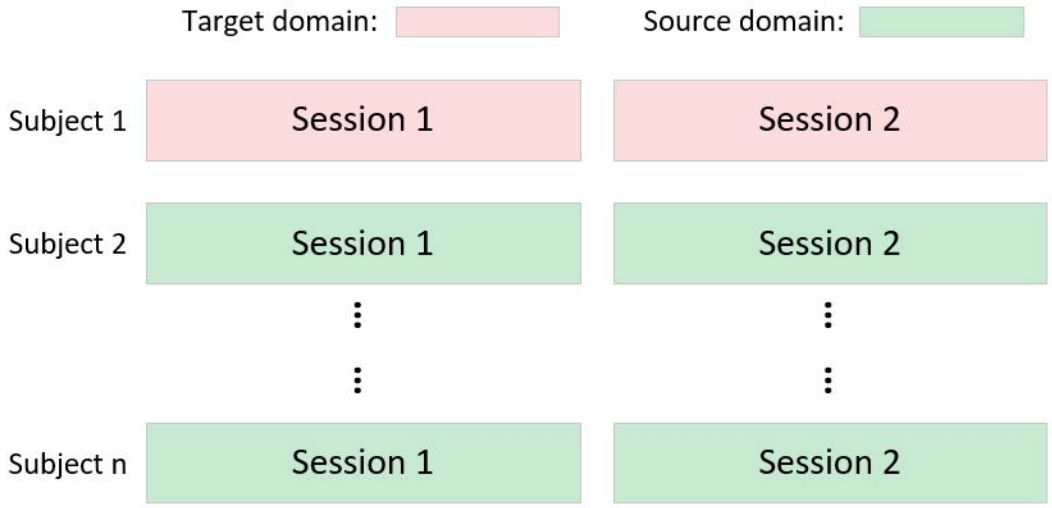


Figure 5.4: The experimental settings of the "leaving one subject out" strategy.

movements. The EEG data was captured using 22 EEG electrodes at a sampling rate of 250 Hz. Then signals underwent bandpass filtering within the range of 0.5 Hz to 100 Hz, along with notch filtering at 50 Hz. Each subject participated in two separate recording sessions on different days, and each session consisted of 288 trials. All sessions are categorized within either the source domain or the target domain.

5.3.1.2 Dataset II

The Korean University (KU) dataset [94] is one of the largest MI datasets, comprising EEG signals from fifty-four healthy subjects. Every subject engaged in 200 trials, with 100 trials dedicated to the left-hand MI task and another 100 to the right-hand MI task. EEG signals were captured from 62 EEG electrodes and initially sampled at a rate of 1,000 Hz. To facilitate equitable comparisons with other techniques, we resampled the raw signals to 250 Hz. Subsequently, 20 channels are chosen from the region associated with motor function based on the previous study [93].

5.3.2 Training Procedure

The "leaving one subject out" (LOSO) strategy (Fig 5.4) is used in our experiment. One subject is selected as the test set in the target domain. The remaining subjects are sent

into the source domain. The subjects in the source domain are divided into k groups, with each group serving as a sub-source domain. To take full advantage of all the data in the source domain, we employ a two-stage training strategy according to [93] and the early-stopping (ES) technique. First, all data will be divided into two parts, with 80% designated for training and 20% for validation. The 5-fold cross-validation is employed in the first training stage. The ES technique regards the validation loss as the criterion and monitors every epoch. Training is terminated when the loss of the validation set does not decrease within a specified number of ES epochs or the number of training epochs exceeds the predefined threshold value. Once the model with the highest validation accuracy is built, the corresponding validation loss is also recorded. Then, to involve all the source domain data in the training process, the model built in the first stage is trained again using both the training and validation set. The validation loss is monitored by the ES. If it falls below the previously recorded loss in stage one, the training will stop. In order to ensure the model's convergence, a maximum limit of 1000 training epochs is imposed for stage one, and 400 for stage two. The Adam optimizer is adopted. In the first stage, the learning rate is configured as 0.001. In the second stage, if the number of epochs is less than 150, the learning rate remains at 0.001. However, if the number of epochs exceeds 150, the learning rate is adapted to 1×10^{-4}.

The computer system utilized in this experiment was equipped with 22 AMD processors and 90 GB of RAM. For training and testing EEG data, a GTX 4090 GPU with 24 GB of memory was employed. The proposed model and baseline models were constructed using PyTorch based on Python 3.8.

5.3.3 Baseline Models

The proposed model is compared with the following benchmarks: traditional machine learning approaches (CSP [138] and FBCSP [70]), the CNNs-based approaches (Shallow ConvNet [132], EEGNet [132] and FBCNet [93])

Table 5.2: Comparison of average classification accuracy (%) and standard deviation (Std) on BCIC-IV-2a dataset.

Subject	CSP	FBCSP	Shallow ConvNet	EEGNet	FBCNet	Proposed Model
1	32.36	42.5	70.78	54.83	49.55	74.65
2	25.8	26.27	37.73	30.94	31.02	44.96
3	35.82	51.49	64.65	60.38	58.68	64.06
4	33.23	31.88	47.97	38.87	41.41	51.73
5	24.91	26.51	29.25	28.8	28.3	52.95
6	26.15	27.01	33.82	26.64	32.17	44.44
7	28.96	23.65	44.58	32.03	28.58	69.27
8	49.53	51.37	70.78	63.29	51.25	74.3
9	32.03	38.35	60.68	54.96	50.49	64.23
Avg	32.09**	35.45**	51.14*	43.42**	41.27**	**60.07**
Std	7.55	10.93	16.04	14.78	11.58	11.86

The * and ** denote the statistically significance between the classification results of the proposed model and the baseline models with * : $p < 0.05$ and ** : $p < 0.01$

Table 5.3: Comparison of average classification accuracy (%) and standard deviation (Std) on KU dataset.

	CSP	FBCSP	Shallow ConvNet	EEGNet	FBCNet	Proposed Model
Avg	56.08**	65.19**	74.62**	72.23**	71.54**	**81.80**
Std	6.82	13.04	12.15	13.93	14.07	10.70

The * and ** denote the statistical significance between the classification results of the proposed model and the baseline models with * : $p < 0.05$ and ** : $p < 0.01$

5.3.3.1 Machine learning approaches

CSP and FBCSP are the most commonly used benchmark models in the traditional machine learning domain. CSP determines the optimal spatial filters by diagonalizing a matrix for data mapping. Building upon the effective extraction of spatial features, FBCSP can mitigate the influence of subject-specific variations in frequency bands by identifying discriminative pairs of them. As described in [70], EEG signals are decomposed into nine frequency bands, each spanning a 4 Hz range from 4 to 40 Hz, utilizing Chebyshev filters in the FBCSP model. For classification, the support vector machine (SVM) with the default radial bias function (RBF) kernel is employed.

5.3.3.2 CNNs-based approaches

Shallow ConvNet first used the CNN layers to extract the temporal-spatial features from the EEG signals. The log, square and pooling operations are adopted to deal with features. Based on this shallow structure, EEGNet utilizes a separable CNN layer to refine temporal-spatial features, making it suitable for classification tasks across various EEG data while ensuring the quality of the classification. FBCNet referred to the core idea of the FBCSP, dividing the EEG signals into nine sub-frequency bands ranging from 4 to 40 Hz. Each subband is fed into the model to capture spatial features. A variance layer is employed with a fully connected layer following to unite features. All three models exhibit excellent performance and robustness on within-subject and cross-subject scenarios of primary MI-EEG datasets.

5.3.4 Experimental Results

The averaged classification accuracy of different methods is shown in Table 5.2 and Table 5.3. The statistical significance tests between benchmarks and the proposed method were conducted. For BCIC-IV-2a dataset, the results obtained by various methods are as follows: 35.09% ($p < 0.01$) for CSP, 35.45% ($p < 0.01$) for FBCSP, 51.14% ($p < 0.05$) for Shallow ConvNet, 43.42% ($p < 0.01$) for EEGNet, 41.27% ($p < 0.01$) for FBCNet, and 60.07% for our proposed model. The proposed method surpasses the best benchmark result by 8.93%. In the KU dataset, the results achieved by different methods are as follows: 56.08% ($p < 0.01$) for CSP, 65.19% ($p < 0.01$) for FBCSP, 74.62% ($p < 0.01$) for Shallow ConvNet, 72.23% ($p < 0.01$) for EEGNet, 71.54% ($p < 0.01$) for FBCNet, and 81.80% ($p < 0.01$) for our proposed model. The proposed method outperforms the best benchmark result by 7.18%. The results tested on two datasets demonstrate our proposed model effectively decodes EEG signals and extracts useful cross-domain information from source data. The trained model successfully achieved excellent classification results in the unseen target domain.

Table 5.4: Ablation study of the proposed model. Comparison of average classification accuracy (%) and standard deviation (SD) of BCIC-IV-2a and KU dataset.

	BCIC-IV-2a (SD)	KU (SD)
w./o Inter	54.61 (10.31)	81.00 (11.12)
w./o Mutual	57.50 (12.28)	80.52 (11.09)
w./o Div	56.19 (12.61)	75.85 (9.34)
w./o General	55.12 (12.00)	79.32 (10.56)
Proposed model	**60.07 (11.86)**	**81.80 (10.70)**

5.3.5 Ablation Study

The proposed model used the knowledge distillation framework and feature alignment method to capture the internally and mutually variant representations. A regularization technique was adopted to separate two kinds of features. To validate the contributions of each component, the ablation experiment was conducted by controlling the losses $\mathscr{L}_{mse}$, $\mathscr{L}_{align}$, and $\mathscr{L}_{div}$ in equation (12). The classification results based on the proposed model without internally-invariant features (w./o Inter), without mutually-invariant features (w./o Mutual), without the divergence maximum between two invariant features (w./o Div) and without the whole generalization improvement part (w./o General) are shown in Table 5.4. Any missing component will indeed lead to a decrease in the accuracy of the proposed model. Among them, the performance of w./o Div drops more significantly than other cases in both datasets, indicating the necessity to maximize the divergence of two invariant features.

5.3.6 Parameter Sensitivity

The number of subdomains is an alterable hyperparameter. We randomly divided all subjects from the source domain into several groups, ensuring that the number of subjects within each group was similar. Each group serves as a separate subdomain. The BCIC-IV-2a only has 9 subjects hence we divided the source domain including 8 subjects into 8 subdomains. The source domain of the KU dataset has 53 subjects so we split them into k groups and tested the influence of the number of the subdomains. As shown in Fig 5.5, the averaged accuracies of the proposed model remain stable with different numbers of

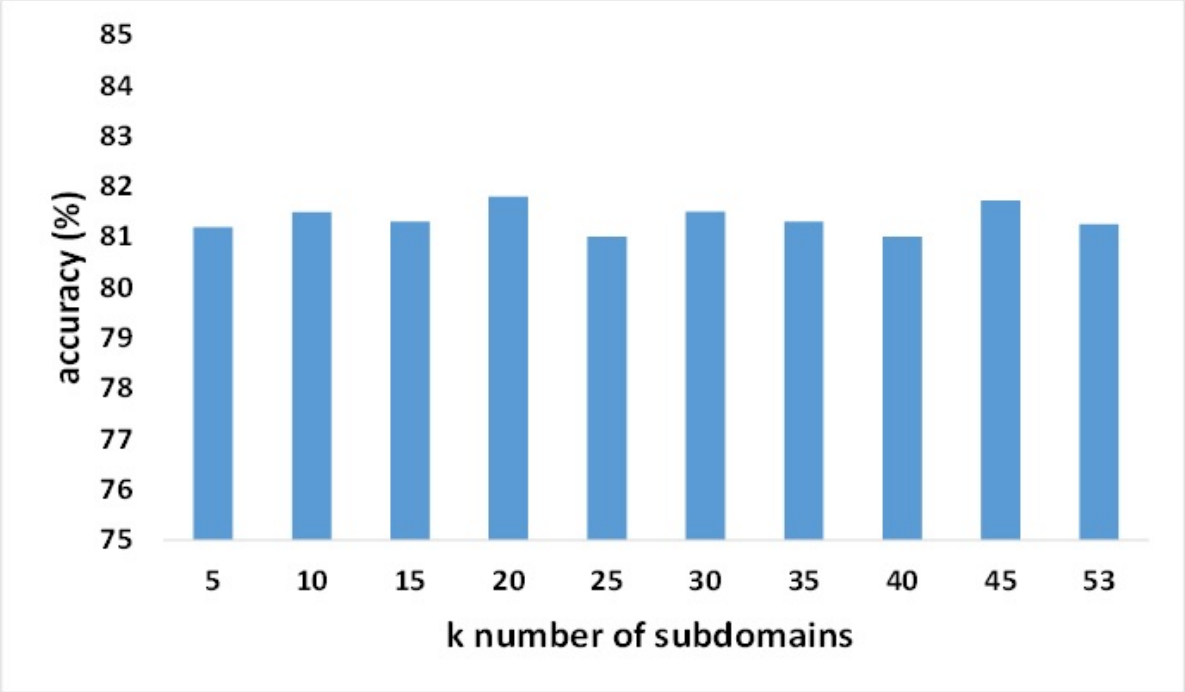


Figure 5.5: Parameter sensitivity of the number of subdomains (KU dataset).

subdomains. We chose $k = 20$ in the KU dataset to get the best performance.

5.3.7 Visualization

To better show the classification performance of the proposed model, we utilized the t-distributed Stochastic Neighbor Embedding (t-SNE) tool to visualize the feature distribution of different parts in the student network of the proposed model. We used the data from subject 8 in the BCIC-IV-2a as the target subject while the other 8 subjects were the source subjects. Fig 5.6 demonstrates the excellent generalization capability in decoding cross-subject MI-EEG signals, without requiring access to unseen target data during the training process. Because the proposed model utilized the feature alignment method to acquire cross-domain knowledge, we also assessed the model's feature aggregation performance in Fig 5.7. The t-SNE visualization in Fig 5.7(a) shows that different subdomains have different data distributions. Fig 5.7(b) is obtained before the fully connected layer in the classifier part of the student network, clearly demonstrating that the proposed model captures the invariant features of cross-domains and reduces

the differences between cross-subjects. Black dashed lines divided the feature maps into four parts corresponding to four MI tasks in the BCIC-IV-2a dataset. In each part, features from different subdomains of the same label are effectively aggregated together which further shows the superior classification and generalization ability of the proposed model.

