Interpretable and Robust AI in Electroencephalogram Systems

ZHOU XINLIANG

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

15th July 2024

Date

ZHOU XINLIANG

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

17th July 2024

Date

Prof. Liu Yang

Authorship Attribution Statement

This thesis contains material from 6 paper(s) published/ under review/ pre-printed in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 2 is partly published as **Xinliang Zhou**, Chenyu Liu, Liming Zhai, Ziyu Jia, Cuntai Guan, Yang Liu. "Interpretable and Robust AI in EEG Systems: A Survey". in **https://doi.org/10.48550/arXiv.2304.10755**, 2023 (preprint)

The contributions of the co-authors are as follows:

- Prof. Yang Liu and Prof. Cuntai Guan provided the initial project direction and edited the manuscript drafts.
- I co-designed the survey study with Mr. Chenyu Liu, and I prepared the manuscript drafts.
- Assis Prof. Ziyu Jia assisted in the selected paper summarizing and analysis, and I organized a long-updated link for related papers.
- Mr. Chenyu Liu and Assoc Prof. Liming Zhai revised the manuscript together.

Chapter 3 is partly published as **Xinliang Zhou**, Chenyu Liu, Jiaping Xiao, Yang Liu,"EEG-based Sleep Staging with Hybrid Attention", In **2023 IEEE Conference on Artificial Intelligence (CAI 2023)**. DOI: 10.1109/CAI54212.2023.00055

The contributions of the co-authors are as follows:

- Prof. Yang Liu provided the initial project direction and edited the manuscript drafts.
- I prepared the drafts of the manuscript. Mr. Chenyu Liu and Mr. Jiaping Xiao revised the manuscript together.
- I performed the data investigation, preprocessing, and conducted data evaluation.
- I designed the neural networks and conducted the experiments as well as the analysis.

Chapter 4 is partly published as (1) Chenyu Liu, **Xinliang Zhou**, Yang Liu, "EENED: End-to-End Neural Epilepsy Detection based on Convolutional Transformer", In **2023 IEEE Conference on Artificial Intelligence (CAI 2023)**.

DOI: 10.1109/CAI54212.2023.00161 and (2) **Xinliang Zhou**, Chenyu Liu, Ruizhi Yang, Liangwei Zhang, Liming Zhai, Ziyu Jia, Yang Liu, "Learning Robust Global-Local Representation from EEG for Neural Epilepsy Detection", In **IEEE Transactions on Artifical Intelligence (TAI)**, 2024. DOI: 10.1109/TAI.2024.3406289

The contributions of the co-authors are as follows:

- Prof. Yang Liu and Assis. Prof. Ziyu Jia provided the initial project direction and edited the manuscript drafts.
- I prepared the drafts of the manuscript. The manuscript was revised together with Mr. Ruizhi Yang and Mr. Liangwei Zhang.
- Mr. Chenyu Liu performed the data investigation, preprocessing, and conducted data evaluation.
- I designed the neural networks and conducted the experiments as well as the analysis. Assoc. Prof. Liming Zhai did the supporting experiment analysis.
- Mr. Ruizhi Yang double-checked my codes before we released them.

Chapter 5 is partly published as **Xinliang Zhou**, Dan Lin, Ziyu Jia, Jiaping Xiao, Chenyu Liu, Liming Zhai, Yang Liu, "An EEG Channel Selection Framework for Driver Drowsiness Detection via Interpretability Guidance", In **2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)**. DOI: 10.1109/EMBC40787.2023.10341126

The contributions of the co-authors are as follows:

- Prof. Yang Liu and Assis. Prof. Ziyu Jia provided the initial project direction and edited the manuscript drafts.
- I prepared the drafts of the manuscript. The manuscript was revised together with Dr. Dan Lin and Mr. Jiaping Xiao.
- I performed the data investigation, preprocessing, and conducted data evaluation.
- I conducted the analysis of the algorithm and the experiments. Assoc. Prof. Liming Zhai did the supporting experiment analysis.
- Mr. Chenyu Liu double-checked my code and proofread the manuscript.

Chapter 6 is partly published as **Xinliang Zhou**, Chenyu Liu, Jiaping Xiao, Liming Zhai, Ziyu Jia, Yang Liu, "Learning Relational Probabilistic Graphs for EEG-based Emotion Recognition", In **IEEE Transactions on Affective Computing (TAFFC)**, 2024 (under revision)

The contributions of the co-authors are as follows:

- Prof. Yang Liu and Assis. Prof. Ziyu Jia provided the initial project direction and edited the manuscript drafts.
- I prepared the drafts of the manuscript. The manuscript was revised together with Mr. Chenyu Liu and Mr. Jiaping Xiao.

- I performed the data investigation, preprocessing, and conducted data evaluation.
- I designed the neural networks and conducted the experiments as well as the analysis together with Mr. Chenyu Liu. Assoc. Prof. Liming Zhai did the supporting experiment analysis.
- Mr. Chenyu Liu double-checked my code and proofread the manuscript.

15th July 2024

Date

ZHOU XINLIANG

Acknowledgements

I wish to express my deepest gratitude to all those who have guided, supported, and helped me throughout my research and life journey.

First and foremost, I would like to thank my academic advisor and life mentor, Prof. Yang Liu. He generously provided me the opportunity to join his research team and patiently guided me in both my research and personal growth. He taught me the importance of responsibility and critical thinking. Without his generous mentorship, my research journey would have been far more difficult and challenging.

I would also like to thank my helpful, friendly, and trusted colleagues. In particular, I am grateful to Assis. Prof. Ziyu Jia, who greatly assisted me in building my domain knowledge and academic writing skills. I am also thankful to my fellow peers, Mr. Jiaping Xiao, Mr. Haoning Wu, and Mr. Chenyu Liu, for helping me develop as a researcher. My sincere thanks go to all my team members for providing such a harmonious and positive work environment. I am also grateful to my department and collaborators. The College of Computing and Data Science supplied ample learning resources that enabled my research.

Moreover, I wish to thank Prof. Cuntai Guan, Assoc. Prof. Liming Zhai, and their teams for the tremendous research and personal support they provided during my PhD journey.

Lastly, I am profoundly thankful to my dear family for their endless help and encouragement in my life decisions. I could not have reached this point without them. In particular, I wish to thank my beloved wife, Miss Chen Jing, for her unwavering love and companionship and for brightening my life. She helped me overcome difficult times and gave me the faith to persevere.

Thank you all!

Abstract

Electroencephalogram (EEG) provides valuable information about brain activities and states in a non-invasive way, making it a crucial research area in human-computer interaction (HCI). With the rapid advancement of artificial intelligence (AI) technologies, EEG systems have increasingly harnessed the power of AI for various clinical, entertainment, and social interaction applications. For example, sleep staging systems combine EEG signals with deep learning to assist physicians in the rapid diagnosis of sleep disorders. Driver monitoring systems employ EEG-based deep neural networks (DNNs) to accurately detect driver fatigue, thereby reducing the risk of car accidents. Additionally, robotic arm control systems use DNNs to translate human thoughts, as reflected by EEG signals, into control signals, enabling disabled individuals to perform basic tasks such as drinking water or moving objects.

Despite the significant progress driven by AI, models, particularly those based on deep learning, remain largely unexplainable due to their black-box nature. This lack of interpretability poses challenges in understanding and trusting the AI's decisions. Furthermore, these models are susceptible to both intentional and unintentional attacks, raising serious concerns about their robustness and reliability. Addressing these issues is crucial for AI's widespread adoption and safe deployment in EEG systems. Researchers are actively exploring methods to enhance the interpretability and robustness of AI models, ensuring they can provide reliable and transparent support in critical applications. As the field evolves, these advancements will be pivotal in realizing the full potential of AI-enhanced EEG systems for improving human life across various domains.

A comprehensive literature review on interpretable and robust AI techniques for EEG systems is presented in this thesis. It begins with an introduction to the background knowledge of EEG signals. Next, it proposes a taxonomy of interpretability, categorizing it into three types: backpropagation, perturbation, and rule-based methods. Additionally, it categorizes robustness based on undesirable

factors into four classes: noise and artifacts, human variability, data acquisition instability, and adversarial attacks. The literature review includes detailed analyses and comparisons for each category. Finally, it identifies several critical challenges for interpretable and robust AI in EEG systems and discusses their future directions. This literature review aims to guide researchers in understanding this field's latest advancements and future trends.

This thesis introduces a novel framework called the Hybrid Attention EEG Sleep Staging (HASS) Framework, designed for cross-subject EEG sleep staging tasks. HASS employs a spatio-temporal attention mechanism to adaptively assign weights to inter-channel and intra-channel EEG segments based on the spatio-temporal relationships of the brain during different sleep stages. Experimental results on the MASS and ISRUC datasets demonstrate that HASS can significantly improve typical sleep staging networks. HASS addresses the difficulties of capturing the spatial-temporal relationships of EEG signals during sleep staging under cross-subject scenarios and holds promise for improving the accuracy and reliability of sleep assessment in both clinical and research settings.

Furthermore, this thesis introduces EENED and GlepNet, novel EEG-based architectures for neural epilepsy detection. To address the challenge of learning robust global-local representations in epilepsy signals, EENED/GlepNet combines temporal convolutional layers with a multi-head attention mechanism. This approach enhances the performance of epilepsy diagnosis, potentially improving patient outcomes. The utilization of Grad-CAM for interpretability further elevates its clinical value, allowing healthcare professionals to validate and visually understand the model's diagnostic process. This advancement not only promises improved epilepsy detection but also contributes significantly to neuropsychological research and the application of machine learning in healthcare.

Additionally, an Interpretability-guided Channel Selection (ICS) framework is proposed for the EEG driver drowsiness detection task. ICS provides a two-stage training strategy to select the key contributing channels with interpretability guidance progressively. ICS trains a teacher network in the first stage using full-head channel EEG data. It then applies class activation mapping (CAM) to the trained teacher model to highlight the high-contributing EEG channels and proposes a channel voting scheme to select the top N contributing channels. In the second stage, ICS trains a student network with the selected channels of EEG data for

driver drowsiness detection. Experiments on a public dataset demonstrate that our method significantly improves the performance of cross-subject driver drowsiness detection.

Finally, by adopting relational thinking theory to transform raw EEG signals into probabilistic graphs, this thesis improves the decoding performance for EEG emotion classification tasks. The proposed method, the relational probabilistic graph convolutional network (RPGCN), effectively models variations in potential emotional states. RPGCN considers relationships among EEG channels and provides interpretability by explaining recognition results consistent with cognitive neuroscience findings. Extensive experiments demonstrate that RPGCN significantly outperforms state-of-the-art approaches for EEG-based emotion recognition. The interpretable modeling of EEG signals opens new possibilities for integrating brain activity analysis to enable more intelligent and personalized human-computer interaction.

Contents

List of Figures

List of Tables

Abbreviations

AASM	American Academy of Sleep Medicine
AI	Artificial intelligence
AR	Augmented Reality
BCI	Brain-Computer Interfaces
BCDC	Bayesian-copula Discriminant Classifier
BS	Bayesian System
CAM	Class Activation Mapping
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
CT	Computed Tomography
CV	Computer Vision
DAGAM	Domain Adversarial Graph Attention Model
DAPr	Deep Attribution Prior
DDA	Dynamic Domain Adaptation
DeepLIFT	Deep Learning Important Features
DNN	Deep Neural Networks
DSF	Dynamic Spatial Filtering
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
EEG	Electroencephalogram
EENED	End-to-End Neural Epilepsy Detection
ELBO	Evidence Lower Bound
EMD	Empirical Mode Decomposition
EMG	Electromyogram
EOG	Electrooculogram
EPs	Evoked Potentials
ERBM	Entropy Rate Bound Minimization Analysis

ERD	Event-Related Desynchronization
ERPs	Event-Related Potentials
ERS	Event-Related Synchronization
FFN	Feedforward Neural Networks
FIS	Fuzzy Inference System
FN	False Negative
fNIRS	Functional Near-infrared Spectroscopy
FP	False Positive
FPA	Flower Pollination Algorithm
GAP	Global Average Pooling
GC	Granger Causality
GCN	Graph Convolutional Network
GLU	Gated Linear Unit
GlepNet	Global Local Epilepsy Neural Network
GNN	Graph Neural Networks
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GSR	Galvanic Skin Response
HASS	Hybrid Attention Sleep Staging
HCI	Human-Computer Interaction
ICA	Independent Component Analysis
ICS	Interpretability-guided Channel Selection
iEEG	Intracranial Electroencephalogram
IMF	Intrinsic Mode Function
ITR	Information Transfer Rate
JDA	Joint Distribution Adaptation
K	Key
KL	Kullback-Leibler
LIME	Local Interpretable Model-agnostic Explanations
LLMs	Large Language Models
LMD	local Mean Decomposition
LN	Layer Normalization
LRP	Layer-wise Relevance Propagation
LSTM	Long and Short-Term Memory

MDI	Mean Decrease Impurity
MEG	Magnetoencephalography
MHSA	Multi-Head Self-Attention
MI	Motor Imagery
MRI	Magnetic Resonance Imaging
MST	Multi-domain Spatial Transformer
NLP	Natural Language Processing
NREMs	Non-rapid Eye Movements
NSGA	Non-dominated Sorting Genetic Algorithm
PCC	Pearson Correlation Coefficient
PFFM	Position-wise Feed-Forward Module
PPG	Photoplethysmography
PSD	Power Spectral Density
PSG	Polysomnography
Q	Query
RAM	Random Access Memory
REMs	Rapid Eye Movements
RF	Random Forest
RMS	Root Mean Square
RNN	Recurrent Neural Networks
RPCA	Robust Principal Component Analysis
RPG	Relational Probabilistic Graph
RPGCN	Relational Probabilistic Graph Convolutional Network
RSP	Respiratory
RSVP	Rapid Serial Visual Presentation
RTN	Relational Thinking Networks
SDAE	Stacked Denoising Autoencoder
SE	Squeeze-and-excitation
SHAP	Shapley Additive Explanations
SNN	Spiking Neural Networks
SNR	Signal-to-Noise Ratio
SOZ	Seizure Onset Zone
SPA	Spatial Feature Extractors
SSEPS	Steady-state Evoked Potentials
SSVEP	Steady-State Visually Evoked Potential

SVM	Support Vector Machine
TEMP	Temperature
TF	Transformer
TL	Transfer Learning
TP	True Positives
TSK	Takagi-Sugeno-Kang
V	Value
VMD	Variational Mode Decomposition
VR	Virtual Reality
VRNN	Variational Recurrent Neural Network
W	Wakefulness
WPT	Wavelet Packet Transform

The abbreviations do not include the baselines and dataset names.

Chapter 1

Introduction

1.1 Background

Electroencephalogram (EEG) systems have long been established as fundamental tools in neuroscience, neurology, and psychology, employed to measure the brain's electrical activity. By placing electrodes on the scalp, EEG systems capture the brain's spontaneous electrical activity, providing invaluable insights into various neurological and cognitive processes. However, the vast and complex data produced by these systems present significant analytical challenges, necessitating advanced methods for effective interpretation.

In recent years, artificial intelligence (AI) has emerged as a transformative tool for EEG data analysis. AI-driven algorithms have demonstrated remarkable efficiency in processing large EEG datasets, detecting intricate patterns, and even predicting neurological disorders or cognitive states. Despite these advancements, significant challenges remain in the realm of AI-based EEG systems, particularly regarding interpretability and robustness. FIGURE 1.1 provides an overview of the challenges and needs in developing interpretable and robust AI EEG systems.

Interpretable AI: Traditional AI models, particularly deep learning models, are renowned for their accuracy but often operate as opaque "black boxes." This lack of transparency poses a significant concern in medical and neuroscientific contexts, where understanding the rationale behind predictions is crucial. For example, in diagnosing neurological disorders using EEG data, clinicians require not only an

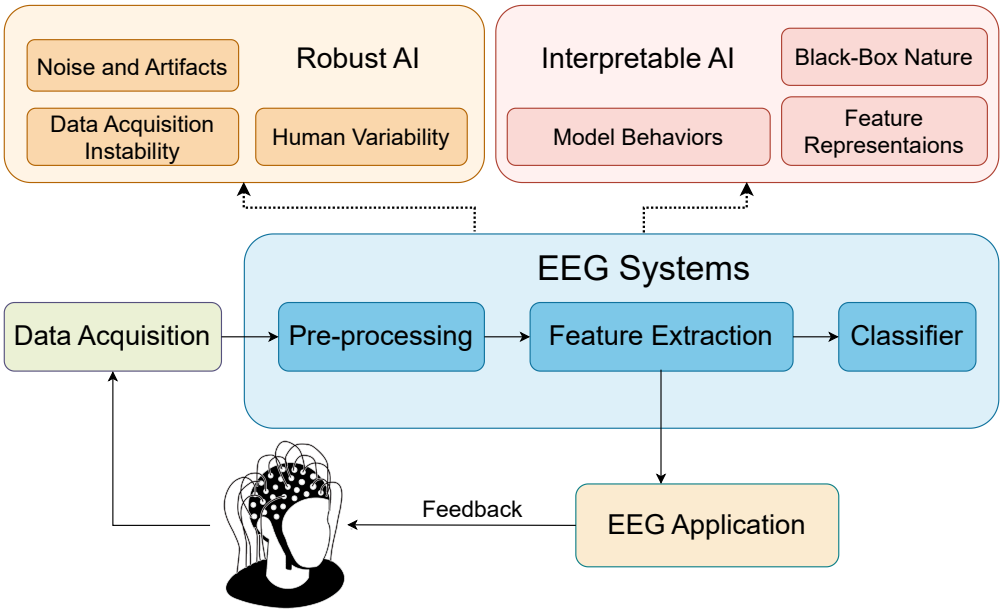

FIGURE 1.1: Overview of the challenges and needs in developing Interpretable and Robust AI EEG Systems. The diagram illustrates the key components of EEG systems, including data acquisition, pre-processing, feature extraction, and classification. It highlights the challenges in creating robust AI models, such as noise and artifacts, data acquisition instability, and human variability. Additionally, it emphasizes the need for interpretable AI, addressing issues related to the black-box nature of models and understanding model behaviors and feature representations. Feedback loops in the system are critical in refining EEG applications for improved accuracy and reliability.

accurate diagnosis but also a clear explanation of the underlying factors driving that diagnosis. Interpretable AI seeks to address this issue by demystifying the decision-making processes of AI models, thereby enhancing trust and usability in clinical settings.

Robust AI: The inherently noisy nature of EEG data further complicates analysis. Factors such as muscle movements, eye blinks, and external electrical interference introduce artifacts into the data, potentially compromising the accuracy of AI models. Robust AI models are designed to withstand such noise and artifacts, ensuring consistent and reliable predictions even when confronted with suboptimal data quality. The necessity for robustness becomes especially critical in real-world applications, where the quality of input data can vary significantly.

The integration of AI into EEG systems also faces additional challenges, including the non-stationary nature of EEG signals, individual differences in brain activity,

and the vast diversity of potential cognitive states or conditions. These factors contribute to the complexity of developing effective AI models for EEG analysis.

The non-stationary nature of EEG signals refers to the fact that brain activity patterns can change over time and vary across different experimental conditions. This variability makes it difficult to develop models that can consistently perform well across different scenarios. Additionally, individual differences in brain activity mean that models trained on one group of subjects may not generalize well to other populations. To address these issues, researchers are exploring techniques such as transfer learning and domain adaptation, which aim to improve model generalization across different contexts.

Furthermore, the vast diversity of potential cognitive states or conditions that can be reflected in EEG data adds another layer of complexity. Accurately classifying and interpreting these states requires models that can capture subtle and complex patterns in the data. This necessitates the development of sophisticated feature extraction and representation learning techniques, which can effectively capture the relevant information from the EEG signals.

This thesis aims to explore the intricate challenges and potential solutions associated with creating interpretable and robust AI models for EEG systems. By addressing these critical aspects, the research seeks to facilitate the development of AI applications that are not only effective but also transparent and reliable. The ultimate goal is to enhance the usability of AI-driven EEG systems in clinical and research settings for innovative and impactful applications in neuroscience, neurology, and psychology.

While AI offers promising prospects for enhancing EEG data analysis, ensuring the interpretability and robustness of these AI models is of great importance. This thesis aims to delve deep into the challenges and solutions associated with creating interpretable and robust AI models for EEG systems, hoping to facilitate the development of more effective, transparent, and reliable novel applications.

1.2 Motivations

The advent of AI-based EEG systems has created revolutionary opportunities across various applications, from medical diagnostics to enhancing human-computer interactions. Despite their potential, these systems often face criticism for their "black-box" nature and instability in practical applications, limiting their reliability and interpretability. These issues hinder the trust and acceptance of such technologies and restrict their practical applicability and effectiveness in critical domains such as healthcare and safety-critical systems.

This thesis is motivated by the urgent need to address these limitations by advancing the development of interpretable and robust AI techniques tailored explicitly for EEG systems. Interpretability is crucial as it provides insights into the decision-making processes of AI models, fostering user trust and enabling practitioners to diagnose and refine the systems effectively. Robustness ensures the reliability and stability of these systems across diverse and challenging real-world conditions, which is essential for applications involving critical decision-making such as driver fatigue detection, sleep staging, and emotion recognition.

The objective of this thesis is not only to contribute significantly to the theoretical and methodological development of interpretable and robust EEG systems but also contribute significantly for the practical application of these enhanced BCI technologies. It seeks to unlock new possibilities for innovation in EEG system-based BCIs, thereby expanding their applicability and effectiveness in improving human lives.

1.3 Objectives

This thesis aims to address critical challenges in the integration of AI with EEG systems, focusing on enhancing system robustness and developing self-interpretable deep learning methods. Specifically, it seeks to understand how interpretable AI techniques can be utilized to bolster EEG systems' resilience against noisy data and variability across sessions/subjects, thereby ensuring reliable analysis. Additionally, it explores the creation of inherently self-explanatory deep learning models that provide insights into their decision-making processes, aiming to make AI in

EEG systems more transparent and trustworthy. The ultimate goal of this thesis is to develop interpretable neural network architectures that not only mitigate the impact of undesirable factors leading to performance decline but also deliver superior and generalized decoding outcomes, thereby advancing the state of EEG analysis and its applications in healthcare and neurotechnology.

1.4 Contributions of The Thesis

FIGURE 1.2 presents an overview of the targets and contributions of this thesis. The objectives, the research areas, the proposed methods, the interpretability & robustness aspects, and the conclusion are included. This thesis has made significant contributions to this field. They include one comprehensive literature review and five compelling series of approaches to tackle the critical challenges and fulfill the needs in this area, as previously mentioned, and are listed as follows.