5.4 Discussions

In this work, we have proposed a cross-subject model with domain generalization for MI-EEG classification. To get excellent decoding performance for each subject, the within-subject model is built with adequate samples from the same subject. However, the high time-consuming calibration and data collection in the within-subject model training procedure limits the implementation of plug-and-play functionality for MI-BCI applications. Therefore, it is necessary to construct a cross-subject model using previously collected data namely source domain data for classifying the target subject MI tasks without the need to collect target data. However, the variability in data distributions among different subjects within the source domain can lead to a decrease in the classification accuracy of cross-subject models. Previous studies [161–163] adopted the approach of collecting a small portion of data exclusively from the target subjects and using adaptive methods based on models trained on the source domain to improve the performance of cross-subject models. However, this DA-based approach still necessitates conducting additional experiments to acquire electroencephalogram (EEG) data from the target subjects, essentially leading to the creation of a new model for each new subject. DG-based approaches train a generalized model through training on multiple datasets in the source domain, enabling it to exhibit strong performance on an unseen domain.

In the proposed model, we employed a domain-invariant feature learning strategy to learn representations that maintain invariance across domains. The invariant features consist of two sides namely internally and mutually sides. The internally invariant features allow the model to focus on the spectral features corresponding to the MI tasks.

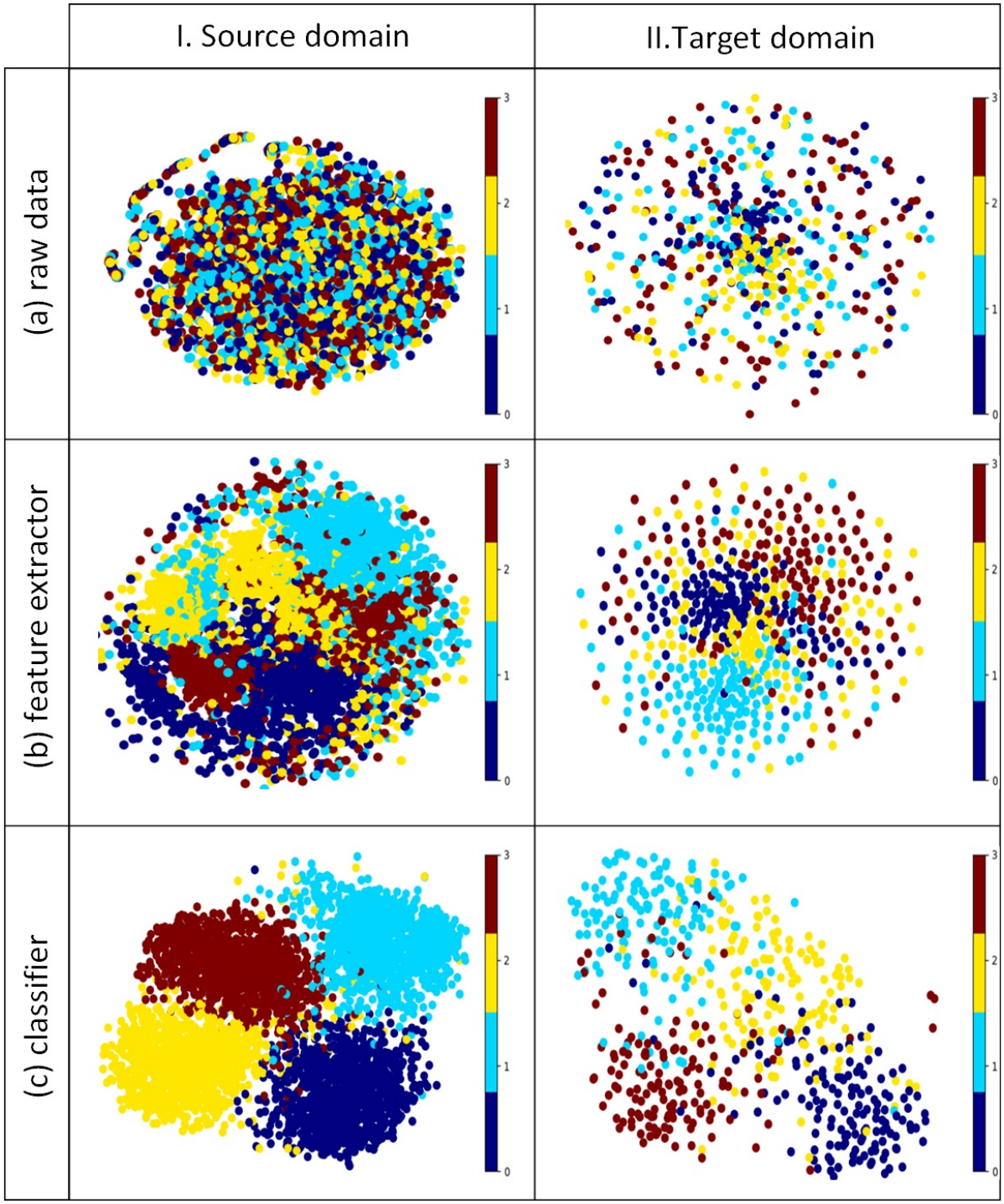
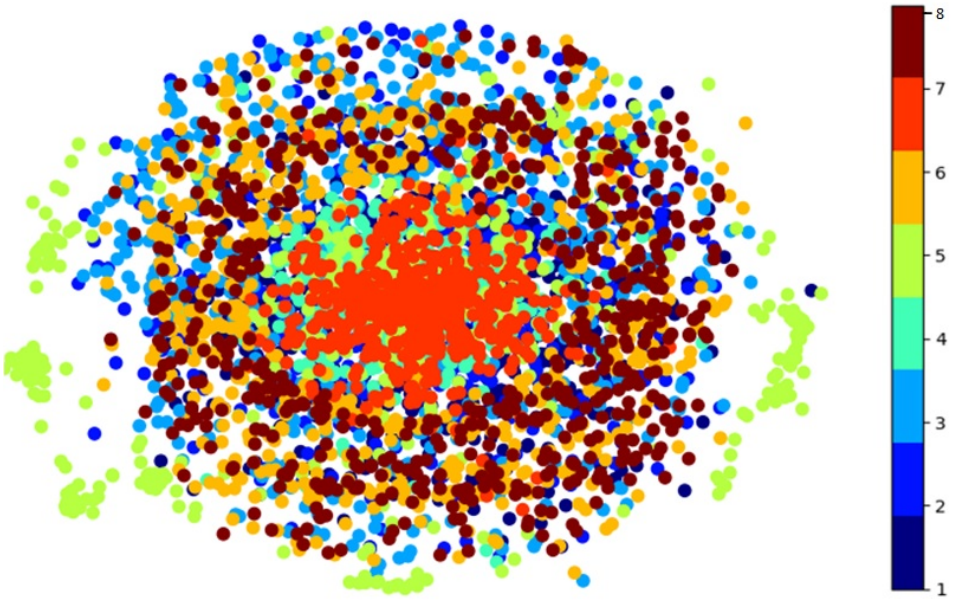
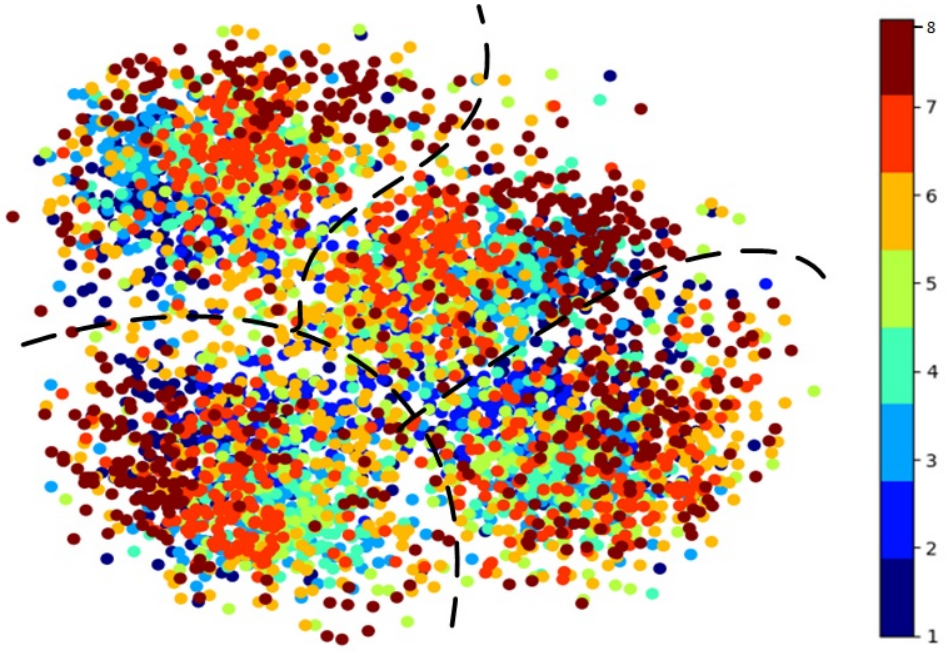


Figure 5.6: The feature maps obtained by t-SNE. Different colors denote different MI classification tasks. Part (a) - (c) is the data distribution of the different parts in the student network of the proposed model. Source domain I includes 8 subdomains namely subjects 1 - 7 and 9 while the target domain comes from the 7^{th} subject from the BCIC-IV-2a dataset.



(a)



(b)

Figure 5.7: The feature maps obtained by t-SNE. Different colors denote 8 different subdomains namely subjects 1 - 7 and 9 which are included in the source domain (a) The data distribution of the raw EEG signals. (b) The feature maps were extracted before the fully connected layer in the proposed model.

We utilized a knowledge distillation framework and trained the teacher and student networks, respectively. The teacher network comprises the spectral features fusion block, feature extractor, and classifier, whereas the student network consists solely of the feature extractor and the classifier. We used the pointwise convolution to adopt cross-frequency interactions corresponding to the MI information, which proves useful for enhancing the robustness of spectral representation. Then in the feature extractor, temporal-spatial convolution is employed to capture the discriminative features in MI EEG. The inclusion of two dense units, creating short connections within the CNN layers, facilitated feature refinement and the extraction of more abstract characteristics. To transfer the internally invariant spectral features from the teacher network to the student network, we employed MSE loss to encourage the student network's features to closely align with those of the teacher network. For mutually invariant features that are exploited from different subdomains, we used the correlation alignment method to align the data distribution and learn the cross-domain transferable knowledge. To eliminate the redundancy and the repeated information among two kinds of features, we use distance regularization to maximize their differences. To better validate the superiority of the proposed model, we conducted the experiments on two public datasets. Based on the data presented in Table 5.2 and Table 5.3, it is evident that our proposed model outperformed the state-of-the-art methods, attaining the highest classification performance. The ablation study in Table 5.4 also demonstrates the usage of two kinds of invariant features and the effect of distance regularization. The visualization results based on t-SNE in Fig 5.6 present the MI-EEG decoding performance of the proposed model. The feature maps obtained in the classifier based on source subjects exhibit very distinct clusters, effectively showing the feature distribution of different labels. Even though the target subject is the unseen domain, the proposed model can effectively classify MI tasks by utilizing acquired generalized information and applying it in a plug-and-play BCI system.

5.5 Conclusions

In this Chapter, the domain generalization technology is applied on the cross-subject MI-EEG decoding model to realize the plug-and-play functionality in BCI applications. The proposed model learns the internally and mutually invariant features from the source domain with no involvement in the target data. For internally invariant features, a knowledge distillation framework is used to fuse the spectral information corresponding to MI tasks and guide the proposed model to capture the invariant representations. For mutually invariant features, the correlation alignment is employed to extract the cross-domain representations. A distance regularization is also adopted to maximize two kinds of invariant features to enhance generalized expression. The proposed method outperforms benchmark models in cross-subject MI-EEG decoding, as evidenced by the classification accuracy and the feature distributions. The results proved that the proposed model achieves an accuracy improvement of 8.93% and 7.18% on the BCIC-IV-2a dataset and KU dataset respectively compared with other advanced deep learning methods.

The proposed model is validated in the same dataset. However, real MI-BCI applications usually have different types of EEG-collected devices which may lead to a huge disparity among data distributions. Meanwhile, the different number of electrodes brings the challenge for transfer learning and limits the usage of DG models. To address this issue, Chapter 6 will introduce a new structure based on the GCN.

CHAPTER 6

CROSS-DATASET MI DECODING - A TRANSFER LEARNING ASSISTED GRAPH CONVOLUTIONAL NETWORK APPROACH

The proliferation of portable EEG recording devices has made it practically feasible to develop MI-BCI. However, the low signal-to-noise ratio of EEG signals for abstract MI tasks, limited data, limited EEG channels, and strong inter- and intra-subject variability pose significant challenges for MI-task recognition. This chapter proposes a transfer learning assisted graph convolutional network (GCN) modeling approach for cross-dataset MI decoding, one of the most challenging issues in this field. In the experiments, a multi-channel dataset with 62 electrodes and a few-channel dataset with 8 electrodes are utilized for cross-dataset modeling. To harness multi-channel information, we utilize the GCN module to aggregate topological features. The pre-trained model is guided with few-channel signals as inputs through a knowledge distillation framework. Subsequently, the pre-trained model is adapted to the few-channel dataset using a transfer learning strategy with minimal data training. Experiment results show that the proposed model achieves up to 7.04% accuracy improvement compared with state-of-the-art models, demonstrating the effectiveness of the proposed approach in cross-dataset MI-EEG

decoding, thus enabling more effective MI-BCI applications.

6.1 Introduction

In the practical application of BCI systems, achieving good classification accuracy and robustness across different subjects with minimal or no retraining on new data poses a significant challenge. Transfer learning (TL) methods such as domain adaptation (DA) and domain generalization (DG) are gradually beginning to be employed to address this issue. However, previous studies only used single datasets for validation, which are often collected using numerous wet electrodes, ensuring higher data quality. In contemporary EEG applications, there is an increasing prevalence of portable EEG acquisition devices [164]. To save experimental time, these devices typically have fewer channels and employ dry electrodes, resulting in lower data quality [165], bringing a significant challenge for decoding MI-EEG signals. Additionally, due to variations in the number of channels, high-quality datasets cannot be directly utilized for transfer learning without channel selection. Research on MI-based cross-dataset studies remains limited. Zaremba and Atyabi [166] used three different datasets and filtered data with the same 11 channels existing across these datasets. However, this method merely consolidated subjects from different datasets for the cross-subject training, following a leave-one-out experiment, and did not address the differences between various datasets. Xu [167] and Xie [168] only choose three channels (C3, CZ, C4) across different datasets. The former method employed Riemannian Procrustes Analysis (RPA) [169] to align the Riemannian center among subjects within different datasets and train the DL-based model. The latter approach adopted fine-tuning technology when applying models to new MI paradigms. However, these methods do not consider the channel differences caused by various datasets or devices, as well as the potential transfer from a multi-channel dataset to one with fewer channels.

To remedy these limitations, we propose a GCN network based on the Knowledge Distillation [170] and fine-tuning methods, extracting the temporal-spatial-spectral features

in the multi-channel public dataset with high data quality, aggregating specific channels information and transferring to the dataset with much fewer channels for decoding MI tasks. In the experiment, the public dataset [94] collected by 62 wet electrodes is regarded as the source domain while the few-channel dataset collected by only 8 dry electrodes is regarded as the target domain. First, we train the proposed model as the teacher network with 62-channel inputs. To reduce the 62-channel data to the same specific 8 channels in the target domain, GCN layers are adopted to aggregate spatial information. Then the student network with these 8-channel inputs in the source domain is guided by the teacher network to learn the aggregated information and effectively harness all the data in the source domain. Finally, the pre-trained student model is validated on the target domain by fine-tuning technology with minimal data of re-training, aiming to transfer the parameters learned from the source domain model. From the experiment results, the proposed model is shown to achieve the highest accuracies among the compared benchmarks. Furthermore, ablation studies and visualization experiments are conducted to understand the effectiveness of GCN layers and transfer learning strategies.

The remainder of the chapter is organized as follows. 6.1 gives two dataset descriptions and the detailed structure of the proposed model. 6.3 details the results including the ablation studies and visualization experiments. Discussions are presented in 6.4 and we conclude them in 6.5.

6.2 Methods

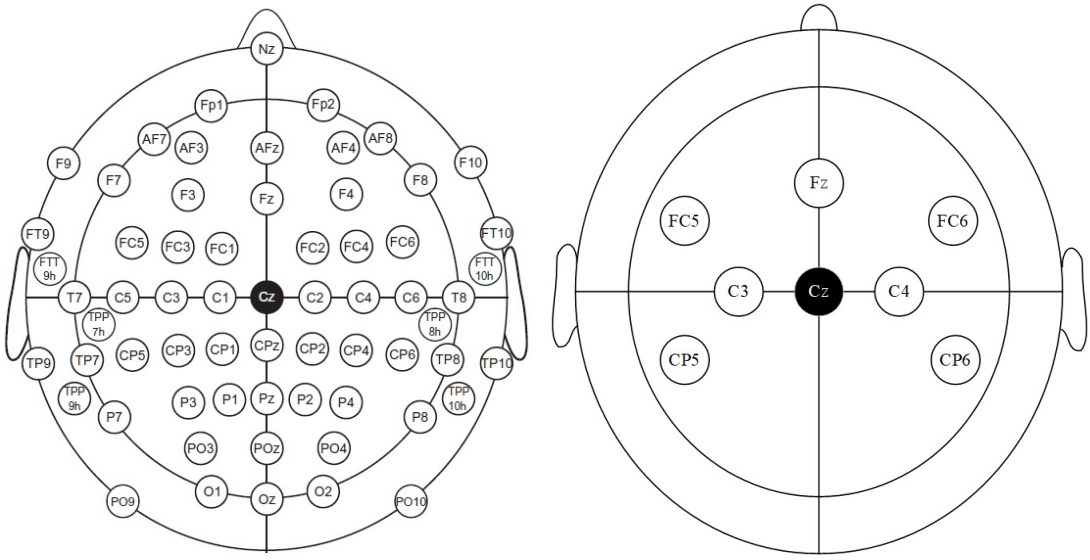
In this chapter, datasets collected from two different EEG devices are introduced. Subsequently, we present the preprocessing steps, the specific details of the proposed model, and the experimental procedure.

6.2.1 Data Description

1) Korean University dataset [94] (KU dataset): The EEG signals were collected using a device with 62 wet electrodes whose impedances were maintained below 10 $k\Omega$. The

EEG channel configuration (Fig 6.1(a)) conformed to the International 10-20 system. This dataset comprised 54 healthy individuals performing left and right-hand motor imagery tasks. Each subject participated in two experimental sessions, with each session consisting of 200 trials. Consequently, each individual contributed a total of 400 trial data. In this experiment, we downsampled the sampling rate from 1000 Hz to 250 Hz.

2) Few-channel dataset (8-channel dataset): The EEG signals were collected using a device with only 8 dry electrodes (Fig 6.1(b)) whose impedances were about 300 $k\Omega$. This device is portable and features a plug-and-play functionality, eliminating the need for bridging the gap between electrode pin and scalp with conductive electrolyte gel and significantly reducing experimental preparation time. However, its impedance is much higher compared to wet electrodes, leading to a decrease in data quality. This dataset included 22 healthy subjects performing left and right-hand motor imagery tasks. Each subject had 80 trials with a sampling rate of 250 Hz.



(a) (b)

Figure 6.1: The channel configuration of the International 10-20 system: (a) KU dataset with 62 channels. (b) Few-channel dataset with 8 channels.

6.2.2 Framework

The residence of the portable device with 8 dry electrodes is much higher than the device with 62 wet electrodes used in the KU dataset. Moreover, the limited amount of data in the 8-channel dataset is insufficient for achieving robust classification results when utilized for within-subject modeling, particularly for deep learning models. In practical BCI applications, there is a greater need for models that can be used immediately or with minimal data calibration. Therefore, transferring model information from the KU dataset with high data quality and abundant data to the 8-channel dataset with fewer channels holds meaningful and valuable implications. However, the two datasets utilized devices with a different number of channels. While it is feasible to only use the 8 channels common to both devices like previous studies [166][167][168], this approach does not fully exploit the additional channel information present in the KU dataset. To address this problem, we divide the whole experiment framework into two parts. The first part employed data distillation, enabling the model to learn a compact representation that captures the task-specific feature representation when using all 62 channels, even though the model was trained with only 8 channels as input. The second step involves fine-tuning of the pre-trained model using a minimal amount of data in the target domain data for training, followed by validation of the remaining target domain data.

6.2.3 Model Structure

The proposed model primarily consists of four components: Temporal Block, Dense Block, Graph Block, and Feature Extraction (shown in Fig 6.2).

6.2.3.1 Temporal Block

The EEG signals are denoted as $E = (X_i, Y_i)|i = 1,2,...,N$, where $X_i \in R^{C \times T}$ represents $i-th$ EEG trial with C channels and T samples. N is the total number of EEG signal trials. First, the EEG signals are sent into three parallel CNN layers with multi-scale temporal kernels. [133] demonstrated that the optimal kernel size differs among subjects and may

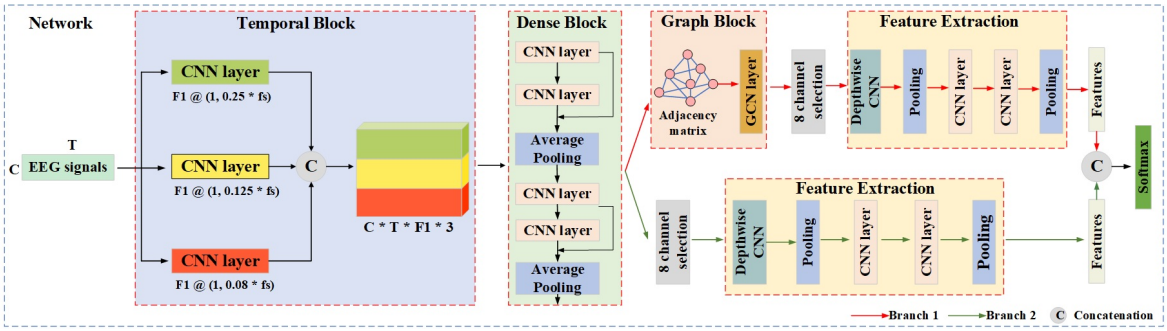


Figure 6.2: The detailed structure of the proposed model.

vary over time for the same subject. Simultaneously conducting convolution at multiple scales and then aggregating allow for the extraction of features at different scales, resulting in richer temporal features and implying more accurate classification judgments during the final decision-making process. Define the ratio as $\alpha = \{\alpha^i \mid i = 1,2,3\}$, where i represents three parallel CNN layers. Hence, the kernel size is denoted as:

$$k^i = \left(1, \alpha^i \cdot f_s\right), i = 1,2,3 \tag{6.1}$$

where f_s is the sampling rate of the EEG signals. Then, three outputs with different scale temporal representations are concatenated and fed into the Dense Block for association and fusion.

6.2.3.2 Dense Block

The Dense Block consists of four CNN layers and two average pooling layers, fusing the concatenated temporal features and further refining useful features. To enhance information flow among the CNN layers, the outputs of each CNN filter were propagated to all subsequent layers, which generated the final output incorporating the extracted features from all preceding layers [126]. The connection between two common CNN layers is:

$$x_l = F_l(x_{l-1}) \tag{6.2}$$

where x_{l-1} and x_l are the input and output of the layer l. In the dense block, the lth layer receives the feature maps from all preceding layers:

$$x_l = F_l([x_0, x_1, ..., x_{l-1}]) \tag{6.3}$$

where $[x_0, x_1, ..., x_{l-1}]$ are the feature maps before the layer l. In the experiment, we used two CNN layers and one pooling layer as one combination. Therefore, the output averaged feature maps from two preceding CNN layers:

$$x_l = F_{Average}(F_{l-1}(x_{l-1}) + F_{l-2}(x_{l-2})) \tag{6.4}$$

Such connections create short paths which enhance the flow of information and feature reuse. Meanwhile, each CNN layer utilizes ELU as the activation function, followed by batch normalization and dropout techniques to suppress the overfitting problem. One CNN layer generates k feature maps contributing to the subsequent layer. Here we set $k = 10$ so that each combination produces 20 feature maps.

6.2.3.3 Graph Block

An undirected and weighted graph can be described as $G = (V, E)$ where V represents the nodes and E denotes the edges among the nodes. In the proposed model, each EEG channel is regarded as a node of the graph while edges are the relationship between channels. The adjacency matrix $W \in R^{N \times N}$ is built to describe the connection relationship between different nodes, where N is the number of channels. Nevertheless, the intricate nature of the activation states in the human brain during MI tasks poses a challenge in constructing an artificial matrix based on prior knowledge. Ma et al [171] learned the channel similarity based on semi-supervised learning and then manually selected 11 channels as inputs. EEG-GENet [172] set the edge between neighboring channels to 1 and those not neighboring to 0. Delvigne et al [173] used the distance as prior knowledge to create an adjacency matrix. To better adapt to the characteristics of end-to-end learning procedure in a DL model, the channel connections based on the temporal features learned for each channel are dynamically learned, ensuring a trainable adjacency matrix.

First, Pearson's correlation matrix (PCM) is adopted to initialize the matrix. PCM is an effective tool to capture the topological information among EEG channels [174][172]. If one trial EEG data is defined as $X \in R^{C \times T}$ with C channels and T temporal features,

the Pearson's correlation coefficient can be obtained by:

$$P_{ij} = \frac{cov\left(X_i, X_j\right)}{\sqrt{var(X_i)var\left(X_j\right)}} \tag{6.5}$$

where i and j denotes the i^{th} and j^{th} channel of the EEG signals. Therefore, the initialized adjacency matrix is:

$$A_{initial} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,C} \\ \vdots & \ddots & \vdots \\ P_{C,1} & \cdots & P_{C,C} \end{bmatrix} \tag{6.6}$$

To make the adjacency matrix trainable and dynamically analyze the similarity among channels, a mask matrix of the same size consisting of trainable parameters is adopted:

$$A_{trainable} = \begin{bmatrix} W_{1,1} & \cdots & W_{1,C} \\ \vdots & \ddots & \vdots \\ W_{C,1} & \cdots & W_{C,C} \end{bmatrix} \tag{6.7}$$

where w is the weight initialized based on xavier uniform [175]. To make the symmetric trainable matrix, $A_{trainable}$ and its transposed are multiplied and applied to $A_{initial}$:

$$A = \Phi_{relu}\left(A_{Initial} \odot \left(A_{trainable} \cdot A^T_{trainable}\right)\right) + I \tag{6.8}$$

where Relu activation is employed to ensure the matrix is non-negative. The degree matrix is $\widetilde{D} = \sum_j A_{ij}, i \neq j$. The normalized adjacency matrix can be calculated as:

$$\widetilde{A} = \widetilde{D}^{-\frac{1}{2}} A \widetilde{D}^{-\frac{1}{2}} \tag{6.9}$$

Once the matrix $\widetilde{A}$ is obtained, a GCN layer with weights and bias vector is applied on the input feature maps:

$$X_{output} = \Phi_{elu}\left(\widetilde{A}XW + bias\right) \tag{6.10}$$

After the graph block operation, each EEG channel includes the information aggregated from other channels.

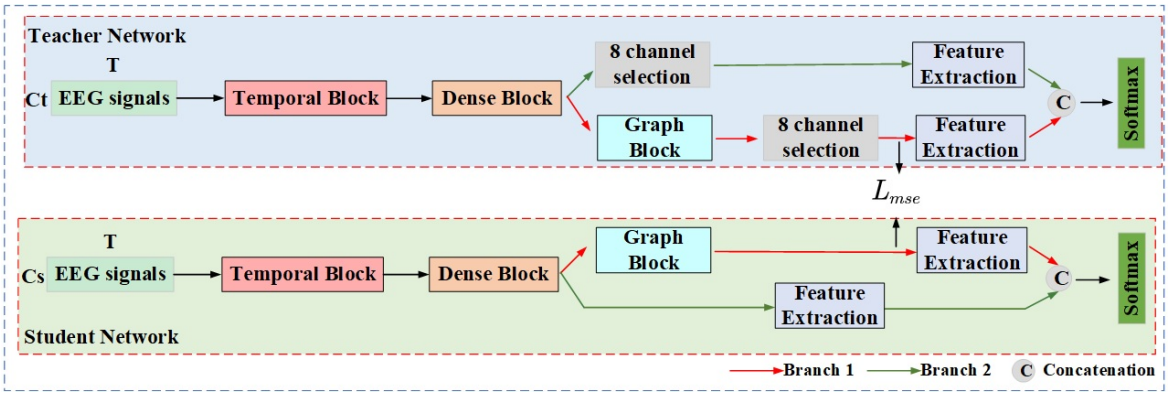


Figure 6.3: The framework of the Knowledge Distillation.