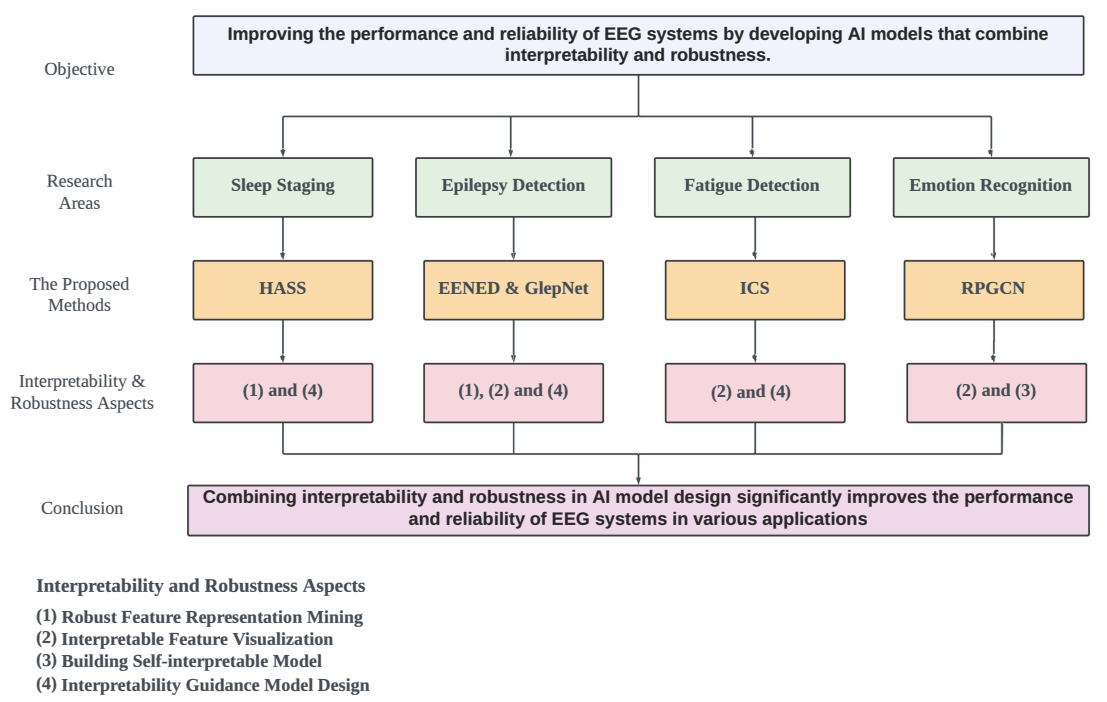

FIGURE 1.2: Overview of the targets and contributions of this thesis.

I. The first contribution of this thesis is to provide a comprehensive literature review focusing on the interpretability and robustness of AI in EEG systems [1]. By highlighting the emerging techniques, it offers insight into the latest trends in this research area. Also, it provides a glimpse into the

enormous potential and implications for the broader field. The thesis serves as an exhaustive guide for understanding, evaluating, and advancing the realm of interpretability and robustness in EEG systems.

II. This thesis introduces the Hybrid Attention EEG Sleep Staging (HASS) Framework for cross-subject EEG sleep staging tasks [2]. HASS employs a spatio-temporal attention mechanism to adaptively assign weights to inter-channel and intra-channel EEG segments based on the spatio-temporal relationships of the brain during different sleep stages. Experimental results demonstrate that HASS can significantly improve typical sleep staging networks and address the difficulties of capturing the spatial-temporal relationships of EEG signals during sleep staging, promising improved accuracy and reliability of sleep assessment.

- **Robust Feature Representation Mining**: Mine robust sleep staging feature via spatial-temporal dimension.
- **Interpretability Guidance Design**: Enhance features with high contributions via the hybrid attention mechanism.

III. This thesis introduces EENED [3] and GlepNet [4], novel EEG-based architectures for neural epilepsy detection. These architectures combine temporal convolutional layers with a multi-head attention mechanism to enhance the accuracy and timeliness of epilepsy diagnosis. The utilization of Grad-CAM for interpretability further elevates their clinical value, allowing healthcare professionals to validate and visually understand the model's diagnostic process, potentially improving patient outcomes and contributing to neuropsychological research.

- **Robust Feature Representation Mining**: Mine robust epilepsy feature via convolution layers and multi-head attention components.
- **Interpretable Feature Visualization**: Apply Grad-CAM for interpretability elevates their clinical value, allowing healthcare professionals to validate and visually understand the model's diagnostic process.
- **Interpretability Guidance Design**: Utilize both global and local information in detecting epilepsy via time-series EEG signals.

IV. This thesis proposes an Interpretability-guided Channel Selection (ICS) framework for the EEG driver drowsiness detection task [5]. ICS provides a two-stage training strategy to select key contributing channels with interpretability guidance progressively. Experiments on a public dataset demonstrate that the proposed method significantly improves the performance of cross-subject driver drowsiness detection.

- **Interpretable Feature Visualization**: Apply the class activation mapping (CAM) to highlight the high-contributing EEG channels.
- **Interpretability Guidance Design**: Design a channel voting scheme to select the top N contributing EEG channels followed by the interpretable feature guidance.

V. By adopting relational thinking theory, this thesis introduces the relational probabilistic graph convolutional network (RPGCN) to improve the decoding performance for EEG emotion classification tasks [6]. RPGCN effectively models variations in potential emotional states by considering relationships among EEG channels and provides interpretability by explaining recognition results consistent with cognitive neuroscience findings. Extensive experiments demonstrate RPGCN's superior performance for EEG-based emotion recognition, opening new possibilities for integrating brain activity analysis into intelligent and personalized human-computer interaction.

- **Interpretable Feature Visualization**: Model and visualize the weight of the GCN's input adjacency matrix as the emotion relation.
- **Building Self-interpretable Model**: Transform raw EEG signals into probabilistic graphs, allowing effective modeling of potential emotional state variations.

The contributions of this thesis are united by a shared focus on advancing interpretability and robustness in EEG-based AI systems, addressing these foundational challenges. Each work contributes to the overarching objectives of this work, demonstrating how diverse approaches can be applied to tackle critical issues in EEG research and applications. Collectively, these contributions form a framework

that enhances the reliability and transparency of EEG models and bridges the gap between theoretical advancements and practical implementations. By integrating innovative methodologies and aligning them with neurophysiological principles, this thesis advances the field of EEG-based AI systems, providing a unified approach to addressing the complex demands of both research and real-world applications.

1.5 Organization of The Thesis

The organization of this thesis is summarized below. In Chapter 2, a comprehensive review of related literature on interpretable and robust AI in EEG systems is provided. In Chapter 3, the thesis analyze Hybrid Attention EEG based Sleep Staging Framework, HASS, to achieve the robust cross-subject sleep staging performance. The novel EEG-based architectures for interpretable neural epilepsy detection, EENED as well as GlepNet, are presented in In Chapter 4. In Chapter 5, the thesis discuss the details of ICS for solving the cross subject issues in driver drowsiness detection under interpretability guidance. Chapter 6 introduces the interpretable RPGCN for EEG Emotion Recognition. Finally, Chapter 7 concludes this thesis, and discuss the further plans and roadmaps for this research field.

1.6 List of Publications

Patents

- **Interpretability-Based EEG Channel Selection Tool for Driver Drowsiness Detection**, Technology Disclosure

Journal Papers

- **Xinliang Zhou**, Chenyu Liu, Ruizhi Yang, Liangwei Zhang, Liming Zhai, Ziyu Jia, and Yang Liu, "Learning Robust Global-Local Representation from EEG for Neural Epilepsy Detection" in **IEEE Transactions on Artificial Intelligence (TAI)**, 2024
- Shaozhe Liu, Leike An, **Xinliang Zhou**, Xiaojun Ning, and Ziyu Jia, "Learning Local to Global Spatial-Temporal Representation for Motor Imagery Classification" in **IEEE Transactions on Systems, Man, and Cybernetics (TSMC)**, 2024 (Under Review)
- **Xinliang Zhou**, Chenyu Liu, Jiaping Xiao, Liming Zhai, Ziyu Jia, and Yang Liu, "Learning Relational Probabilistic Graphs for EEG-based Emotion Recognition" in **IEEE Transactions on Affective Computing (TAFFC)**, 2024 (Under Revision)
- Ziyu Jia, Junyu Ji, **Xinliang Zhou**, and Yuhan Zhou, "Hybrid Spiking Neural Network for Sleep Electroencephalogram Signals" in **Science China Information Sciences (SCIS)**, 2022

Conference Papers

- Chenyu Liu[†], **Xinliang Zhou**[†], Jiaping Xiao, Liming Zhai, Ziyu Jia, and Yang Liu, "VSGT: Variational Spatial and Gaussian Temporal Graph Models for EEG-based Emotion Recognition" in **The 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)**

- Chenyu Liu[†], **Xinliang Zhou**[†], Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu, "VBH-GNN: Variational Bayesian Heterogeneous Graph Neural Networks for Cross-subject Emotion Recognition" in **The 12th International Conference on Learning Representations (ICLR 2024)**

- **Xinliang Zhou**, Dan Lin, Ziyu Jia, Jiaping Xiao, Chenyu Liu, Liming Zhai*, and Yang Liu, " An EEG Channel Selection Framework for Driver Drowsiness Detection via Interpretability Guidance" in **The 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)**

- **Xinliang Zhou**, Chenyu Liu, Jiaping Xiao, and Yang Liu, " EEG-based Sleep Staging with Hybrid Attention" in **The 2023 IEEE Conference on Artificial Intelligence (IEEE CAI 2023)**

- Chenyu Liu, **Xinliang Zhou**, and Yang Liu, " EENED: End-to-End Neural Epilepsy Detection based on Convolutional Transformer" in **The 2023 IEEE Conference on Artificial Intelligence (IEEE CAI 2023)**

[†]: Co-first Authors.

Chapter 2

Literature Review

2.1 Background

EEG provides valuable information about activities and states of the brain in a non-invasive way, being one of the active research areas in human-computer interaction (HCI). With the blossoming of recent AI technologies, EEG systems have increasingly embraced the power of AI for various clinical, entertainment and social interaction applications. For example, sleep staging systems combine EEG signals with deep learning to assist physicians in rapid diagnosis [7]. Driver monitoring systems employ EEG-based deep neural networks (DNNs) to accurately detect driver fatigue to reduce the risk of car accidents [8]. Robotic arm control systems use DNNs to translate human thoughts (reflected by EEG signals) into control signals, helping the disabled perform basic tasks, such as drinking water or moving objects [9]. Although significant progress has been made by AI, the AI models (especially the deep learning-based ones) still remain unexplainable due to their black-box nature and are also susceptible to intentional or unintentional attacks, raising serious concerns for the interpretability [10, 11] and robustness [12] of AI in EEG systems.

Interpretability refers to understanding why and how the AI models make decisions and predictions. Specific to AI-based EEG systems, the interpretability allow researchers to gain insights into EEG dynamics and the link between brain states and cognitive functions, and also make it easier to identify potential biases and failure

This chapter is being prepared for further manuscript submission.

modes of EEG systems. From another point of view, the interpretability can foster user trust and acceptance of EEG systems, enabling users to build confidence in the validity and value of EEG systems.

Robustness refers to the degree to which the decisions and predictions of AI models are free from attacks and perturbations. Unlike traditional HCI data such as image, audio and video, EEG data derived from brain tends to be noisy and variable across individuals, resulting in a lower signal-noise ratio (SNR). This is because EEG signals are easily interfered by biological and environmental artifacts (for example, muscle movements, eye blinks, heartbeat, electrical devices, and so on), and the same stimuli also evoke different EEG responses in different people which has unique neural rhythms.

While the interpretability and robustness in AI based EEG systems have raised serious concerns, and despite tremendous efforts made by researchers to address them, an exhaustive survey summarizing the state of knowledge on these two critical topics remains lacking. There are surveys on interpretable and robust AI in general, but none of them specifically focuses on EEG Systems. To fill this gap, in this thesis we present a systematic literature review covering the following aspects.

We first introduce the background of EEG signals from the EEG categories, EEG applications and EEG datasets. We then elaborate the interpretability and robustness of AI in EEG systems, respectively. For the interpretability, we classify the interpretable AI into three types from the implementation perspective of interpretability methods. The first type is post-hoc methods based on back-propagation, which obtains the contribution of the features to results by back-propagating the prediction results. The second type is perturbation-based methods that explain initial models using local models trained with data perturbation. The last type is rule-based methods, and it applies models based on logical rules to make predictions. For the robustness, we categorize the robust AI into four classes: signal-component-related, subject-related, device-related, and the latest adversarial-attack-related challenges, covering all threats to the stability and security of AI-based EEG systems in practical applications. Within each category, we summarize the common features and shared methodologies, describe representative works, and analyze their differences. Finally, we discuss the potential directions for future research and propose practical suggestions.

The contributions of this literature review are as follows:

- This is a comprehensive literature review focusing on the interpretability and robustness of AI in EEG systems.
- We propose a novel taxonomy of interpretability and robustness for EEG systems.
- We summarize and highlight the emerging and most representative interpretable and robust AI works related to EEG systems.

2.2 Understanding EEG Signals: Categories, Applications and Datasets

In this section, we provide an overview of the EEG paradigms, including EEG signal categories, EEG signal applications and typical EEG datasets. The EEG paradigm is a widely used approach in the field of BCI, and it involves measuring the brain's electrical activity through electrodes placed on the scalp. Compared with other paradigms in BCI, such as the ECoG paradigm [13], the advantages of the EEG paradigm include its non-invasive nature, high temporal resolution, and relative ease of use. However, it has poor spatial resolution and is susceptible to interference from external sources such as muscle activity. Despite these limitations, the EEG paradigm continues to be a valuable tool in developing BCI technology.

2.2.1 EEG Signal Categories

EEG signals fall into three general categories: spontaneous EEG, evoked potentials, and event-related desynchronization/synchronization (ERD/ERS). The spontaneous EEG do not involve external stimuli presented to subjects, while the evoked potentials elicit the subjects' EEG responses to specific external stimuli. In contrast, the ERD/ERS is stimuli-irrelevant, only reflecting subjects' mental activities. The EEG signal categories are shown in FIGURE 2.1.

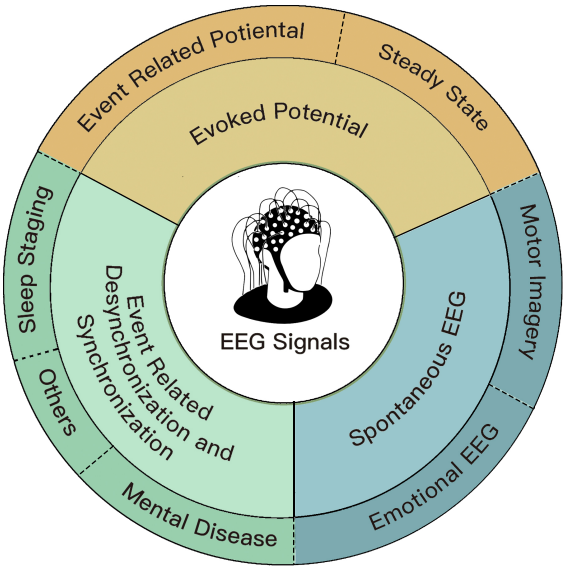

FIGURE 2.1: Summary of EEG Signal Categories.

2.2.1.1 Spontaneous EEG

The most widely used EEG, in general, is spontaneous EEG. It refers to the measurement of brain waves obtained without external stimuli. Some common spontaneous EEG signals are obtained from the scenarios where test subjects are engaged in experiencing fatigue and sleeping, suffering from a brain disorder (for example, Autism, Seizure), and performing motor imagery (MI) tasks [14].

2.2.1.2 Evoked Potentials

Evoked Potentials (EPs), also called evoked responses, are EEG signals elicited by non-spontaneous event stimuli. Depending on different kinds of stimuli, there are two forms of EPs signals: event-related potentials (ERPs) and steady-state evoked potentials (SSEPs). The ERPs record the EEG signals elicited by specific and isolated stimulus events. While the SSEPs can reflect subjects' perception of pressure, touch, temperature and pain. On this basis, both ERPs and SSEPs contain somatosensory, auditory, and visual potentials according to the subjects' senses. All EPs signals, such as $P300$ [15], rapid serial visual presentation (RSVP) [16], and error-evoked potentials, are more robust than spontaneous signals because the amplitude and frequency of EPs are typically higher.

2.2.1.3 Event-Related Desynchronization and Synchronization

The ERD/ERS reflects a relative power decrease/increase of EEG in a specific frequency band during physical motor executions and mental activities. The ERD/ERS does not require any external stimuli. However, to gather high-quality ERD/ERS signals, participants must undergo lengthy training that may last several weeks. In addition, the ERD/ERS signals are prone to fluctuate with different participants and thus have low stability.

2.2.2 EEG Signal Applications

EEG signals have various applications. We list the six most typical applications shown in FIGURE 2.2.

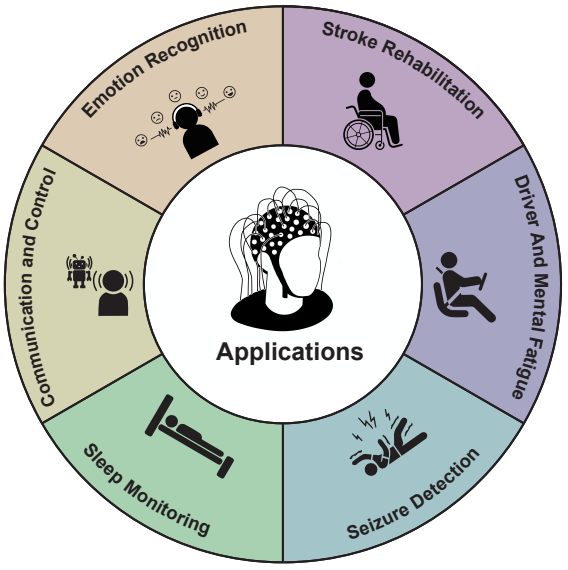

FIGURE 2.2: Typical EEG Applications.

2.2.2.1 Sleep Monitoring

Sleep monitoring plays a crucial role in the early diagnosis and intervention of sleep-related diseases. According to the sleep staging criteria proposed by the American Academy of Sleep Medicine (AASM) [17], the complete sleep process can be divided into three stages: wakefulness (W), non-rapid eye movements (NREMs), and rapid eye movements (REMs). Different sleep stages are reflected in different prominent

waveforms in EEG signals. For example, spindles and K-complex waveforms are prominent features in the NREMs stage [18]. To save manpower and time, employing EEG signals for automatic sleep monitoring has gradually become a hot research topic [19–21].

2.2.2.2 Seizure Detection

The characteristics of seizure activities can be observed from EEG signals, which provide essential information about the type and severity of the seizure and help identify the location of the seizure focus in the brain (for example, temporal lobe) [22, 23]. The EEG signals are crucial for developing effective treatment plans and monitoring the patient's health condition.

2.2.2.3 Fatigue Detection

EEG signals can be used for detecting and monitoring fatigue, such as driver fatigue [24, 25] and mental load fatigue. One common way is through spectral power analysis, for which the changes in the power of different frequency bands can be used as fatigue indicators. Another way is to analyze the ERPs. When fatigue happens, the amplitude of certain ERPs (For example, $P300$) will decrease [26]. EEG signals can also be combined with other techniques, such as eye tracking and reaction time tests, providing a more complete picture of the cognitive and neural changes associated with fatigue.

2.2.2.4 Communication and Control

EEG signals representing human intent can be decoded into language or control signals used to communicate with people or intelligent devices. A typical application is the $P300$ speller, which enables users to type without any motor systems and converts their intention into text [15, 27, 28]. Besides, some intelligent environment stuff can be linked to and controlled by EEG systems, for example, assistive robots in smart homes. These applications can be achieved by detecting SSVEP owing to the advantages of less training time, excellent recognition performance, and high information translation rate (ITR) [29, 30]. Current research in

this scenario focuses mainly on controlling robots [31–33], wheelchairs [32, 34], and so on

2.2.2.5 Emotional Recognition

EEG systems can be applied to assess and understand changes in the brain related to mental and physical states [35–37]. Studies have shown differences in the activity of specific brain regions, such as the amygdala and prefrontal cortex, when a person is experiencing fear or happiness. Activities in these brain regions are directly reflected in the related EEG signals, which could be applied to mental health [38], cognitive psychology [39], and affective computing research [11, 40].

2.2.2.6 Stroke Rehabilitation

Stroke has a high mortality rate and leads to long-term disability in up to 50% of survivors. Therefore, motor rehabilitation is a top priority for post-stroke treatment [41]. Unlike traditional stroke rehabilitation treatments, EEG systems do not rely on patients' residual motor ability. In contrast, EEG systems create a direct communication pathway between the brain and an external device [42], bypassing the traditional neuromuscular pathway. During EEG system-assisted rehabilitation, the system collects the patient's EEG signals and then decodes the patients' motor intentions into commands through signal processing. These commands drive the robotic device to move the patient's paralyzed limb to complete the rehabilitation exercise. Studies have reported motor cortex activation in patients who underwent EEG Systems-based rehabilitation and statistically significant improvements in patients' motor abilities during subsequent motor assessments [43, 44].

2.2.3 Typical EEG Datasets

The selection and utilization of datasets is a critical foundation of physiological signal processing, greatly affecting the interpretability and robustness of derived conclusions. Table 2.1 summarizes the public EEG datasets employed for various tasks, including emotion recognition, fatigue detection, seizure detection, sleep monitoring and motor imagery.

For example, the emotion recognition task engages datasets such as Deap [45], Dreamer [46], and ASCERTAIN [47], which incorporate various physiological signals, encompassing EEG, Electromyography (EMG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), Temperature (TEMP), and Respiratory (RSP). These datasets cater to various age demographics and compile a substantial number of subjects, thereby augmenting the robustness of the analytic methodologies applied. Likewise, fatigue detection and seizure detection tasks leverage diverse datasets, underscoring the resilience and adaptability of the ensuing models.

Differently, sleep monitoring and motor imagery domains utilize multiple datasets collated from a wide demographic range, reinforcing the inferred models' broad applicability. The comprehensive nature of these datasets buttresses the interpretability of the findings by providing a detailed understanding of the behavior of different physiological signals under an array of conditions.

The broad spectrum of datasets outlined in Table 2.1 underscores the extensive application of these tasks across various ages and subject populations, highlighting the inherent robustness of these approaches. Besides, these datasets contribute to the interpretability by fostering a nuanced and comprehensive understanding of the field, thus improving the transparency in physiological signal processing.

2.3 Interpretable AI in EEG Systems

AI interpretability refers to explaining the decisions and actions of AI models in a manner that humans can understand. For interpretable AI in EEG systems, it means the internal logic and workings of AI models conform to physiological principles. For example, in motor imagery (MI) tasks, the EEG signals that contribute to predictions are derived from electrodes around the motor cortex, as depicted in FIGURE 2.3.