6.2.3.4 Channel Selection and Feature Extraction

If the input EEG data has 62 channels, a channel selection step is required; otherwise, it is not necessary. The channel selection procedure chooses the specific 8 channels (FC6, C4, Fz, C3, FC5, CP6, Cz and CP5)(Fig 6.1(b)) same in the target domain data. Subsequently, both the branch with the graph block and the branch without the graph block undergo further feature extraction. The depthwise CNN layer helps the model to extract global spatial features while reducing computational complexity compared with common CNN layers. Then two average pooling layers and CNN layers follow to fuse the feature maps and decrease dimensionality. The one with graph block aggregates other channels' features and topological information while the other one focuses on mining temporal-spatial features. The different views of feature representations brought by two parallel branches help enhance the model's robustness and classification performance.

6.2.4 Training Procedure

The training procedure was initially conducted on the source domain dataset, employing Knowledge Distillation and a two-stage training strategy [176] to build a pre-trained model. Then this model was validated on the target domain with fine-tuning technology.

In the source domain, the data consists of 62 channels, while the target domain has only 8 channels available. To leverage the extensive data in the source domain and transfer model parameters to the target domain, our proposed model is designed to

learn feature representations and distributions by aggregating information from all 62 channels. However, the model is currently configured to accept 62-channel inputs, making it unsuitable for direct use in the target domain with only 8 channels. Training the model with the limited 8-channel data in the source domain would result in the loss of channel aggregation information. To address this, we employ a Knowledge Distillation framework, constructing both a teacher network and a student network with different numbers of channels as input (Fig 6.3). First, all subject data in the source domain are divided into training data and validation data based on 5-fold cross-validation (CV). In the first stage, the teacher network was trained with 62-channel data from the training set in the source domain as inputs and incorporated the channel selection step during training. The early-stopping tool monitored the validation set accuracy and stopped the training if there was no increase in the next 150 epochs. Then the best validation accuracy and the corresponding validation data are saved for the next stage. In the second stage, the student network was trained with 8-channel data from all data (training set + validation set) in the source domain. During this procedure, the student network did not need the channel selection step but guidance from the teacher network to align the feature distributions between the two networks. A Mean Squared error loss was used to calculate the distance among the feature maps after the graph block:

$$\mathscr{L}_{mse} = \frac{1}{n}\sum_{i-1}^{n}(x_{teacher} - x_{student})^2 \tag{6.11}$$

where n is the number of trials and x are the feature maps.$x_{teacher}$ had the channel aggregation knowledge based on 62-channel while $x_{student}$ had no other channels information. By minimizing $\mathscr{L}_{mse}$, the student network learned the feature representations with aggregation knowledge and abundant topological information extracted by the graph block in the teacher network. When it was fine-tuned in the target domain, useful pre-trained parameters were transferred effectively. The second stage used the saved validation set and stopped when the validation accuracy was higher than the one recorded in stage one. Even if the final accuracy can not reach the same value, the model will stop with the early stopping criteria to prevent the occurrence of infinite training.

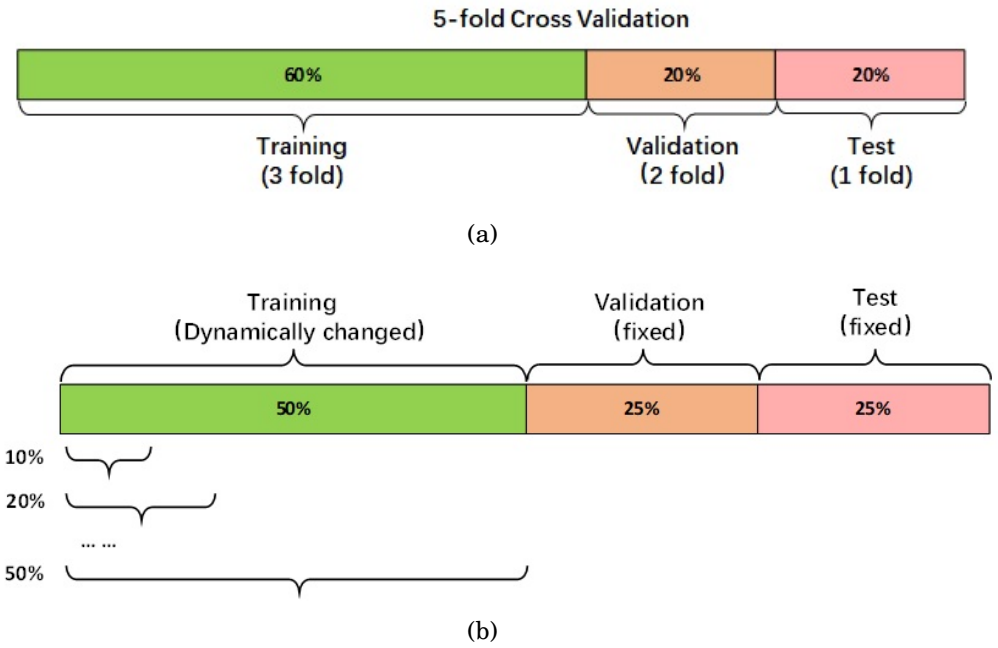


Figure 6.4: Two scenario descriptions. (a) Scenario 1 with 5-fold CV, (b) Scenario 2 with fixed validation and test set

Subsequently, the student network was used as a pre-trained model being validated in the target domain. Despite the model learning to classify MI tasks using only 8 channels in the source domain, calibration based on limited target data was still necessary due to the utilization of two datasets from entirely different devices. The source domain dataset employs wet electrodes with good data quality, while the target domain dataset uses dry electrodes with an impedance reaching 300 $k\Omega$, resulting in poor data quality. Directly using the pre-trained model without verification yields poor classification model accuracy. Hence, fine-tuning is applied, retraining the parameters of the model based on little target domain data. We designed two scenarios in the target domain validation: 1) A 5-fold CV was employed with 3 folds for training, 1 fold for validation, and the rest for testing (Fig 6.4a). 2) The dataset is split in half, with one part designated as a fixed validation and test set, and the other part used as the training set. In the training set, not all the data is used at once; instead, experiments are conducted by incrementally adding 10% of the data each time, and results are recorded until the training set reaches half of the total dataset size (Fig 6.4b).

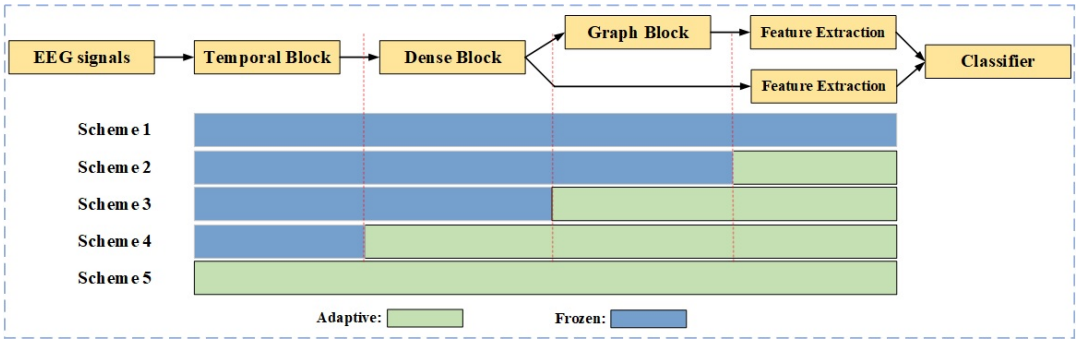


Figure 6.5: The schemes of the fine-tuning framework.

Since each model contains several operation layers, whether the parameters were frozen based on the functionality of each block. 5 schemes for fine-tuning strategy were conducted in the experiment (Figure 6.5). Different schemes froze different blocks in the pre-trained model while the rest blocks were adaptive and re-trained based on limited data in the target domain. Scheme 1 froze all layers namely no data in the target domain involved in training. This Scheme no longer required any new data, greatly reducing validation time. However, it resulted in a decrease in classification accuracy due to significant data distribution differences caused by different EEG acquisition devices. Scheme 5 made all layers adaptive namely all parameters in the model were updated to match the target data.

6.2.5 Training Setup

The cross-entropy function was adopted to evaluate the distance between the probability distribution of the model prediction values y_p and the true labels y_t:

$$L\left(y_p, y_t\right) = -\sum_m y_{p,m} \log y_{t,m}. \tag{6.12}$$

where m is the index of y. Adam optimizer was used with 0.001 as the learning rate. The computer used in this experiment had 22 Intel processors and 80 GB RAM. GTX 4090 GPU with 24 GB memory was used for training and testing MI-EEG signals. Pytorch 1.10.0 was used for building the proposed model.

6.3 Results

We use two traditional machine learning methods (CSP [138] and FBCSP [70]), three CNN-based models (Shallow ConvNet [85], Deep ConvNet [85], and EEGNet [132]) and two GCN-based models (EEG-GENet [172] and EEG-ARNN [177])as benchmarks to demonstrate the effectiveness of our proposed method. All the baseline methods used the parameters and structures suggested by their authors for a fair comparison. The details of the baseline models are described as follows:

1. Machine learning methods: CSP and FBCSP are two classic machine learning algorithms. The core idea is to find a set of optimal spatial filters that can separate features after projection. FBCSP goes a step further by dividing the data into multiple sub-bands and identifying informative and discriminative pairs of sub-bands. Both models are lightweight, easily modifiable, and widely applied. In the experiment, the Support Vector Machine (SVM) was employed as the classifier.

2. CNN-based models: Shallow ConvNet, Deep ConvNet, and EEGNet have excellent performance on MI-EEG classification and robustness. They utilize a 1-D CNN and a deepwise CNN layer to extract temporal-spatial features. Then, the Deep ConvNet model combines several common CNN layers and pooling layers before the classifier while the Shallow ConvNet only adopts a squaring layer with the log operation. EEGNet uses the pointwise CNN layer to reduce amounts of calculation resources while ensuring informative learned features.

3. GCN-based models: The EEG-GENet model is built based on the structure of EEGNet. After extracting the temporal features by a CNN layer, a GCN layer is followed to capture the topology information according to the EEG electrodes. EEG-ARNN combines one CNN layer and one average pooling layer as a module. The GCN layers are added after each module with a trainable adjacency matrix which is initialized with one. Both of them perform well on the public BCICIV-2a dataset [131].

Table 6.1: Comparison of classification accuracy (%) and standard deviation (Std) on the 8-channel dataset.

Subject	CSP	FBCSP	Shallow ConvNet	Deep ConvNet	EEGNet	EEG-GENet	EEG-ARNN	Proposed model
1	65.00	75.00	72.50	45.00	47.50	77.50	57.50	**78.75**
2	56.25	81.25	67.50	58.75	66.25	72.50	68.75	**86.25**
3	58.75	**70.00**	61.25	65.00	55.00	67.50	47.50	65.00
4	86.25	**93.75**	82.50	65.00	76.25	57.50	78.75	83.75
5	56.25	42.50	53.75	63.75	72.50	61.25	72.50	**72.50**
6	45.00	61.25	51.25	73.75	82.50	43.75	81.25	**86.25**
7	45.00	**56.25**	43.75	47.50	50.00	47.50	45.00	48.75
8	50.00	57.50	52.50	50.00	55.00	51.25	58.75	**58.75**
9	50.00	55.00	47.50	48.75	42.50	55.00	53.75	**61.25**
10	56.25	55.00	58.75	60.00	76.25	51.25	70.00	**80.00**
11	51.25	70.00	71.25	53.75	55.00	63.75	63.75	**73.75**
12	47.50	53.75	61.25	55.00	**63.75**	58.75	52.50	62.50
13	67.50	**90.00**	86.25	50.00	58.75	80.00	60.00	82.50
14	56.25	66.25	52.50	81.25	**91.25**	60.00	88.75	88.75
15	55.00	70.00	73.75	71.25	70.00	58.75	**80.00**	75.00
16	45.00	43.75	45.00	56.25	53.75	58.75	55.00	**66.25**
17	47.50	51.25	50.00	55.00	58.75	56.25	52.50	**58.75**
18	57.50	51.25	53.75	45.00	48.75	57.50	51.25	**60.00**
19	60.00	60.00	77.88	62.50	**82.50**	58.75	66.25	80.00
20	**60.00**	52.50	46.25	50.00	55.00	53.75	51.25	52.50
21	80.00	**95.00**	82.50	72.50	86.25	85.00	83.75	93.75
22	51.25	**60.00**	51.25	56.25	42.50	41.25	48.75	51.25
Avg	56.70**	64.15**	61.04**	58.47**	63.18**	59.89**	63.07**	**71.19**
Std	10.60	15.11	13.47	9.98	14.52	11.09	13.13	13.30
Acc≥ 75	2	5	4	1	6	3	5	**10**

The improvement of the proposed model over the baseline methods with $^{*}: p < 0.05$ and $^{**}: p < 0.01$

6.3.1 Overall performance

We conducted the experiments based on the 5-fold CV on the 8-channel dataset. The statistical significance tests including an Analysis of Variance (ANOVA) test and paired t-tests between each baseline model and the proposed model. The result (shown in Table 6.1) demonstrated that the proposed model achieved the highest accuracy of 70.23%, which was 14.49% ($p < 0.001$), 7.04% ($p < 0.001$), 10.15% ($p < 0.001$), 12.72% ($p < 0.001$), 8.01% ($p < 0.001$), 11.30% ($p < 0.001$), and 8.12% ($p < 0.001$) higher than the CSP, FBCSP, Shallow ConvNet, Deep ConvNet, EEGNet, EEG-GENet and EEG-ARNN respectively. In the traditional machine learning methods, FBCSP performed best, surpassing even other DL models. The possible reason is that the 8-channel dataset is more sensitive to frequency band filtering. FBCSP was the only method among these baseline models that divided the original data into multiple sub-bands and performed band selection. In the DL models, EEGNet and EEG-ARNN performed better and got an accuracy of approximately 63%. However, regardless of the model used, there is a significant

Table 6.2: Comparison of classification accuracy (%) and standard deviation (Std) based on the proposed model with two cases.