Interpretability is essential for EEG systems because it assesses whether the AI model has learned physiologically meaningful features. Foremost, interpretability allows checking whether the predictive logic of AI models conforms to specific proven physiological rules, since the predictive accuracy scores of the AI models can be deceptive. For example, in MI tasks, the model making decisions may pay more attention to the noises generated by subjects' involuntary muscle movements rather

TABLE 2.1: Summary of public datasets used in EEG systems.

Task	Dataset	Physiological Signal	Subject Number	Subject Age
ER	Deap [45]	EEG, EMG ...	32	19 – 37
	SEED [48]	EEG	15	Mean 23.27
	SEED-IV [49]	EEG	15	20 – 24
	SEED-V [50]	EEG, SMI	20	N/A
	Dreamer [46]	EEG, ECG	23	20 – 33
	HCI-Tag [51]	EEG, ECG...	30	19 – 40
	ASCER-TAIN [47]	EEG, ECG ...	58	Mean 30
	AMIGOS [52]	EEG, ECG ...	40/37	N/A
	Enterface 06 [53]	EEG, fNIRS	16	N/A
	Imagined Emotion [54]	EEG	31	18 – 38
Fatigue Detection	MEDT [55]	EEG	27	22 – 28
	DDDE [56]	EEG	13	44.5 ∓ 18.8
	FatigueSet [57]	EEG, ECG ...	12	N/A
Seizure Detection	CHB-MIT [58]	EEG	23	1.5 – 22
	Bonn University [59]	EEG	10	N/A
	Freiburg Seizure [60]	EEG	21	Adults and Children
	Helsinki University [61]	EEG	79	Infants
	EPILEP-SIAE [60]	EEG	275	N/A
	Temple University [62]	EEG	80	6 – 56
	NMT [63]	EEG	N/A	1 – 90
Sleep Monitoring	Sleep-EDF [64]	EEG, EMG ...	197	18 – 90
	SHHS [65]	EEG, EMG ...	6441	Above 40
	MASS [66]	EEG, EMG ...	200	18 – 76
	ISRUC [67]	EEG, EMG ...	118	Adults and Children
	HMC [68]	EEG, EMG ...	105	20 – 80
	CAP Sleep [69]	EEG, ECG ...	102	18 – 70
	Sleep Cassette [70]	EEG, EOG ...	100	18 – 65
	SIESTA [71]	EEG, EOG ...	72	18 – 80
Motor Imagery	BCI IV 2a* [72]	EEG, EOG	9	N/A
	BCI IV 2b* [72]	EEG, EOG	9	N/A
	OpenMBI [73]	EEG	54	24 – 45
	Stroke [74, 75]	EEG	21	Mean 54.2
	MOABB [76]	EEG	104	18 – 55

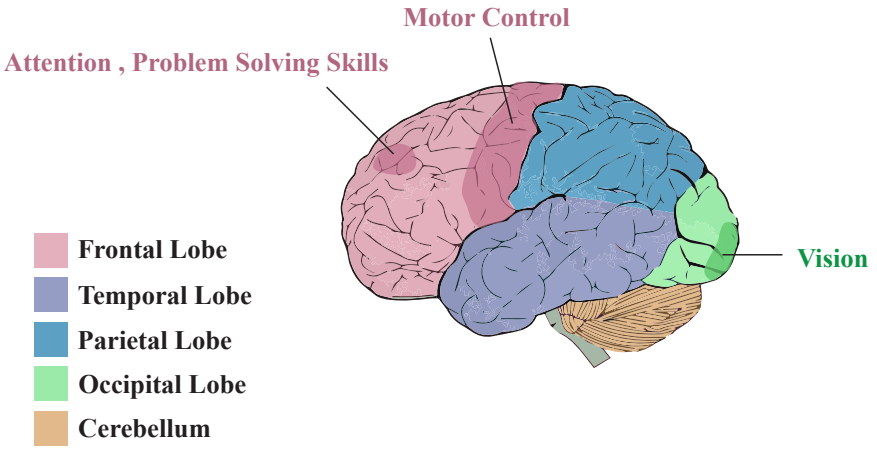

FIGURE 2.3: Key brain regions related to the motor imagery task. These include the frontal lobe for cognitive skills and motor function, the temporal lobe for sensory input processing, the parietal lobe for sensory information integration, the occipital lobe for vision, and the cerebellum for coordination of voluntary movements. Each region contributes to the successful execution of the task.

TABLE 2.2: Summary of Interpretable AI in EEG Systems.

Interpetability Categories	Methods	Coverage	Explanation Type	Representative Works
Backpropagation-based Methods	LRP	Local/Global	Attribution	[78–82]
	DeepLIFT	Local/Global	Attribution	[83–87]
	CAM	Local	Attribution	[24, 88–91]
	Grad-CAM	Local	Attribution	[4, 92–96]
Perturbation-based Methods	LIME	Local	Attribution	[97–101]
	SHAP	Local	Attribution	[85, 102–105]
Rule-based Methods	RF	Global	Decision Rules	[106–110]
	FIS	Global	Fuzzy Rules	[111–115]
	BS	Global	Bayesian Rules	[116–120]

than EEG signals that truly originate from cranial nerve movements. Furthermore, interpretability methods can uncover patterns that inform brain signal research. For example, when predicting subjects' sleep states, models identified that the signals of peripheral EEG channels generated by regular eye movements during deep sleep are highly correlated with sleep status [77], even though these EEG signals had long been overlooked.

After a thorough review of existing literatures, we divide the interpretability methods applied in EEG systems into three categories from the perspective of the implementation: backpropagation-based methods, perturbation-based methods and rule-based methods. We summarize the interpretability categories and their representative works in Table 2.2.

Type and Coverage of interpratable AI in EEG Systems: In EEG systems,

the interpretability of AI models can be broadly categorized as either local or global, influenced by feature attribution or logic rules. Local interpretability aims to explain individual predictions by illuminating why a model correlates a specific EEG pattern with a particular condition. Techniques like Layer-wise Relevance Propagation (LRP) [121], Deep Learning Important Feature (DeepLIFT) [122], Class Activation Mapping (CAM) [123], Gradient-weighted Class Activation Mapping (Grad-CAM) [124], Local Interpretable Model-Agnostic Explanations (LIME) [125], and Shapley Additive Explanations (SHAP) [126] provide local interpretability.

In contrast, global interpretability illuminates the overall behavior of a model, revealing how it operates across multiple instances. Methods like random forest (RF) [127], fuzzy inference system (FIS) [128] and Bayesian system (BS) [129] are frequently utilized for the global interpretability.

Feature attribution, assigning importance values to input features for a model's decision, is prevalent in methods such as LRP, DeepLIFT, CAM, Grad-CAM, LIME and SHAP, potentially highlighting key brain activity patterns. Meanwhile, logic rules that provide clear criteria for classifying EEG data conditions are also used in RF, FIS and BS.

The explanation type (attribution or logic rule) and interpretability scope (local or global) provide unique and crucial insights into AI decision-making in EEG systems. The selection depends on the required level of interpretability and the most appropriate explanation type for the given data and task.

2.3.1 Backpropagation-based Methods

Backpropagation-based methods decompose the model predictions by first back-propagating the gradients from the predictions into input feature space and then visualizing the weights of these features (for example, time-frequency patterns and electrode regions) in raw EEG signals that contribute to predictions.

2.3.1.1 Layer-wise Relevance Propagation

LRP [121] provides insight into the neurophysiological phenomena behind EEG models' predictions by backpropagating results. The LRP aims to determine the contribution (measured by the relevance value) of individual elements within the input signal (corresponding to each sample point of the EEG signal) to the output prediction. It allows EEG models to integrate temporal information and brain-topography-related spatial information by producing heatmaps.

To implement the LRP, a neural network like a convolutional neural network (CNN) is first be trained to process EEG signals. Let $R_i^{(l)}$ denotes the relevance value of neuron i in layer l ($l = 1, 2, \ldots, L$). At the output layer (for example, $l = L$), the relevance value is equivalent to the model's predicted score:

$$R_i^{(L)} = f_i, \tag{2.1}$$

where f_i is the activation value of neuron i in the output layer of the neural network. For each layer l, the backpropagation is used to propagate $R_i^{(l)}$ to the subsequent layer, $R_j^{(l-1)}$. In the LRP, this process is defined as

$$R_j^{(l-1)} = \sum_i \frac{f_j^{(l-1)} w_{ji}^{(l)}}{z_i^{(l)}} R_i^{(l)}, \tag{2.2}$$

where $f_j^{(l-1)}$ represents the activation value of neuron j in layer $(l-1)$, $w_{ji}^{(l)}$ denotes the weight between neuron i in layer l and neuron j in layer $(l-1)$, and $z_i^{(l)} = \sum_j f_j^{(l-1)} w_{ji}^{(l)}$ denotes the sum of the inputs to neuron i in layer l.

Backpropagating through each layer allows us to calculate the relevance values for every sample point in the input layer (for example, the original EEG signal). A higher relevance value indicates that the data contributed more to the prediction, while a lower value indicates less contribution. Visualizing these relevance values reveals how the neural network extracts useful information from EEG signals. This offers insights to improve neural network models or understand brain signals more thoroughly.

The LRP reveals whether the models focus on task-relevant EEG signals. Ellis *et al.* [79] used the LRP to generate heatmaps highlighting the local and global signals. The heatmaps show that the local signals with higher relevance values

are highly related to the sleep state of the human brain, matching neurophysiological expectations. Similarly, the LRP can clarify the contributions of noise and neurophysiological factors. Nagarajan *et al.* [80] applied LRP to select the high contributing EEG channels in MI tasks, confirming that the model indeed learns features from the electrodes at action-related brain regions.

The LRP also reveals how EEG signals from different brain dimensions correlate with model decisions over time. Sturm *et al.* [81] used the LRP to track how a DNN's attention shifted between feature regions during action switching in MI tasks, revealing the physiological principles underlying the model predictions. Moreover, Wang *et al.* [130] and Bang *et al.* [78] leveraged the LRP to explain 3D-CNN model predictions, for which the heatmaps simultaneously highlighted the contributions of frequency ranges, time intervals and spatial locations of relevant signals.

2.3.1.2 Deep Learning Important Features

DeepLIFT aims to check if the model's decisions align with known neurophysiological phenomena, and provide guidance for finding generalizable EEG features. Similar to LRP, the DeepLIFT uses backpropagation to calculate how each input feature correlates with the model prediction for each trial.

The DeepLIFT is based on reference activation, enabling comparison of a feature's importance against a predefined reference point. Its core principle is to calculate a contribution score for each input feature. The contribution score can be computed using the following equation:

$$C_i = (f_i - f_i^0) \times \frac{\partial y}{\partial f_i}, \tag{2.3}$$

where C_i denotes the contribution score for feature i, f_i represents the actual activation of feature i, f_i^0 signifies the reference activation for feature i, $\frac{\partial y}{\partial f_i}$ denotes the gradient of the output with respect to the activation of feature i.

For the DNNs composed of multiple layers, the chain rule is employed to compute the contribution score for each input feature. The chain rule for DeepLIFT can be

expressed as

$$C_i = \sum_j C_{i,j} = \sum_j \frac{\partial f_j}{\partial f_i} C_j, \tag{2.4}$$

where $C_{i,j}$ indicates the contribution score of feature i to feature j, $\frac{\partial f_j}{\partial f_i}$ denotes the gradient of the activation of feature j with respect to the activation of feature i, and C_j represents the contribution score for feature j.

The most common application of the DeepLIFT method in EEG systems is to verify whether the prediction logic of models conforms to physiological principles. For example, Lawhern *et al.* [83] used the DeepLIFT to prove that their proposed EEGNet can learn to focus on EEG channels near task-related brain regions in different EEG classification tasks. Similarly, Ju *et al.* [131] used the DeepLIFT to interpret the spatiotemporal frequency information learned by their Tensor-CSPNet in MI tasks, and found it match the key frequency components existing in the left and right hands.

On the other hand, the DeepLIFT can discover certain feature patterns from model predictions to guide brain research. For seizure detection, the high gamma frequency is known to be a key feature for distinguishing pre-ictal and inter-ictal segments. However, Gabeff *et al.* [84] interpreted the models using DeepLIFT, revealing that some low amplitude patterns were also detected as ictal. This finding complements the established conclusion, and verifies that the resting EEG features can also be helpful, providing counter-balancing information for seizure detection.

2.3.1.3 Class Activation Mapping

CAM is a technique that produces heatmaps by visualizing the importance of each input feature in the final classification decision. For EEG systems, the input features are often the EEG signals from multiple channels and time points. To describe the CAM method mathematically, let $f_k(x)$ denote the activation of the k-th feature map in the last convolutional layer of the network, given an input x. The class-specific weights w_c^k for class c are learned during the training process. The class activation map for class c can be computed as

$$M_c(x) = \sum_k w_c^k \cdot f_k(x), \tag{2.5}$$

where $M_c(x)$ represents the heatmap for class c. This heatmap can be visualized as an overlay on the input EEG signals, highlighting the most relevant spatial and temporal features that contribute to the final classification decision.

To obtain the final classification score, the global average pooling (GAP) layer is applied on the feature maps, and then a softmax activation function is used to generate the probability distribution over classes:

$$p_c(x) = \frac{e^{S_c(x)}}{\sum_{c=1}^{C} e^{S_c(x)}}, \tag{2.6}$$

where $S_c(x) = \sum_i \sum_j f_k^c(x)_{i,j}$ is the sum of the activation values of the k-th feature map for class c, and C denotes the total number of classes.

CAM analysis connects the deep-layered features to the biologically meaningful features. In the work of Cui *et al.* [24], the features from the last convolutional neural network layer were traced back to the bursts in the θ band and the spindles in the α band, which strongly relate to drowsiness. Similarly, Yildiz *et al.* [88] analyzed the seizure detection model using CAM, and found that low-frequency EEG signals are critical for distinguishing seizures.

2.3.1.4 Gradient-weighted Class Activation Mapping

Grad-CAM extends the CAM approach by considering the gradient information flowing into the last convolutional layer of the network, offering a more precise and high-resolution visualization of relevant features in AI-based EEG systems. The Grad-CAM method computes the importance weight α_c^k for the k-th feature map with class c in the last convolutional layer as follows:

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial S_c(x)}{\partial f_k(x)_{i,j}}, \tag{2.7}$$

where Z is the total number of spatial locations in the feature map, and $\frac{\partial S_c(x)}{\partial f_k(x)i,j}$ denotes the gradient of the class score $S_c(x)$ with respect to the activation $f_k(x)_{i,j}$ at the spatial location (i, j). The Grad-CAM heatmap for class c can then be computed as

$$M_c^{\text{Grad}}(x) = \text{ReLU}\left(\sum_k \alpha_c^k \cdot f_k(x)\right), \tag{2.8}$$

where ReLU is the rectified linear unit function, ensuring that only positive contributions are considered. Compared to the CAM, the Grad-CAM gives a more nuanced understanding of how the model behaves, since it takes into account both the positive and negative influences of the input features.

The Grad-CAM have been utilized to provide insight into the features learned by classifiers. Fei *et al.* [90] revealed that higher frequency bands are particularly useful for emotion recognition. Jonas *et al.* [132] identified key EEG features for prognostication in comatose patients after cardiac arrest. Similarly, Aslan *et al.* [133] used the Grad-CAM to visualize model outputs and clarify the relationship between frequency components in seizure patients versus healthy individuals. Additionally, applying Grad-CAM for channel selection can enhance decoding efficacy and achieve an optimal balance between model performance and channel utilization [93].

2.3.2 Perturbation-based Methods

Perturbation-based methods perturb individual EEG samples and observe the impact on subsequent network neurons and predictions, trying to reveal correlations between samples and model outputs. Similar to backpropagation-based methods, they are also post-hoc methods that interpret the models by attribution. However, the perturbation-based methods are model-agnostic, building local models to approximate the predictions of the original models based on perturbed inputs. In other words, the local models establish the connection between biological features and original model predictions.

2.3.2.1 Interpretable Model-agnostic Explanations

LIME explains target model predictions by approximating them locally with interpretable models. To be specific, the LIME approximates the complex model's behavior near a specific input point x using a simple and locally linear model, which quantifies the contribution of individual elements within the input signal to

the prediction $f(x)$. The original input x is perturbed to create similar inputs x_i with $i = 1, 2, \ldots, N$, and their corresponding predictions $f(x_i)$ are obtained from the trained model. Weights w_i for the perturbed inputs inputs x_i are computed using an exponential kernel

$$w_i = \exp\left(-\frac{d_i^2}{\sigma^2}\right), \tag{2.9}$$

where d_i denotes the distances between the x and x_i, and σ is a scaling factor.

The simple linear model (for example, linear regression) is then trained using the perturbed inputs x_i, predictions $f(x)$, and weights w_i. The coefficients β_i of this simple linear model represent the contributions of each sample point in the input signal to the output prediction:

$$f(x) \approx \sum_{i=1}^{N} \beta_i x_i, \tag{2.10}$$

Visualizing these feature contributions provides insights into how the model processes EEG signals and extracts relevant information. This can guide researchers to improve neural network models or interpret brain signals more comprehensively.

Locally interpretable models offer a direct way to map initial model predictions onto EEG features. Giudice *et al.* [97] used local models to explain DNN predictions of voluntary/involuntary blinks, revealing that the peaks and troughs in signals correspond to voluntary and involuntary eye blink behaviors, respectively. Similarly, Alsuradi *et al.* [98] also utilized the LIME to explain active/inactive action predictions, and find that the action trial can be identified as active for strong desynchronization in the α and β bands, and passive for the synchronization in those bands.

Some models, containing specialized layers like SAGpooling, impede the visualization of feaure contributions through backpropagation. To tackle this issue, Xu *et al.* [99] applied the LIME to construct local interpretability models for the domain adversarial graph attention model (DAGAM). They identify that the symmetry of EEG activities between the left and right hemispheres is a critical feature of neutral emotions.

2.3.2.2 Shapley Additive Explanation Values

SHAP quantifies the contribution of each input features to prediction based on the Shapley values from game theory. The Shapley value refers to the marginal contribution of the EEG feature, which is the difference between prediction results before and after the feature is added. The SHAP value for feature i in a model f is defined as

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)], \tag{2.11}$$

where N is the set of all input features, S is a subset of features without feature i, and $|S|$ denotes the cardinality of S. The term $f(S \cup \{i\}) - f(S)$ represents the marginal contribution of feature i when added to the subset of features S.

SHAP values have three key properties: local accuracy, missingness, and consistency. Local accuracy ensures that the sum of SHAP values for each input feature and the expected model output equals the model prediction for a specific instance. Missingness idictates that if a feature is missing or has no impact on the model prediction, its SHAP value will be zero. Consistency guarantees that if a feature contributes more in a new model compared to an old one, the SHAP value of that feature should not decrease.

The SHAP is often utilized to explain complex AI models in EEG systems. Tahmassebi *et al.* [85] constructed a real-time DNN model to monitor patients' eye states. To ensure model interpretability in practical scenarios, they employ the SHAP to build locally interpretable models that reveal the relationship between EEG features and the eye state. Raab *et al.* [102] also used the SHAP to explain the feature contributions of initial models, which are DNN models of different dimensions (1-D and 3-D) for seizure detection.

2.3.3 Rule-based Methods

Unlike post-hoc interpretability methods which use feature contributions as the explanation, the rule-based methods apply particular logic rules, such as decision rule, fuzzy rule, Bayesian rule, as the interpretations of the EEG systems, resulting in high interpretability.

2.3.3.1 Random Forest

RF is an rule-based method based on the "IF-THEN-ELSE" logic rule (for example, decision rule). The interpretability of RF arises from two main aspects: feature importance and decision paths.

Feature Importance: Feature importance in RF is typically calculated using the Gini importance or mean decrease impurity (MDI). For each feature i derived from the EEG signals, the Gini importance $I_G(i)$ is given by

$$I_G(i) = \sum_{t \in T_i} \frac{n_t}{n} \left(1 - \sum_{c=1}^{C} p_{tc}^2 \right), \tag{2.12}$$

where T_i represents the set of nodes in the RF that split on feature i, n_t is the number of samples reaching node t, n is the total number of samples, C is the number of classes, and p_{tc} is the proportion of samples with class c in node t. The Gini importance measures the decrease in node impurity, which is weighted by the probability of reaching each node. The features of the EEG signals can be ranked based on their Gini importance, so that the most relevant features (also the most relevant EEG signals) contributing to the model's decision-making process can be identified.

Decision Paths: RF consists of multiple decision trees, each of which is trained on a bootstrapped sample of the original data. The decision path of an instance in a tree is the sequence of nodes from the root to a leaf node, which corresponds to the class assigned by the tree.

RF model interpretability can be achieved by analyzing decision paths for an instance across all trees. This reveals common patterns and rules leading to predictions, providing insights into the decision-making process.

Visualizing feature importance and decision paths provides a deeper understanding of how input EEG signals relate to model predictions, and can aid model refinement and more comprehensive brain signal interpretation. For example, Abdulhay *et al.* [134] employed the RF to classify the Shannon entropy of instantaneous values associated with each intrinsic mode function (IMF), directly responding to the association between the instantaneous amplitudes and frequencies of the IMF and seizures. Li *et al.* [135] followed a similar approach for seizure classification. The

difference is that the authors classify the electrodes directly by RF to find the seizure onset zone (SOZ).

2.3.3.2 Fuzzy Inference System

FIS is a computational framework based on the fuzzy set theory, fuzzy logic or fuzzy reasoning. It offers inherent interpretability by using human-readable rules and transparent reasoning processes.

A typical FIS consists of four main components: fuzzification, fuzzy rule base, fuzzy inference engine, and defuzzification. The *fuzzification* involves converting crisp input values into fuzzy sets using membership functions. For each input variable x_i, a membership function $\mu_{A_i}(x_i)$ is used to determine the degree of membership of x_i to the fuzzy set A_i:

$$\mu_{A_i}(x_i) : x_i \to [0, 1], \tag{2.13}$$

where $\mu_{A_i}(x_i)$ represents the degree of membership of x_i to the fuzzy set A_i.

The *fuzzy rule base* is a collection of human-readable "IF-THEN" rules that describe the relationships between input and output fuzzy sets. A fuzzy rule can be expressed as:

$$R_k : \text{ IF } x_1 \text{ is } A_{k1} \text{ AND } \cdots \text{ AND } x_n \text{ is } A_{kn} \text{ THEN } y \text{ is } B_k, \tag{2.14}$$

where x_i are the input variables, A_{ki} and B_k are fuzzy sets, and R_k represents the k-th fuzzy rule.

The *fuzzy inference engine* combines the fuzzified input values and fuzzy rules to produce fuzzy output sets. The firing strength or weight w_k of each rule R_k is computed as the product of the membership degrees of the input values to their corresponding fuzzy sets:

$$w_k = \prod_{i=1}^{n} \mu_{A_{ki}}(x_i), \tag{2.15}$$

The fuzzy output sets are then generated by aggregating the weighted consequent fuzzy sets B_k:

$$B'_k(x) = w_k \cdot B_k(x), \tag{2.16}$$

where $B'_k(x)$ is the weighted fuzzy output set for rule R_k.