	With Fine-tuning	W/O Fine-tuning
Avg	71.19	66.36
Std	13.30	14.89

variation in the classification results for each individual, further emphasizing that the model's performance varies from subject to subject. The proportion of subjects who had accuracy over 75% was 9.1% (2 of 22), 13.63% (3 of 22), 18.18% (4 of 22), 4.54% (1 of 22), 27.27% (6 of 22), 13.63% (3 of 22), 22.72% (5 of 22), 45.45% (10 of 22) for CSP, FBCSP, Shallow ConvNet, Deep ConvNet, EEGNet, EEG-GENet, EEG-ARNN and proposed model respectively, demonstrating notable classification performance of the proposed model.

6.3.2 Analysis of Different Schemes

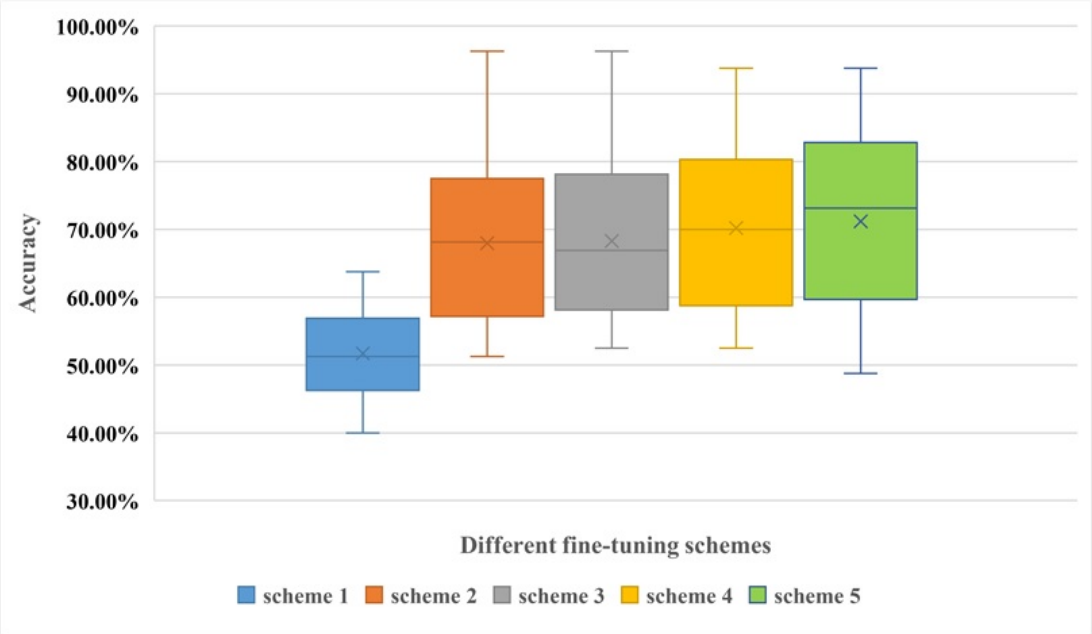


Figure 6.6: Different schemes of the fine-tuning framework.

We first compared the results (Table 6.2) of the proposed model using fine-tuning method and modeling solely using the target domain data (w/o fine-tuning). Even with-

out learning knowledge from the source domain data and pre-training the model, the proposed model structure still achieved a classification performance in the target domain that is at least 2% higher than benchmarks. To further validate the influence of different blocks in the model during the fine-tuning process, we conducted experiments based on different schemes (Fig 6.5). The box plot (Fig 6.6) illustrated that the more parameters involved in adaptive tuning, the better the model's performance. Scheme 1 froze all layers so that no weights could be updated to adapt to the target domain, leading to the worst classification accuracy. The data distribution divergence across datasets and devices limited the model's performance and robustness. Although updating all parameters increases the computational load, achieving a 71% accuracy in a limited dataset collected from only 8 dry electrodes is worthwhile and makes it effectively applicable to portable devices.

6.3.3 Analysis of Training Proportion

In practical applications of BCI, obtaining a large amount of data has always been challenging. Therefore, calibrating the model with little or no data is necessary. In the 8-channel dataset, only 80 trials were collected for each individual, significantly less than the data in public datasets. Following Scenario 2, we divided the entire 8-channel dataset into two halves, fixing the validation and test data set. The training set was used to adapt the parameters of the pre-trained model based on fine-tuning method. The training data started from 0% and increased by 10% of the total data volume in each experiment, up to 50%, for a total of 6 experiments. The result (shown in Fig 6.7) indicated that with more training data used for adaptation, the model's classification performance improved. If no data in the 8-channel dataset were used to adapt, the model only reached 51.7% which could not discriminate any MI tasks. Even with 10% of the data namely 9 trials used for adaptation, the model's classification accuracy can improve by 7.39%. When the training data volume reached 40% of the whole dataset, the classification accuracy was 67.77% which was higher than the results of baseline models based on 5-fold CV. Therefore, with the assistance of fine-tuning techniques, the proposed model can achieve

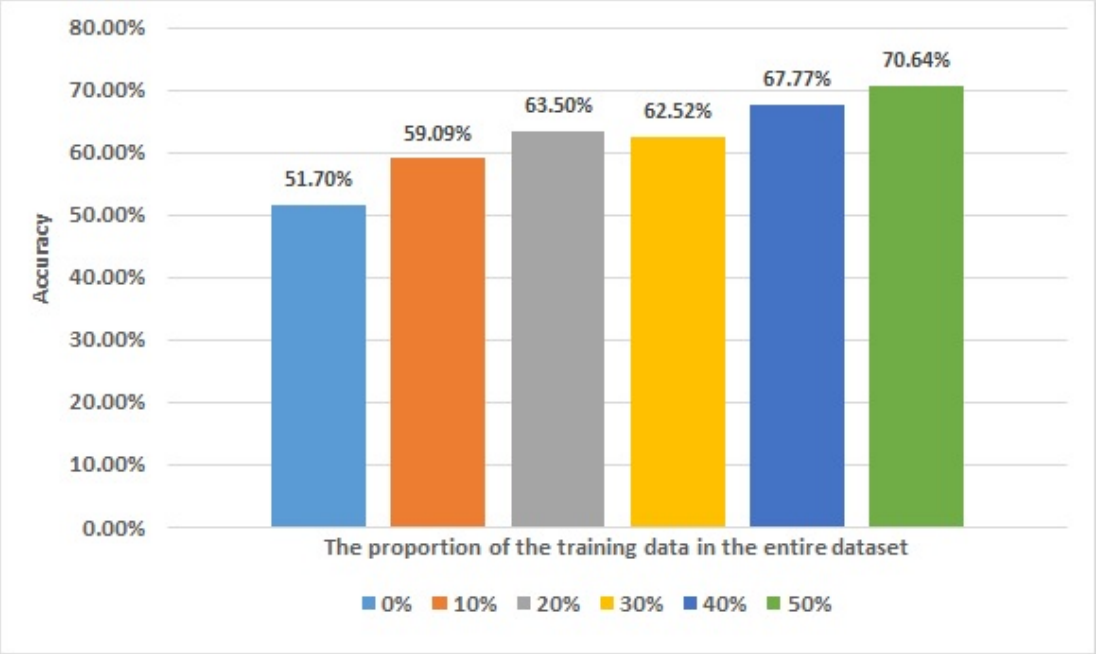


Figure 6.7: The results using different training data volumes.

Table 6.3: The classification accuracy (%) of the ablation study.

	W/O D_Block	W/O G_Blcok	Proposed model
Avg	65.79	70.11	71.19
Std	13.18	13.20	13.30

performance surpassing the use of within-subject models in the target domain, even with the adaptation and calibration using a small amount of target domain data, further demonstrating the practicality of the proposed model.

6.3.4 Ablation Study

To validate the contribution of the Dense Block and Graph Block which were important components in the proposed model, an ablation study was conducted: 1) Without Dense Block (W/O D_Block): Dense blocks were utilized to capture information from the fused feature maps extracted by the Temporal Block. The Dense Block included 4 CNN layers and 2 average pooling layers. All of them were abandoned in the ablation experiment. 1) Without Graph Block (W/O G_Block): Graph Block was adopted to learn the topological information based on electrodes and transfer the knowledge from data with 62 channels

to data with 8 channels. In the ablation experiment, we prohibited the transmission of topological information learned by the graph block in the knowledge distillation framework. The results in Table 6.3 showed that the proposed model had an accuracy of 5.4% ($p = 0.02$) and 1.08% ($p = 0.03$) higher than the W/O D_Block model and W/O G_Block, respectively. The Dense Block contributed more because it refined the temporal block and involved more parameters while only one GCN layer existed in the Graph Block.

6.3.5 Influence of Aggregated Channels

The source domain has 62 channels while the target domain only has 8 channels. To transmit the topological information learned from 62-channel data, the features from the rest of the k channels were aggregated in the GCN layer. To validate the influence of K-aggregated channels, we adjusted the adjacency matrix. First, C_s defined as one of the 8 specific channels was selected. Then we sorted the other channels that were not included in these 8 channels in descending order based on the weights obtained by the trainable adjacency matrix. The m channels with the smallest weights, specifically those least correlated with these 8 specific channels, were set to 0 in the adjacency matrix, ensuring that the GCN layer did not consider their information when aggregating channel features. We conducted 7 experiments from using all 62 channels to only 9 channels. The result in Fig 6.8 indicated that when the number of fused channels decreased and the limited information was captured, the overall classification performance of the model also decreased accordingly.

6.3.6 Visualization

1) Adjacency matrix visualization: We recorded the weights of the trainable adjacency matrix to validate the channel relationships learned by the proposed model. Fig 6.9 shows the heatmaps of the adjacency matrix in the teacher network model which was trained based on the source domain. In Fig 6.9(a), These 8 channels were only connected to themselves because the initialization matrix included a self-loop step. Following training

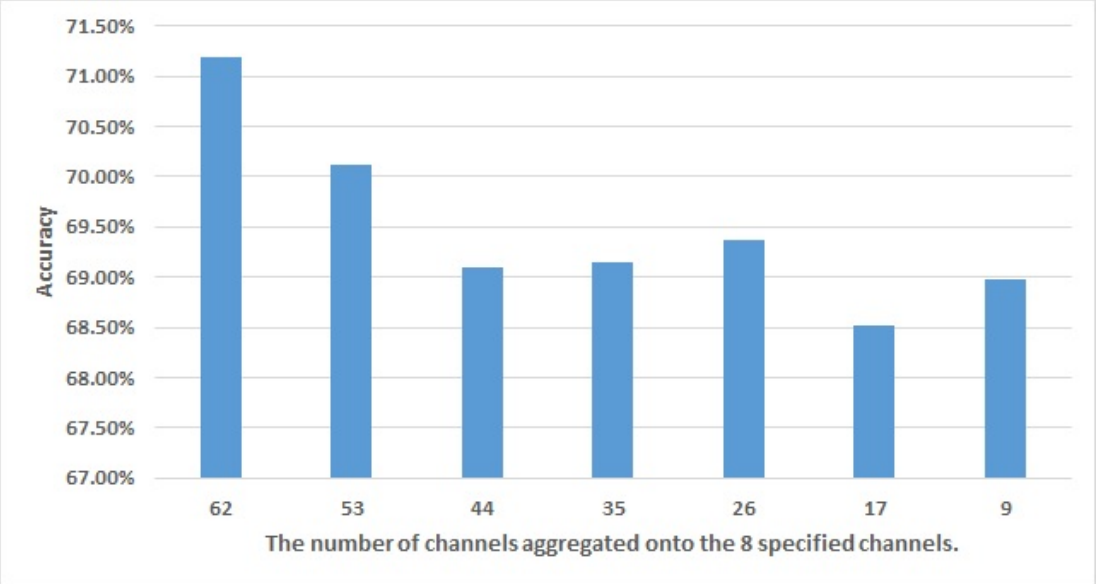
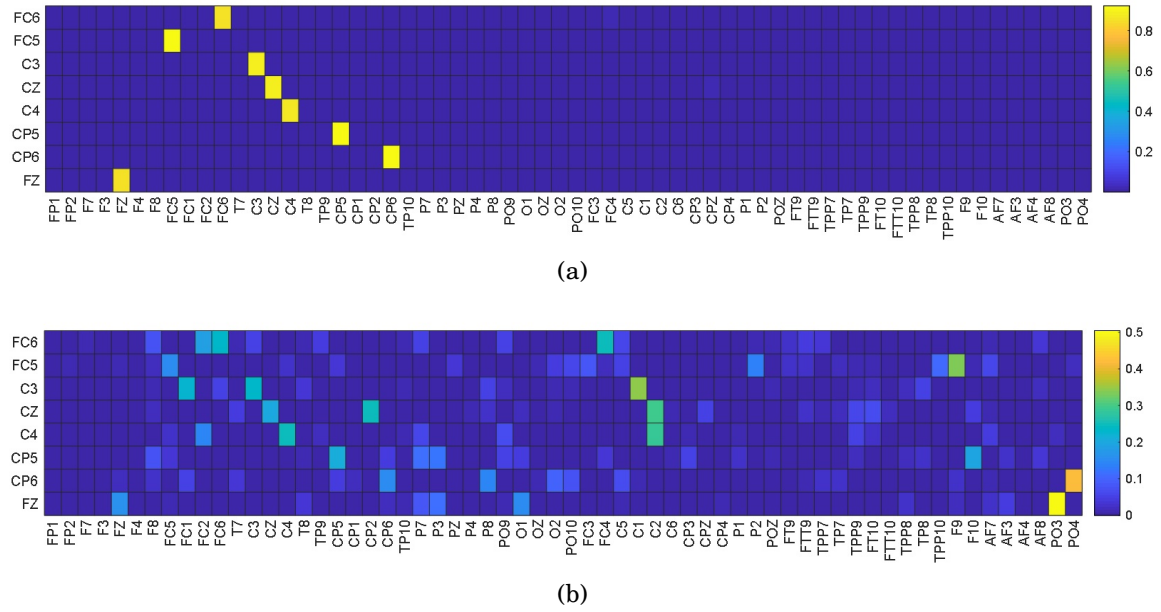


Figure 6.8: The accuracy of the proposed model with different numbers of aggregated channels.

and validation based on early stopping criteria, the channels established relationships with each other and aggregated based on the trained adjacency matrix, contributing to the final classification of MI tasks (Fig 6.9(b)). FC6, FC5, C3, C4, and Cz have stronger connections with their neighboring electrodes. For instance, connections between 1) FC6 and (FC4 and FC2), 2) FC5 and F9, 3) C3 and (C1 and FC1), 4) Cz and (CP2 and C2), and 5) C4 and (FC2 and C2). CP5, CP6, and Fz have more relations with channels P7, P3, PO3, and PO4 which belong to the parietal and parieto-occipital lobes. It can be observed that many channels strongly correlated with these specific 8 channels are not part of the same set. Without aggregating by the GCN layer, the information associated with these correlated channels will be missing, leading to a decrease in model classification accuracy.