The *defuzzification* process converts the fuzzy output sets back into crisp output values. One common method is the centroid defuzzification, which obtains the crisp output value y by calculating the centroid of aggregated fuzzy output sets

$$y = \frac{\sum_x B'_k(x) \cdot x}{\sum_x B'_k(x)}. \tag{2.17}$$

FIS provides human-readable rules that give insights for model refinement and brain signal interpretation. Feng *et al.* [111] developed a Takagi-Sugeno-Kang (TSK) FIS based on joint distribution adaptation (JDA) to simultaneously reduce the difference between the marginal distribution and the conditional distribution of the EEG training sets and test sets. This approach can be extended to multi-categorical EEG seizure detection tasks. Furthermore, Jiang *et al.* [112] applied the TSK-FIS to driver fatigue detection, and proposed an online multi-view & transfer TSK-FIS for driver drowsiness estimation. This FIS is inherently interpretable, enabling direct tracing of the EEG channels associated with fatigue.

2.3.3.3 Bayesian System

The BS uses Bayesian theorem to model the relationship between EEG features and predictions:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}, \tag{2.18}$$

where θ denotes the model parameters, and D represents the observed data (for example, EEG signals and associated cognitive states). $P(\theta)$ is the prior distribution capturing our prior beliefs about the parameters that relate EEG signals to cognitive states. $P(D|\theta)$ is the likelihood function that quantifies the probability of observing the EEG data given the model parameters. Depending on the problem, this might involve a linear regression model, a neural network, or any other appropriate models that capture the. $P(\theta|D)$ is the posterior distribution representing the updated beliefs about the model parameters after observing the EEG data.

The interpretability of Bayesian Systems in EEG applications stems from the explicit representation of uncertainty through probability distributions. By analyzing the posterior distribution $P(\theta|D)$, it becomes possible to explain the relationships between EEG signals and cognitive states, as well as the uncertainty associated with the model's predictions.

TABLE 2.3: Comparison of different interpretability methods in EEG Systems.

	Backpropagation-based Method	**Perturbation-based Method**	**Rule-based Method**
Mechanism	Analyze the feature contribution by backpropagating the gradients from predictions.	Explain the original model's behavior with local surrogate models.	Explain model using specific logic rules
Explanation Stage	Post-hoc	Post-hoc	Ante-hoc
Model Dependence	Model-specific	Model-agnostic	Model-agnostic
Flexibility	Low	High	High
Application Scenario	Differentiable models	Tolerable of high computational costs	Availability of priori knowledge
Limitation	Gradient dependency; Narrow applicability	Computationally intensive; Prone to overfitting	Oversimplify complex EEG systems; Require domain expertise

Qian *et al.* [116] introduced a Bayesian-copula discriminant classifier (BCDC) to study the relationship between drowsiness and nap. The BCDC shows a better understanding of the periodical rhymes of physiological states, and enhances the interpretability of driver alertness. Wu *et al.* [117] proposed a separation and recovery Bayesian method, finding that the predictive emotion features originate from the lateral temporal area and distribute in γ and β bands.

2.3.4 Discussion on Interpretable AI in EEG Systems

Selecting an appropriate interpretability method is crucial for understanding the decision-making mechanisms of AI models in EEG systems. Each has its own strengths and limitations, which we will explore in more detail. The comparative overview of these methods is shown in Table 2.3.

Backpropagation-based methods represent a popular subset of interpretability techniques, encompassing methods like LRP, DeepLIFT, CAM, and Grad-CAM. These are post-hoc methods, which mean that they are applied after the model has made its predictions to identify the time-frequency patterns and electrode regions that were influential in the decision-making process. They can offer valuable insights into the model's behavior by illuminating the inner workings of its hidden layers.

However, they rely on gradient information which may be unavailable. Moreover, their model-specific nature limits broader applicability.

In contrast, perturbation-based methods such as LIME and SHAP provide greater flexibility as they are model-independent. These methods explain the model's predictive behavior by creating local surrogate models that approximate the original model's behavior in a particular instance's neighborhood. This can reveal how different features contribute to a prediction, which is valuable in EEG systems analysis to understand crucial brain regions and time-frequency features. However, these methods can be computationally intensive, especially for complex models or large datasets. Moreover, fitting local models may raise overfitting concerns, potentially leading to misinterpretations.

Rule-based methods like RF, FIS, and BS prioritize interpretability by using logic rules or mathematical statements. Their transparency allows direct insight into feature contributions, promoting trust and understanding. However, they may oversimplify complex EEG systems by reducing them to incomplete rules. Moreover, accurate interpretation often requires domain expertise, and a balance between interpretability and performance evaluation is necessary, as these methods might not consistently yield the highest predictive accuracy.

In conclusion, selecting an appropriate interpretability method involves the following aspects: the specificity of backpropagation-based method, the computational cost and overfitting risks of perturbation-based methods, and the incompleteness of rule-based methods. The selection should match the specific needs and constraints of the EEG system, the computing resources available, the required level of interpretability, and the expertise of the users.

2.4 Robust AI in EEG systems

AI robustness refers to the ability of AI models to consistently and accurately perform their designated tasks when faced with unexpected conditions. For robust AI in EEG systems, it means effectively countering uncontrollable disturbances across the entire spectrum of EEG signal processing, spanning from the signal sampling phase to the signal input phase. Specifically, the robust AI should adapt to changing brain activity patterns, resist environmental noise, and fill channel

TABLE 2.4: Summary of Robust AI in EEG Systems.

Undesirable Factors	Subcategory	Methods and Representative Works
Noise and Artifacts	External Noise	Traditional Signal Processing [136, 137]
	Internal Artifacts	Models' Self-Robustness [91, 138]
Human Variability	Cross-subject Issues	Transfer Learning [139, 140], Dynamic Domain Adaptation [141]
	Cross-session Issues	Transfer Learning [142, 143], Robust Feature Extraction [141, 144]
Data Acquisition Instability	Resistance Change	Attention Mechanism [145, 146]
	Channel Missing & Broken	Missing Data Reconstruction [137, 147–150]
Adversarial Attacks	Evasion & Manipulation	Adversarial Training [151–154]

gaps from electrode resistance fluctuations. Moreover, model robustness ensures model accuracy in practical applications. For example, physicians rely on precise EEG diagnostics to inform brain disease treatment. Without robust AI providing accurate predictions, the EEG systems lack diagnostic value.

In this section, we classify the undesirable factors that affect EEG systems into four categories: noise and artifacts, human variability, data acquisition instability and adversarial attacks. Based on these four categories, we elaborate the techniques that can alleviate the adverse effects of undesirable factors and improve the robustness of EEG systems. We summarize the robust AI and the representative works in Table 2.4.

2.4.1 Noise and Artifacts in Signals

Brain activity measurement, especially through the EEG signals, is often susceptible to external noises from various sources (for example, electromagnetic interference) and internal artifacts from the human body (for example, muscle movements and eye blinks). These factors distort or interfere with EEG signals, impacting the accuracy and reliability of the EEG system. To address this issue, two types of methods have been developed to minimize the effects of noise and artifacts in EEG signals.

2.4.1.1 Signal Processing

The first type involves traditional signal processing denoising techniques, relying on filtering algorithms developed from prior knowledge to separate and remove noise efficiently. For instance, Kaur *et al.* [136] compared two signal denoising techniques based on discrete wavelet transform (DWT) and wavelet packet transform (WPT)

combined with variational mode decomposition (VMD). The VMD first decomposes the signals into diverse components, and then the DWT and WPT are used to denoise the artifactual components. The WPT with VMD provides a more refined frequency decomposition, facilitating better noise separation and artifact removal compared to its DWT counterpart.

2.4.1.2 Learning-based Denoising

The second type involves the use of modular adaptive denoising techniques in EEG systems, which rely on specialized network structures to automatically denoise the signals. Hussein *et al.* [138] utilized a long short-term memory (LSTM) network to leverage the temporal dependencies in the time series EEG data, and the LSTM can be expressed as

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f), \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i), \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o), \\ \tilde{c}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c), \\ c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t, \\ h_t &= o_t * \tanh(c_t) \end{aligned} \tag{2.19}$$

where f_t, i_t and o_t represent the forget, input and output gates at timestamp t, $W.$ and $b.$ refer to the weight and bias for the respective gates, c_t and $\tilde{c}_t$ denote the cell state and the cell input activation, $\sigma(\cdot)$ denotes the Sigmoid function, $*$ denotes the element-wise product, and h_t is the output hidden state. The robustness of [138] lies in the LSTM's ability to learn long-term dependencies and capture relevant temporal features, thus improving signal-to-noise ratio (SNR) and reducing artifacts.

Building on the inception-time network backbone, Zhang *et al.* [91] proposed an end-to-end framework that takes raw EEG signals as input, eliminating the need for complex signal preprocessing. This noise-insensitive method can capture robust features of motor imagery (MI) tasks and effectively eliminate noise interference. The inception-time network is described as

$$F(x) = \sum_{i=1}^{N} w_i f_i(x), \tag{2.20}$$

where $F(x)$ represents the output of the inception layer, w_i denotes the weights, $f_i(x)$ and N refers to the i-th convolutional layer and the total layer numbers. The robustness of this method mainly stems from the network's ability to learn hierarchical representations of EEG signals and adaptively select relevant features, enhancing the noise suppression capabilities and improving the overall signal quality.

Both traditional signal processing denoising techniques and learning-based denoising methods contribute significantly to enhancing the robustness of EEG data. By focusing on improving signal quality and robustness, it is expected that more accurate and reliable data can be obtained for various applications, including the diagnosis and treatment of neurological disorders, brain-computer interfaces (BCI) and cognitive research.

2.4.2 Human Variability

Variations in EEG signals across different subjects and states pose significant challenges for cross-subject and cross-session EEG systems. These variations mainly arise from the differences in brain anatomy, brain function, and other individual characteristics. Additionally, changes in mental state, fatigue, or recording conditions for the same subject can also lead to substantial disparities in EEG data. To mitigate such impact, AI models need to extract stable features from EEG signals across sessions and subjects.

One common approach to address the EEG variations caused by human variability is transfer learning (TL). For cross-subject EEG, Li *et al.* [139] proposed a multi-source TL method, which utilizes transfer mapping to reduce the difference between known and new subjects, by minimizing the following divergence:

$$D(\mathcal{S}, \mathcal{T}) = \sum_{i=1}^{n} ||\mathbf{w}_i^{\mathcal{S}} - \mathbf{w}_i^{\mathcal{T}}||_2^2, \tag{2.21}$$

where $\mathcal{S}$ and $\mathcal{T}$ denote the source domain and target domain, $\mathbf{w}_i^{\mathcal{S}}$ and $\mathbf{w}_i^{\mathcal{T}}$ represent the weights in the corresponding domains, and $||\cdot||_2$ denotes the ℓ_2-norm.

For cross-session EEG, Lin *et al.* [142] proposed a robust principal component analysis (RPCA)-embedded TL approach, aiming to generate a personalized cross-session model with less labeled data while alleviating intra-session and inter-session differences. The loss function of the TL is given by

$$\min_{\mathbf{L},\mathbf{E}} ||\mathbf{L}||_* + \lambda||\mathbf{E}||_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{L} + \mathbf{E}, \tag{2.22}$$

where $\mathbf{X}$ is the data matrix of EEG signals, $\mathbf{L}$ is the low-rank matrix, $\mathbf{E}$ is the sparse error matrix, and λ is the regularization parameter. $\|\cdot\|_*$ denotes the matrix nuclear norm, which is the sum of the singular values, and $\|\cdot\|_1$ denotes the ℓ_1-norm, which is the sum of the absolute values of entries.