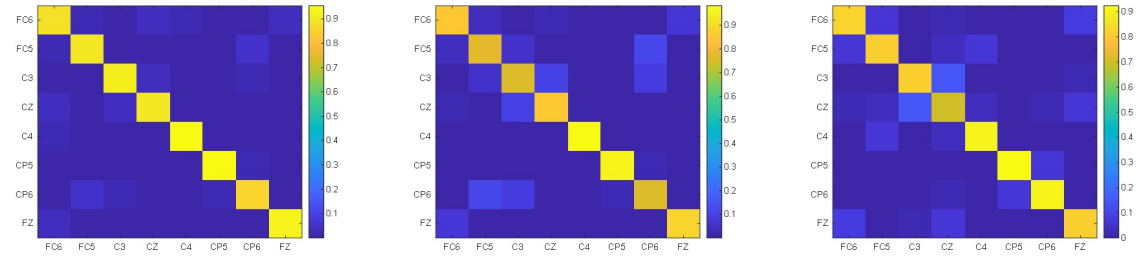
During the knowledge distillation procedure, the student network was guided by the teach network and better extracted the features based on the 8-channel inputs in the source domain. Compared with the trained adjacency matrix (Fig 6.10(b)) and untrained matrix (Fig 6.10(a)), some connections were strengthened like C3 and Cz. Fig 6.10(c) is the fine-tuned model of the 21^{st} subject which reached an accuracy of



(a)

(b)

Figure 6.9: The heatmaps of the adjacency matrix in the teacher network model: (a) Untrained model, (b) Trained model.



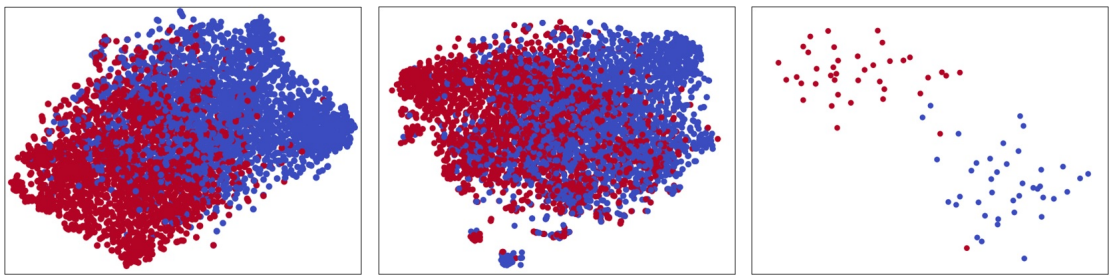
(a) (b) (c)

Figure 6.10: The heatmaps of the adjacency matrix: (a) Untrained model (Student network), (b) Trained model (Student network), (c) Fine-tuned model (the 21^{st} subject in the 8-channel dataset).

93.75% in the 8-channel dataset. Compared with the pre-trained model (Fig 6.10(b)), the fine-tuning method allowed the adjacency matrix to further adapt to the 8-channel dataset and reconstruct the whole relations among channels. The channels Cz and C3 still maintained a strong connection while the relationships between CP6 and FC5, CP6 and C3 decreased. Some connections were activated like Cz and Fz, CP5 and CP6, and C4 and FC5. Due to the fine-tuning method applied to each subject in the pre-trained model, the reconstructed relationships among channels varied. However, further research is

needed to explore the relationship between the activated channels and the classification performance of each within-subject model.

2) Feature Visualization: The t-distributed Stochastic Neighbor Embedding (t-SNE) method was utilized to visualize the feature maps of the fully connected layer before the final classifier of the proposed model. Fig 6.11(a) and Fig 6.11(b) are the teacher network and student network trained in the source domain. Fig 6.11(c) is the feature map of the fine-tuned model based on the 21^{st} subject in the target domain. Each subject in the target domain only has 80 trials so the limited points are shown in Fig 6.11(c). Based on the t-SNE analysis, the proposed model demonstrated strong capabilities in EEG signal classification. The classification boundaries of the teacher network's feature map appear more distinguishable than those of the student network. One reason for this is that the teacher network took the data with 62 channels as input, incorporating more information, while the student network only had 8 channels as input. Although the student network learned topological features from the teacher network, there was still a slight loss in classification performance.



(a) (b) (c)

Figure 6.11: The feature map of the proposed model: (a) Teacher network, (b) Student network, (c) Fine-tuned model.

6.4 Discussions

With the proliferation of portable devices, the research and application of BCI has gained much more momentum. In real-life applications, it is challenging to collect large amounts of high-quality data, and there is a strong demand for reducing experimental preparation

time. Therefore, it is crucial to ensure excellent accuracy in MI-task classification while reducing calibration time and the amount of required training data. DL models have shown promising results in decoding EEG signals, and transfer learning has been effectively applied to shorten verification times. However, few models can be generalized across datasets, especially when the datasets are collected using different devices with different channels. The target data in our experiment were collected using dry electrode devices, which have limited quantity, lower quality, and a very restricted number of channels, making it challenging to directly use the models trained with past public datasets. Therefore, we first utilized GCN to learn the topological knowledge of EEG channels on a public dataset with 62 channels. Subsequently, through a knowledge distillation framework, the feature distribution obtained from classifying MI tasks based on 62-channel data in the source domain was adopted to guide the proposed model with 8-channel inputs. Finally, the pre-trained model employed fine-tuning for adapting target domain data.

In our experiments, the proposed model achieved the highest classification accuracy compared with machine learning methods, CNN-based and other GCN-based models. To better validate the practicability of the model, scenario 2 with fixed validation and test data was conducted to examine how the model performed in the case of training with a small amount of data. By altering the training data volume, we found that the model achieved a classification accuracy of 67.77% when the training data constituted 40% of the entire dataset, which was higher than the results of baseline models using a 5-fold CV, where 60% of the entire dataset was employed for training. The model with fine-tuning technology built based on the source domain has a classification accuracy 4.83% higher than the model built only based on the target domain, demonstrating the effectiveness of the transfer learning. The different schemes of fine-tuning also influenced the model performance. The more parameters involved in the adaptation, the better the model performed. Besides that, we also validated how the number of aggregated channels affects the model performance. When the teacher network captures features from more channels and guides the student network, the final fine-tuned model will

achieve higher classification accuracy.

6.5 Conclusions

This chapter has proposed GCN based transfer learning method for cross-dataset MI EEG decoding. The proposed model combines both the CNN and GCN layers, aggregating topological information from 62 channels into only 8 specific channels and guiding a pre-trained model by knowledge distillation. Fine-tuning technology has been used to adapt the target dataset. The results show that the proposed model achieved an accuracy of 71.19% based on across-dataset, 7.04% higher than the state-of-the-art approaches. The feature visualization and heatmaps indicate excellent performance of the proposed model on EEG decoding and channel relation reconstruction, demonstrating its potential to enhance the effectiveness of the BCI applications with portable devices.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This thesis has developed new deep learning methods to decode MI-EEG for rehabilitation. Decoding MI signals aids in identifying patients' motor intentions. The classification results can be utilized for controlling external devices or as neural feedback to encourage active participation in rehabilitation training. Moreover, previous research has indicated that higher classification accuracy correlates with better rehabilitation outcomes. Therefore, achieving high accuracy in MI task classification is crucial. This study has primarily focused on enhancing model performance through deep learning and enabling the model to be effective in various real-world application scenarios.

7.1 Conclusions

The thesis has addressed several key issues such as poor accuracy, limited generalization, and inadequate practicality in EEG decoding. Corresponding models and frameworks have been proposed to tackle these challenges. Experimental results have confirmed that the proposed methods enhance the performance of EEG decoding models which are validated across different datasets or application scenarios, further improving the

practicality of MI-BCI in real-world applications and providing more effective assistance for patient rehabilitation. The summary of the work is presented in the following aspects:

In Chapter 2, the principles of MI-BCI and EEG decoding algorithms are introduced which aim at the importance of MI task recognition for rehabilitation. The key is to design a model with high classification performance, excellent generalization ability, and practicality such as less calibration time or training costs. Then, a comprehensive review of the traditional machine learning methods and deep learning methods is presented.

In Chapter 3, a novel approach for encoding MI-EEG signals using a multi-view CNN architecture is presented. Initially, multiple frequency sub-band MI-EEG signals are generated by employing bandpass filters that target specific brain rhythms, serving as inputs for our CNN model. Subsequently, temporal and spatial features are extracted from both the entire frequency band and the filtered sub-band signals. Moreover, leveraging two dense blocks with multi-CNN layers enhances the learning capabilities of the model and promotes efficient information propagation. The proposed method achieves an average accuracy of 75.16% on the publicly available Korea University EEG dataset, which comprises data from 54 healthy subjects performing two-class motor imagery tasks.

Chapter 4 is a continual work in building the within-subject model for MI classification. Compared with the research in Chapter 3, the new model harnesses a local and global Transformer decoders to make up for the shortcomings of the CNN model. The integration of a global transformer encoder with a Densely Connected Network is proposed to enhance information flow and reuse within the model. Spatial features from all channels, as well as differences between hemispheres, are incorporated to bolster the model's robustness. Three experimental scenarios, namely within-session, cross-session, and two-session, are designed. Results have demonstrated that compared to current state-of-the-art models, the proposed approach yields accuracy improvements of up to 1.46%, 7.49%, and 7.46% in the respective scenarios using the public Korean dataset. Additionally, for the BCI-IV-2a dataset, the proposed model achieves improvements of 2.12% and 2.21% in the cross-session and two-session scenarios, respectively.

Chapter 5 focuses on cross-subject modeling to achieve plug-and-play functionality in the MI-BCI system. In this research, domain-invariant features from source subjects are extracted and a knowledge distillation framework is employed to acquire internally invariant representations by fusing spectral features. Subsequently, a correlation alignment approach is utilized to align mutually invariant representations across each pair of sub-source domains. Additionally, distance regularization is applied to two types of invariant features to enhance generalizable information. Experimental results show that compared to current state-of-the-art models, our proposed approach achieves accuracy improvements of 8.93% and 4.4% on the public Korean dataset and BCI-IV-2a dataset, respectively.

Chapter 6 focuses on cross-dataset modeling to achieve good classification results with a model on the device with high impedance, poor collected data quality, and limited electrode channels. In this research, a GCN for cross-device MI-EEG decoding is proposed, utilizing transfer learning techniques. By leveraging multi-channel information, the GCN module aggregates topological features. A knowledge distillation framework is adopted to guide the pre-trained model with few-channel signals as inputs. Subsequently, we adapt the model to the few-channel dataset using a transfer learning strategy with dynamically inputting different amounts of data for training. When using only 40% of the training data, the proposed model achieved an accuracy of 67.77%, surpassing the baseline model's 64.15% accuracy obtained with 60% of the data for training. Experimental results indicate a significant improvement in accuracy, up to 7.04%, compared to state-of-the-art models. These findings underscore the effectiveness of our approach in cross-dataset MI-EEG decoding, thereby enhancing the practicality of MI-BCI applications.

Overall, this thesis have proposed several deep learning models for improving MI task classification accuracy, model generalization ability, and practicality.

7.2 Future Work

Although the proposed models have been proven useful for MI-EEG decoding and validated on different datasets and scenarios, there are still several challenges that require further investigation.

The model in Chapter 3 incorporates five sets of inputs, extracting features individually from each set of input signals. Despite utilizing average pooling layers and 1×1 CNN to reduce computed feature maps, training the complex model remains time-consuming. Enhancing model compactness entails identifying the most relevant rhythms during training and pruning redundant layers and neurons. Therefore, future endeavors will focus on identifying the most suitable sub-bands and implementing neuron pruning methods to reduce model size. Meanwhile, the proposed model leverages both temporal and spatial features of MI-EEG signals, with channel selection primarily informed by previous studies and experience. Nonetheless, disparate subjects may exhibit variance in motor imagery areas. Hence, future efforts will introduce automatic channel selection methods to enhance the model's adaptability.

The proposed model in Chapter 4 solely incorporates the transformer encoder for processing time-series data, overlooking potential spatial features extracted through the self-attention mechanism. This omission is deliberate due to the extensive length of the sequence after extracting temporal features. Utilizing a transformer to learn correlations between features of each channel and replacing deepwise convolutions could lead to severe overfitting issues. Additionally, while the complexity of our model based on trainable parameters is not high, computational time remains significant due to the time-consuming nature of the local transformer sliding to process inputs similar to CNNs. Moreover, in contrast to other transformer-based models, the proposed model does not consider the selection of position encoding methods. Hence, future work will explore more efficient model structures.

In Chapter 5, real-world problems entail not only variations across subjects but also practical demands across diverse scenarios and devices. Models should not solely learn domain-invariant features at a high-level abstraction but also engage in optimization and

weight redistribution learning across channels or time periods. Moreover, the proposed model only involves testing model performance and visualizing feature distribution. In the future, interpretable techniques can be utilized in deep learning models to elucidate invariant features and propose their specific physical meanings, thereby mutually corroborating them with relevant neural mechanisms.

In Chapter 6, while the adjacency matrix was trainable to dynamically capture topological information, the initialization of the graph solely relied on the Pearson correlation coefficient, lacking prior knowledge such as the relationship between the motor brain area and other brain areas, as well as individual channel connectivity. Despite employing knowledge distillation, the transfer of knowledge from graph convolutions to temporal-spatial features remains insufficient. Further investigation is necessary to enhance the integration of multi-channel information into a reduced number of channels. Moreover, the model attains an accuracy of over 67% with minimal data from the target domain during training, but it has not fully achieved zero-shot learning. Improvements are required to enhance the model's generalization ability and robustness without the need for additional data.

Journal Papers

1. J. Zhang and K. Li, "A multi-view cnn encoding for motor imagery eeg signals," *Biomedical Signal Processing and Control*, vol. 85, p. 105063, 2023

2. J. Zhang, K. Li, B. Yang, and X. Han, "Local and global convolutional transformer-based motor imagery eeg classification," *Frontiers in Neuroscience*, vol. 17, 2023

Conference Paper

1. J. Zhang and K. Li, "A pruned deep learning approach for classification of motor imagery electroencephalography signals," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 4072–4075

BIBLIOGRAPHY

[1] K. Blinowska and P. Durka, "Electroencephalography (eeg)," *Wiley encyclopedia of biomedical engineering*, 2006.

[2] M. Diykh, Y. Li, and S. Abdulla, "Eeg sleep stages identification based on weighted undirected complex networks," *Computer methods and programs in biomedicine*, vol. 184, p. 105116, 2020.

[3] P. Boonyakitanont, A. Lek-Uthai, K. Chomtho, and J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using eeg," *Biomedical Signal Processing and Control*, vol. 57, p. 101702, 2020.

[4] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive workload recognition using eeg signals and machine learning: A review," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[5] A. H. Meghdadi, M. Stevanović Karić, M. McConnell, G. Rupp, C. Richard, J. Hamilton, D. Salat, and C. Berka, "Resting state eeg biomarkers of cognitive decline associated with alzheimer's disease and mild cognitive impairment," *PloS one*, vol. 16, no. 2, p. e0244180, 2021.

[6] R. Mane, T. Chouhan, and C. Guan, "Bci for stroke rehabilitation: motor and beyond," *Journal of neural engineering*, vol. 17, no. 4, p. 041001, 2020.

[7] N. Vivaldi, M. Caiola, K. Solarana, and M. Ye, "Evaluating performance of eeg data-driven machine learning for traumatic brain injury classification," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 11, pp. 3205–3216, 2021.

[8] S. Saha, K. A. Mamun, K. Ahmed, R. Mostafa, G. R. Naik, S. Darvishi, A. H. Khandoker, and M. Baumert, "Progress in brain computer interface: Challenges and opportunities," *Frontiers in Systems Neuroscience*, vol. 15, p. 578875, 2021.

[9] K. K. Ang and C. Guan, "Brain-computer interface in stroke rehabilitation," *Journal of Computing Science and Engineering*, vol. 7, no. 2, pp. 139–146, 2013.

[10] R. Scherer and C. Vidaurre, "Motor imagery based brain–computer interfaces," in *Smart wheelchairs and brain-computer interfaces*. Elsevier, 2018, pp. 171–195.

[11] J. Decety and D. H. Ingvar, "Brain structures participating in mental simulation of motor behavior: A neuropsychological interpretation," *Acta psychologica*, vol. 73, no. 1, pp. 13–34, 1990.

[12] O. Mokienko, L. Chernikova, A. Frolov, and P. Bobrov, "Motor imagery and its practical application," *Neuroscience and Behavioral Physiology*, vol. 44, no. 5, pp. 483–489, 2014.

[13] A. Vourvopoulos, S. Bermudez i Badia, and F. Liarokapis, "Eeg correlates of video game experience and user profile in motor-imagery-based brain–computer interaction," *The Visual Computer*, vol. 33, pp. 533–546, 2017.

[14] J. H. Gruzelier, "Eeg-neurofeedback for optimising performance. i: A review of cognitive and affective outcome in healthy participants," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 124–141, 2014.

[15] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: communication and restoration of movement in paralysis," *The Journal of physiology*, vol. 579, no. 3, pp. 621–636, 2007.

[16] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by eeg/meg: methodological, practical and conceptual challenges," *Signal processing*, vol. 85, no. 11, pp. 2190–2212, 2005.

[17] K. K. Ang, C. Guan, K. S. Phua, C. Wang, L. Zhou, K. Y. Tang, G. J. Ephraim Joseph, C. W. K. Kuah, and K. S. G. Chua, "Brain-computer interface-based robotic end effector system for wrist and hand rehabilitation: results of a three-armed randomized controlled trial for chronic stroke," *Frontiers in neuroengineering*, vol. 7, p. 30, 2014.

[18] G. J. Hankey, "The global and regional burden of stroke," *The lancet global health*, vol. 1, no. 5, pp. e239–e240, 2013.

[19] P. W. Duncan, L. B. Goldstein, D. Matchar, G. W. Divine, and J. Feussner, "Measurement of motor recovery after stroke. outcome assessment and sample size requirements." *Stroke*, vol. 23, no. 8, pp. 1084–1089, 1992.

[20] F. Hummel, P. Celnik, P. Giraux, A. Floel, W.-H. Wu, C. Gerloff, and L. G. Cohen, "Effects of non-invasive cortical stimulation on skilled motor function in chronic stroke," *Brain*, vol. 128, no. 3, pp. 490–499, 2005.

[21] J. C. Grotta, E. A. Noser, T. Ro, C. Boake, H. Levin, J. Aronowski, and T. Schallert, "Constraint-induced movement therapy," *Stroke*, vol. 35, no. 11_suppl_1, pp. 2699–2701, 2004.

[22] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, p. 1423, 2019.

[23] D. Elstob and E. L. Secco, "A low cost eeg based bci prosthetic using motor imagery," *arXiv preprint arXiv:1603.02869*, 2016.

[24] G. Müller-Putz, P. Ofner, A. Schwarz, J. Pereira, G. Luzhnica, C. di Sciascio, E. Veas, S. Stein, J. Williamson, and R. M. Murray-Smith, "Restoration of upper limb function in individuals with high spinal cord injury by multimodal neuroprostheses for interaction in daily activities," in *Proceedings of the 7th Graz Brain-Computer Interface Conference, Graz, Austria*, 2017, pp. 18–22.

[25] J. Feng, I. Spence, and J. Pratt, "Playing an action video game reduces gender differences in spatial cognition," *Psychological science*, vol. 18, no. 10, pp. 850–855, 2007.

[26] T. Li, J. Zhang, T. Xue, and B. Wang, "Development of a novel motor imagery control technique and application in a gaming environment," *Computational intelligence and neuroscience*, vol. 2017, no. 1, p. 5863512, 2017.

[27] I. Martišius and R. Damaševičius, "A prototype ssvep based real time bci gaming system," *Computational intelligence and neuroscience*, vol. 2016, no. 1, p. 3861425, 2016.

[28] Z. Wang, Y. Yu, M. Xu, Y. Liu, E. Yin, and Z. Zhou, "Towards a hybrid bci gaming paradigm based on motor imagery and ssvep," *International Journal of Human–Computer Interaction*, vol. 35, no. 3, pp. 197–205, 2019.

[29] C. M. Krause, A. H. Lang, M. Laine, M. Kuusisto, and B. Pörn, "Event-related. eeg desynchronization and synchronization during an auditory memory task," *Electroencephalography and clinical neurophysiology*, vol. 98, no. 4, pp. 319–326, 1996.

[30] G. Pfurtscheller, "Eeg event-related desynchronization (erd) and synchronization (ers)," *Electroencephalography and Clinical Neurophysiology*, vol. 1, no. 103, p. 26, 1997.

[31] G. Pfurtscheller and F. L. Da Silva, "Event-related eeg/meg synchronization and desynchronization: basic principles," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[32] S. K. Lal and A. Craig, "Electroencephalography activity associated with driver fatigue: Implications for a fatigue countermeasure device." *Journal of Psychophysiology*, vol. 15, no. 3, p. 183, 2001.

[33] P. Krishnan, S. Yaacob, A. P. Krishnan, M. Rizon, and C. K. Ang, "Eeg based drowsiness detection using relative band power and short-time fourier transform." *J. Robotics Netw. Artif. Life*, vol. 7, no. 3, pp. 147–151, 2020.

[34] J. Drummond, C. Brann, D. Perkins, and D. Wolfe, "A comparison of median frequency, spectral edge frequency, a frequency band power ratio, total power, and dominance shift in the determination of depth of anesthesia," *Acta Anaesthesiologica Scandinavica*, vol. 35, no. 8, pp. 693–699, 1991.

[35] J. D. Williams and J. H. Gruzelier, "Differentiation of hypnosis and relaxation by analysis of narrow band theta and alpha frequencies," *International Journal of Clinical and Experimental Hypnosis*, vol. 49, no. 3, pp. 185–206, 2001.

[36] A. Wróbel *et al.*, "Beta activity: a carrier for visual attention," *Acta neurobiologiae experimentalis*, vol. 60, no. 2, pp. 247–260, 2000.

[37] C. S. Herrmann, M. H. Munk, and A. K. Engel, "Cognitive functions of gamma-band activity: memory match and utilization," *Trends in cognitive sciences*, vol. 8, no. 8, pp. 347–355, 2004.

[38] H. Jasper and W. Penfield, "Electrocorticograms in man: effect of voluntary movement upon the electrical activity of the precentral gyrus," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 183, no. 1, pp. 163–174, 1949.

[39] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal, "Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14 681–14 722, 2023.

[40] L. Tong, Y. Qian, L. Peng, C. Wang, and Z.-G. Hou, "A learnable eeg channel selection method for mi-bci using efficient channel attention," *Frontiers in Neuroscience*, vol. 17, p. 1276067, 2023.

[41] P. Gaur, K. McCreadie, R. B. Pachori, H. Wang, and G. Prasad, "An automatic subject specific channel selection method for enhancing motor imagery classification in eeg-bci using correlation," *Biomedical Signal Processing and Control*, vol. 68, p. 102574, 2021.

[42] J. Jin, C. Liu, I. Daly, Y. Miao, S. Li, X. Wang, and A. Cichocki, "Bispectrum-based channel selection for motor imagery based brain-computer interfacing," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2153–2163, 2020.

[43] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3d convolutional neural network for eeg-based motor imagery classification," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 27, no. 10, pp. 2164–2177, 2019.

[44] O. Avilov, S. Rimbert, A. Popov, and L. Bougrain, "Optimizing motor intention detection with deep learning: towards management of intraoperative awareness," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 3087–3097, 2021.

[45] M. Riyad, M. Khalil, and A. Adib, "Mi-eegnet: A novel convolutional neural network for motor imagery classification," *Journal of Neuroscience Methods*, vol. 353, p. 109037, 2021.

[46] N. Shajil, S. Mohan, P. Srinivasan, J. Arivudaiyanambi, and A. Arasappan Murrugesan, "Multiclass classification of spatially filtered motor imagery eeg signals using convolutional neural network for bci based applications," *Journal of Medical and Biological Engineering*, vol. 40, pp. 663–672, 2020.

[47] T.-W. Lee and T.-W. Lee, *Independent component analysis*. Springer, 1998.

[48] J. Yang, Z. Ma, J. Wang, and Y. Fu, "A novel deep learning scheme for motor imagery eeg decoding based on spatial representation fusion," *IEEE Access*, vol. 8, pp. 202 100–202 110, 2020.

[49] J.-H. Jeong, B.-H. Lee, D.-H. Lee, Y.-D. Yun, and S.-W. Lee, "Eeg classification of forearm movement imagery using a hierarchical flow convolutional neural network," *IEEE Access*, vol. 8, pp. 66 941–66 950, 2020.

[50] J. Yang, S. Yao, and J. Wang, "Deep fusion feature learning network for mi-eeg classification," *Ieee Access*, vol. 6, pp. 79 050–79 059, 2018.

[51] B. Xu, L. Zhang, A. Song, C. Wu, W. Li, D. Zhang, G. Xu, H. Li, and H. Zeng, "Wavelet transform time-frequency image and convolutional network-based motor imagery eeg classification," *IEEE Access*, vol. 7, pp. 6084–6093, 2018.

[52] X. Ma, S. Qiu, W. Wei, S. Wang, and H. He, "Deep channel-correlation network for motor imagery decoding from the same limb," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 297–306, 2019.

[53] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[54] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains," *International Scholarly Research Notices*, vol. 2014, 2014.

[55] W. J. Freeman, M. D. Holmes, B. C. Burke, and S. Vanhatalo, "Spatial spectra of scalp eeg and emg from awake humans," *Clinical Neurophysiology*, vol. 114, no. 6, pp. 1053–1068, 2003.

[56] A. H. Jahidin, M. M. Ali, M. N. Taib, N. M. Tahir, I. M. Yassin, and S. Lias, "Classification of intelligence quotient via brainwave sub-band power ratio features and artificial neural network," *Computer methods and programs in biomedicine*, vol. 114, no. 1, pp. 50–59, 2014.

[57] P. Singh, S. D. Joshi, R. K. Patney, and K. Saha, "The fourier decomposition method for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2199, p. 20160871, 2017.

[58] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2013.

[59] C.-J. Peng, Y.-C. Chen, C.-C. Chen, S.-J. Chen, B. Cagneau, and L. Chassagne, "An eeg-based attentiveness recognition system using hilbert–huang transform and support vector machine," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 230–238, 2020.

[60] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[61] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[62] J. Pardey, S. Roberts, and L. Tarassenko, "A review of parametric modelling techniques for eeg analysis," *Medical engineering & physics*, vol. 18, no. 1, pp. 2–11, 1996.

[63] Y. Zhang, B. Liu, X. Ji, and D. Huang, "Classification of eeg signals based on autoregressive model and wavelet packet decomposition," *Neural Processing Letters*, vol. 45, pp. 365–378, 2017.

[64] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "Eeg-based bci emotion recognition: A survey," *Sensors*, vol. 20, no. 18, p. 5083, 2020.

[65] P. Jahankhani, V. Kodogiannis, and K. Revett, "Eeg signal classification using wavelet feature extraction and neural networks," in *IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06)*. IEEE, 2006, pp. 120–124.

[66] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background eeg," *Brain topography*, vol. 2, no. 4, pp. 275–284, 1990.

[67] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial eeg," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[68] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Optimizing spatio-temporal filters for improving brain-computer interfacing," *Advances in Neural Information Processing Systems*, vol. 18, 2005.

[69] Q. Novi, C. Guan, T. H. Dat, and P. Xue, "Sub-band common spatial pattern (sbcsp) for brain-computer interface," in *2007 3rd International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2007, pp. 204–207.

[70] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 2390–2397.

[71] H. Higashi and T. Tanaka, "Simultaneous design of fir filter banks and spatial patterns for eeg signal classification," *IEEE transactions on biomedical engineering*, vol. 60, no. 4, pp. 1100–1110, 2012.

[72] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI communications*, vol. 30, no. 2, pp. 169–190, 2017.

[73] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[74] C. M. Bishop, "Neural networks and their applications," *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.

[75] M. Bentlemsan, E.-T. Zemouri, D. Bouchaffra, B. Yahya-Zoubir, and K. Ferroudji, "Random forest and filter bank common spatial patterns for eeg-based motor imagery classification," in *2014 5th International conference on intelligent systems, modelling and simulation*. IEEE, 2014, pp. 235–238.

[76] J. Luo, Z. Feng, J. Zhang, and N. Lu, "Dynamic frequency feature selection based approach for classification of motor imageries," *Computers in biology and medicine*, vol. 75, pp. 45–53, 2016.

[77] S. Aggarwal and N. Chugh, "Signal processing techniques for motor imagery brain computer interface: A review," *Array*, vol. 1, p. 100003, 2019.

[78] C.-Y. Chen, C.-W. Wu, C.-T. Lin, and S.-A. Chen, "A novel classification method for motor imagery based on brain-computer interface," in *2014 International joint conference on neural networks (IJCNN)*. IEEE, 2014, pp. 4099–4102.

[79] R. Fu, Y. Tian, T. Bao, Z. Meng, and P. Shi, "Improvement motor imagery eeg classification based on regularized linear discriminant analysis," *Journal of medical systems*, vol. 43, no. 6, pp. 1–13, 2019.

[80] M. R. Islam, T. Tanaka, M. S. Akter, and M. K. I. Molla, "Classification of motor imagery bci using multiband tangent space mapping," in *2017 22nd International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[81] G. Sagee and S. Hema, "Eeg feature extraction and classification in multiclass multiuser motor imagery brain computer interface u sing bayesian network and ann," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. IEEE, 2017, pp. 938–943.

[82] M. Hamedi, S.-H. Salleh, A. M. Noor, and I. Mohammad-Rezazadeh, "Neural network-based three-class motor imagery classification using time-domain

features for bci applications," in *2014 IEEE region 10 symposium*. IEEE, 2014, pp. 204–207.

[83] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[84] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, "A survey on deep learning based brain computer interface: Recent advances and new frontiers," *arXiv preprint arXiv:1905.04149*, p. 66, 2019.

[85] 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*, vol. 38, no. 11, pp. 5391–5420, 2017.

[86] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[87] F. Li, F. He, F. Wang, D. Zhang, Y. Xia, and X. Li, "A novel simplified convolutional neural network classification algorithm of motor imagery eeg signals based on deep learning," *Applied Sciences*, vol. 10, no. 5, p. 1605, 2020.

[88] P. Kant, S. H. Laskar, J. Hazarika, and R. Mahamune, "Cwt based transfer learning for motor imagery classification for brain computer interfaces," *Journal of Neuroscience Methods*, vol. 345, p. 108886, 2020.

[89] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, "Eeg classification of motor imagery using a novel deep learning framework," *Sensors*, vol. 19, no. 3, p. 551, 2019.

[90] W. Huang, Y. Xue, L. Hu, and H. Liuli, "S-eegnet: Electroencephalogram signal classification based on a separable convolution neural network with bilinear interpolation," *IEEE Access*, vol. 8, pp. 131 636–131 646, 2020.

[91] M. Dätig and T. Schlurmann, "Performance and limitations of the hilbert–huang transformation (hht) with an application to irregular water waves," *Ocean Engineering*, vol. 31, no. 14-15, pp. 1783–1834, 2004.

[92] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion," *Future Generation computer systems*, vol. 101, pp. 542–554, 2019.

[93] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multi-view cnn with novel variance layer for motor imagery brain computer interface," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 2950–2953.

[94] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "Eeg dataset and openbmi toolbox for three bci paradigms: an investigation into bci illiteracy," *GigaScience*, vol. 8, no. 5, p. giz002, 2019.

[95] C. Ju and C. Guan, "Tensor-cspnet: A novel geometric deep learning framework for motor imagery classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[96] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[97] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[98] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[99] S. U. Amin, H. Altaheri, G. Muhammad, W. Abdul, and M. Alsulaiman, "Attention-inception and long-short-term memory-based electroencephalography classification for motor imagery tasks in rehabilitation," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5412–5421, 2021.

[100] C. Zhang, Y.-K. Kim, and A. Eskandarian, "Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046014, 2021.

[101] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[102] M.-A. Li and D.-Q. Xu, "A transfer learning method based on vgg-16 convolutional neural network for mi classification," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 2021, pp. 5430–5435.

[103] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, M. Liu, F. Chen, and W. Che, "A deep transfer convolutional neural network framework for eeg signal classification," *IEEE Access*, vol. 7, pp. 112 767–112 776, 2019.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[105] Z. Khademi, F. Ebrahimi, and H. M. Kordy, "A transfer learning-based cnn and lstm hybrid deep learning model to classify motor imagery eeg signals," *Computers in biology and medicine*, vol. 143, p. 105288, 2022.

[106] S. Jia, Y. Hou, Y. Shi, and Y. Li, "Attention-based graph resnet for motor intent detection from raw eeg signals," *arXiv preprint arXiv:2007.13484*, 2020.

[107] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[108] A. Fernandez, R. B. H. Bunke, and J. Schmiduber, "A novel connectionist system for improved unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, 2009.

[109] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[110] Z. Tayeb, J. Fedjaev, N. Ghaboosi, C. Richter, L. Everding, X. Qu, Y. Wu, G. Cheng, and J. Conradt, "Validating deep neural networks for online decoding of motor imagery movements from eeg signals," *Sensors*, vol. 19, no. 1, p. 210, 2019.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[112] Y. Du, Y. Xu, X. Wang, L. Liu, and P. Ma, "Eeg temporal–spatial transformer for person identification," *Scientific Reports*, vol. 12, no. 1, p. 14378, 2022.

[113] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, and X. Chen, "Eeg-based emotion recognition via transformer neural architecture search," *IEEE Transactions on Industrial Informatics*, 2022.

[114] S. Bagchi and D. R. Bathula, "Eeg-convtransformer for single-trial eeg-based visual stimulus classification," *Pattern Recognition*, vol. 129, p. 108757, 2022.

[115] X. Pu, P. Yi, K. Chen, Z. Ma, D. Zhao, and Y. Ren, "Eegdnet: Fusing non-local and local self-similarity for eeg signal denoising with transformer," *Computers in Biology and Medicine*, vol. 151, p. 106248, 2022.

[116] Y. Ma, Y. Song, and F. Gao, "A novel hybrid cnn-transformer model for eeg motor imagery classification," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[117] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[118] Y. Tao, T. Sun, A. Muhamed, S. Genc, D. Jackson, A. Arsanjani, S. Yaddanapudi, L. Li, and P. Kumar, "Gated transformer for decoding human brain eeg signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 125–130.

[119] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2126–2136, 2022.

[120] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in eeg signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, 2021.

[121] R. Santana, L. Marti, and M. Zhang, "Gp-based methods for domain adaptation: using brain decoding across subjects as a test-case," *Genetic Programming and Evolvable Machines*, vol. 20, no. 3, pp. 385–411, 2019.

[122] P. Wang, J. Lu, B. Zhang, and Z. Tang, "A review on transfer learning for brain-computer interface classification," in *2015 5th International Conference on Information Science and Technology (ICIST)*. IEEE, 2015, pp. 315–322.

[123] H. H. Jasper and H. L. Andrews, "Electro-encephalography: Iii. normal differentiation of occipital and precentral regions in man," *Archives of Neurology & Psychiatry*, vol. 39, no. 1, pp. 96–115, 1938.

[124] M. Ahn, H. Cho, S. Ahn, and S. C. Jun, "High theta and low alpha powers may be indicative of bci-illiteracy in motor imagery," *PloS one*, vol. 8, no. 11, p. e80886, 2013.

[125] L. R. Trambaiolli, P. J. Dean, A. M. Cravo, A. Sterr, and J. R. Sato, "On-task theta power is correlated to motor imagery performance," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3937–3942.

[126] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[127] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648*, 2016.

[128] T. Liu and D. Yang, "A densely connected multi-branch 3d convolutional neural network for motor imagery eeg decoding," *Brain Sciences*, vol. 11, no. 2, p. 197, 2021.

[129] Z. Yu, W. Chen, and T. Zhang, "Motor imagery eeg classification algorithm based on improved lightweight feature fusion network," *Biomedical Signal Processing and Control*, vol. 75, p. 103618, 2022.

[130] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 10, pp. 3839–3852, 2019.

[131] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in neuroscience*, p. 55, 2012.

[132] 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*, vol. 15, no. 5, p. 056013, 2018.

[133] G. Dai, J. Zhou, J. Huang, and N. Wang, "Hs-cnn: a cnn with hybrid convolution scale for eeg motor imagery classification," *Journal of neural engineering*, vol. 17, no. 1, p. 016025, 2020.

[134] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[135] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[136] R. Zhang, Q. Zong, L. Dou, and X. Zhao, "A novel hybrid deep learning scheme for four-class motor imagery classification," *Journal of neural engineering*, vol. 16, no. 6, p. 066004, 2019.

[137] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[138] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[139] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set a," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[140] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network," *Neural Networks*, vol. 136, pp. 1–10, 2021.

[141] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[142] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[143] M. Martin, K. Nitschke, L. Beume, A. Dressing, L. E. Bühler, V. M. Ludwig, I. Mader, M. Rijntjes, C. P. Kaller, and C. Weiller, "Brain activity underlying tool-related and imitative skills after major left hemisphere stroke," *Brain*, vol. 139, no. 5, pp. 1497–1516, 2016.

[144] R.-A. Müller, R. D. Rothermel, M. E. Behen, O. Muzik, T. J. Mangner, and H. T. Chugani, "Differential patterns of language and motor reorganization following early left hemisphere lesion: a pet study," *Archives of Neurology*, vol. 55, no. 8, pp. 1113–1119, 1998.

[145] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2011.

[146] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[147] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[148] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[149] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2023.

[150] T. Grubinger, A. Birlutiu, H. Schöner, T. Natschläger, and T. Heskes, "Domain generalization based on transfer component analysis," in *Advances in Computational Intelligence: 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part I 13*. Springer, 2015, pp. 325–334.

[151] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010.

[152] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 1. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 10–18. [Online]. Available: https://proceedings.mlr.press/v28/muandet13.html

[153] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1414–1430, 2017.

[154] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[155] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[156] D. Freer and G.-Z. Yang, "Data augmentation for self-paced motor imagery classification with c-lstm," *Journal of neural engineering*, vol. 17, no. 1, p. 016041, 2020.

[157] I. Raoof and M. K. Gupta, "Domain-independent short-term calibration based hybrid approach for motor imagery electroencephalograph classification: a comprehensive review," *Multimedia Tools and Applications*, pp. 1–46, 2023.

[158] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, "Domain-invariant feature exploration for domain generalization," *arXiv preprint arXiv:2207.12020*, 2022.

[159] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[160] J. Wang, L. Yao, and Y. Wang, "Ifnet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1900–1911, 2023.

[161] F. Wei, X. Xu, T. Jia, D. Zhang, and X. Wu, "A multi-source transfer joint matching method for inter-subject motor imagery decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1258–1267, 2023.

[162] Y. Liang and Y. Ma, "Calibrating eeg features in motor imagery classification tasks with a small amount of current data using multisource fusion transfer learning," *Biomedical Signal Processing and Control*, vol. 62, p. 102101, 2020.

[163] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[164] F. Liu, P. Yang, Y. Shu, N. Liu, J. Sheng, J. Luo, X. Wang, and Y.-J. Liu, "Emotion recognition from few-channel eeg signals by integrating deep feature aggregation and transfer learning," *IEEE Transactions on Affective Computing*, 2023.

[165] M. Soufineyestani, D. Dowling, and A. Khan, "Electroencephalography (eeg) technology applications and available devices," *Applied Sciences*, vol. 10, no. 21, p. 7453, 2020.

[166] T. Zaremba and A. Atyabi, "Cross-subject & cross-dataset subject transfer in motor imagery bci systems," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[167] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming, "Cross-dataset variability problem in eeg decoding with deep learning," *Frontiers in human neuroscience*, vol. 14, p. 103, 2020.

[168] Y. Xie, K. Wang, J. Meng, J. Yue, L. Meng, W. Yi, T.-P. Jung, M. Xu, and D. Ming, "Cross-dataset transfer learning for motor imagery signal classification via multi-task learning and pre-training," *Journal of Neural Engineering*, vol. 20, no. 5, p. 056037, 2023.

[169] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian procrustes analysis: Transfer learning for brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2390–2401, 2018.

[170] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[171] W. Ma, C. Wang, X. Sun, X. Lin, and Y. Wang, "A double-branch graph convolutional network based on individual differences weakening for motor imagery eeg classification," *Biomedical Signal Processing and Control*, vol. 84, p. 104684, 2023.

[172] H. Wang, H. Yu, and H. Wang, "Eeg_genet: A feature-level graph embedding method for motor imagery classification based on eeg signals," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 3, pp. 1023–1040, 2022.

[173] V. Delvigne, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, "Phydaa: Physiological dataset assessing attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2612–2623, 2021.

[174] Y. Hou, S. Jia, X. Lun, Z. Hao, Y. Shi, Y. Li, R. Zeng, and J. Lv, "Gcns-net: a graph convolutional neural network approach for decoding time-resolved eeg motor imagery signals," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[175] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[176] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "Fbcnet: A multi-view convolutional neural network for brain-computer interface," *arXiv preprint arXiv:2104.01233*, 2021.

[177] B. Sun, Z. Liu, Z. Wu, C. Mu, and T. Li, "Graph convolution neural network based end-to-end channel selection and classification for motor imagery brain-computer interfaces," *IEEE transactions on industrial informatics*, 2022.

[178] J. Zhang and K. Li, "A multi-view cnn encoding for motor imagery eeg signals," *Biomedical Signal Processing and Control*, vol. 85, p. 105063, 2023.

[179] J. Zhang, K. Li, B. Yang, and X. Han, "Local and global convolutional transformer-based motor imagery eeg classification," *Frontiers in Neuroscience*, vol. 17, 2023.

[180] J. Zhang and K. Li, "A pruned deep learning approach for classification of motor imagery electroencephalography signals," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 4072–4075.