Apart from the TL methods, self-adaptive methods are also used to extract robust EEG features to deal with the human variability. Li *et al.* [141] emphasized the importance of aligning EEG data within the same emotion class for generalizable and discriminative features. They proposed a dynamic domain adaptation (DDA) algorithm, where global and local divergences are handled by minimizing their subdomain discrepancies:

$$\min_{\mathbf{F}_s,\mathbf{F}_t} \mathcal{L}(\mathbf{F}_s, \mathbf{F}_t) = \sum_{c=1}^{C} D_c(\mathbf{F}_s, \mathbf{F}_t) + \alpha \sum_{c=1}^{C} D_{\text{local}}(\mathbf{F}_s^c, \mathbf{F}_t^c), \tag{2.23}$$

where $\mathbf{F}_s$ and $\mathbf{F}_t$ represent the feature representations in the source domain and target domain, $\mathbf{F}_s^c$ and $\mathbf{F}_t^c$ represent the feature representations for the c-th class in two domains, C indicates the total number of classes, α denotes the regularization parameter, D_c computes the global divergence between the two domains for class c, and D_{local} computes the local divergence for each class between two domains. The DDA intends to harmonize the feature representations between the source and target domains.

Furthermore, motivated by the effectiveness of deep learning approaches for stable feature abstractions at higher levels, Yin and Zhang [144] developed an adaptive stacked denoising autoencoder (SDAE) to extract cross-session EEG features. Within this framework, the weights of the first hidden layer are directly connected to the input layer and are updated iteratively. This process accounts for the shifts in the statistical properties of EEG power features observed over consecutive days.

Consequently, the SDAE model is endowed with the proficiency to capture a precise EEG data distribution at a high level.

2.4.3 Data Acquisition Instability

Data acquisition instability refers to the unstable connection between the EEG acquisition equipment and the subject, resulting in the loss of EEG channels. One factor that leads to this instability is the hardening of the glue connecting the electrodes to the scalp over time, thus increasing the resistance of the electrodes. Besides, sweating on the subject's scalp can have a similar effect. In the above two cases, the changes of electrode impedance are difficult to detect, so they cause undetectable channel loss in EEG signals. The primary solution to this issue is to identify the missing channels.

Banville *et al.* [147] proposed dynamic spatial filtering (DSF), a multi-head attention mechanism that focuses on good channels and ignore bad ones. The DSF computes the attention weights for each channel as follows:

$$\alpha_i = \frac{\exp(\mathbf{W}_a \mathbf{X}_i)}{\sum j = 1^N \exp(\mathbf{W}_a \mathbf{X}_j)}, \tag{2.24}$$

where α_i represents the attention weight for the i-th channel, $\mathbf{W}_a$ is the learnable attention matrix, $\mathbf{X}_i$ is the feature vector for the i-th channel, and N is the total number of channels.

Estimating and reconstructing the missing channels is another promising way to address this issue. Bahador *et al.* [145] estimated and reconstructed the data segments in missing channels based on the information near the missing segments. They used a linear weighted interpolation method

$$\mathbf{Y}_m = \sum i = 1^N w_i \mathbf{Y}_{n,i}, \tag{2.25}$$

where $\mathbf{Y}_m$ is the estimated missing channel data, $\mathbf{Y}_{n,i}$ represents the i-th neighboring channel data, w_i is the corresponding weight, and N is the total number of neighboring channels.

2.4.4 New Emerging: Adversarial Attacks

Adversarial attacks on EEG systems have become a growing concern in neuroscience and cybersecurity. EEG systems are widely used in medical-related fields, including pathological diagnosis, control of bionic prosthetic limbs, and communication of severely disabled individuals (for example, amyotrophic lateral sclerosis patients). Since these scenarios involve patient privacy and safety, the vulnerability of EEG systems to adversarial attacks may cause severe medical accidents.

Adversarial attacks on EEG systems typically consist of evasion and manipulation. Evasion involves crafting misleading EEG signals to cause the EEG systems to yield incorrect predictions [154, 155]. For example, an attacker could interfere with users' EEG in a MI task to make bionic prosthetic lose control, which may potentially injure the users or bystanders. As for the manipulation, it means simulating the user's EEG to deceive EEG systems into misinterpreting the user's intentions. In such a scenario, EEG systems may leak individual's personal information or initiate unauthorized financial transactions.

However, real-world adversarial attacks on EEG systems are rather difficult, and thus the defensive methods against these attacks have just begun to be investigated [151]. Adversarial training is a defensive technique that incorporates adversarial examples into the training process to enhance model robustness. Given an EEG input $\mathbf{x}$ and its corresponding label y, the adversarial training aims to minimize the loss function $\mathcal{L}$:

$$\min_{\theta} \mathbb{E}(\mathbf{x}, y)[\mathcal{L}(f(\mathbf{x}_{\text{adv}}; \theta), y)], \tag{2.26}$$

where $\mathbf{x}_{\text{adv}}$ denotes the adversarial example, f represents the model with model parameter θ, and the expectation is taken over the distribution of training data $(\mathbf{x}, y)$. By minimizing the loss function under adversarial perturbations, the model's resilience to adversarial attacks is improved, ensuring the safety and privacy of patients using EEG systems.

2.4.5 Discussion on Robust AI in EEG Systems

Understanding and mitigating the factors that impact the robustness of AI-based EEG systems is vital in the burgeoning field of neural engineering and AI. These factors, including noise and artifacts, human variability, data acquisition instability and adversarial attacks, can critically degrade the performance and reliability of the EEG systems. Each factor arises from unique sources and presents distinct challenges, necessitating a comprehensive and multi-pronged approach to address them effectively.

Noise and artifacts from external and internal physiological sources significantly impede the quality of EEG signal acquisition and interpretation. Externally, diverse noise sources like electromagnetic interference or muscle movements, detrimentally affect the EEG signal fidelity. Internally, noise and artifacts from heart rhythms, eye movements and other biological phenomena can compromise the SNR of the EEG recordings, making them more difficult to analyze. Mitigating such issues requires stringent experimental protocols and robust signal processing algorithms that filter out noise without compromising the integrity of the underlying neural signals.

The inherent variability among human subjects poses another challenge to AI-based EEG systems. This variability can manifest in numerous ways, through the differences in skull thickness and scalp conductivity, cognitive states, and other biological factors. Furthermore, temporal variations, such as changes in a person's mental state or fatigue level, can also impact the EEG signals. Therefore, designing AI-based EEG systems that can generalize across inter- and intra-individual differences is paramount. Solutions involve developing sophisticated machine learning models that account for individual variations or implementing adaptive algorithms capable of adjusting to temporal variations.

Data acquisition instability is another significant factor impacting the robustness of AI-based EEG systems. This instability can stem from technical issues such as changes in electrode-skin resistance, missing data due to broken or disconnected channels, or malfunctioning recording devices. These issues lead to data loss or degradation, significantly hampering the quality and interpretability of the EEG data. Therefore, solutions typically focus on the improvements in hardware and

software, including more stable EEG devices, improved electrode design and materials, and more efficient error detection and error correction algorithms.

Lastly, adversarial attacks substantially threaten the security and integrity of AI-based EEG systems. These attacks often exploit the vulnerabilities in AI models by intentionally manipulating input data, leading to incorrect predictions or classifications. Proactive defenses are necessary to resist such threats, including improving model robustness through adversarial training, implementing rigorous data integrity checks, and developing robust cybersecurity measures.

2.5 Summary

The interpretability and robustness of AI EEG systems is growing in importance and urgency. They ensure the trustworthiness and reliability of EEG systems, and greatly contribute to understanding the models and reproducing the results. This literature review pioneers a comprehensive overview of the interpretable and robust AI techniques designed explicitly for EEG systems. We provide a systemic perspective of this critical field, summarize a wide range of available techniques and tools, and offer an authoritative reference for researchers and practitioners. We introduce new and innovative taxonomies for interpretability and robustness in EEG systems. Throughout the literature review, we summarize the most representative works based on their distinctive contributions, inventive mechanisms, or potential influence on the development of EEG systems. We analyze the technical details, properties and limitations of different works within each category, and also compare their differences across different categories. Highlighting these emerging techniques offers insight into the latest trends of this research area, and also provides a glimpse into their enormous potential and implications for the broader field. In conclusion, this literature review serves as an exhaustive guide for understanding, evaluating and advancing the realm of interpretability and robustness in EEG systems.

Chapter 3

EEG-based Cross Subject Sleep Staging with Hybrid Attention

3.1 Introduction

Sleep staging is a crucial process in evaluating sleep quality and diagnosing sleep disorders, which involves dividing a sleep period into several periodical stages [19, 21, 156, 157]. However, manual sleep staging is time-consuming, subjective, and requires professional expertise, which can lead to unstable and unreliable further sleep disorder diagnosis. Therefore, automated sleep staging methods, including deep learning-based approaches, have been developed to improve efficiency and accuracy.

Despite the progress in automated sleep staging, accurately capturing the spatio-temporal relationships within EEG signals during different sleep stages remains challenging. Previous studies have proposed several methods, such as earlier traditional machine learning methods and current deep learning methods, to enhance the performance of sleep staging systems. However, these methods mainly consider extracting only spatial or temporal features inside the EEG signals, and they have limitations in capturing the complex spatio-temporal relationships of the brain.

To address the above challenges, this thesis proposes a novel hybrid attention EEG sleep staging (HASS) framework. The framework employs a well-designed encoder based on the attention mechanism that adaptively assigns weights to different EEG

segments and channels based on their spatio-temporal relationships during sleep stages. Specifically, the proposed encoder internally contains two components based on the attention mechanism: intra-channel and inter-channel attention. The two components are responsible for capturing the spatial and temporal relationships in sleep EEG signals, respectively. The captured spatial and temporal relationships are further integrated as the spatio-temporal relationships, which can effectively improve sleep staging networks' performance.

The proposed HASS framework has shown promising results in improving the F1 score and accuracy of typical sleep staging networks, as demonstrated by experimental results on the MASS [66] and ISRUC [67] datasets. By capturing the complex spatio-temporal relationships of EEG signals during sleep staging, the HASS framework shows excellent potential for improving the stability and reliability of sleep assessment in both clinical and research settings.

3.2 Related Work

Recognizing different sleep stages is crucial for diagnosing and treating sleep disorders. In the past, support vector machines (SVM) [158] and random forests (RF) [159] were widely used for sleep staging. However, they require extensive prior knowledge and manual feature engineering as well as suffer limited performance. Nowadays, deep learning approaches are the primary method for sleep staging and illustrate better performance.

In the early stage, deep learning methods, such as convolutional neural networks (CNN), are utilized to extract temporal features from sleep signals [160]. For instance, Dong *et al.* [160] proposed the multivariate CNN to capture the temporal features for sleep staging. After that, the recurrent neural networks (RNN) [161, 162], the combination of RNN and CNN [163], and spiking neural networks (SNN) [19] are also applied to extract the temporal representation in different sleep stages. Apart from mining the temporal features to achieve sleep stages classification, some studies have tried to extract the spatial features in the sleep data. Liu *et al.* [164] set electrodes as nodes and mine the spatial features between different channels using relational thinking networks (RTN).

The aforementioned methods only extract temporal and spatial features in sleep data separately, which ignores the spatio-temporal correlation between features. In order to better utilize the spatio-temporal relationships in sleep data, this thesis proposes a HASS framework to enhance the typical sleep staging networks' performance.

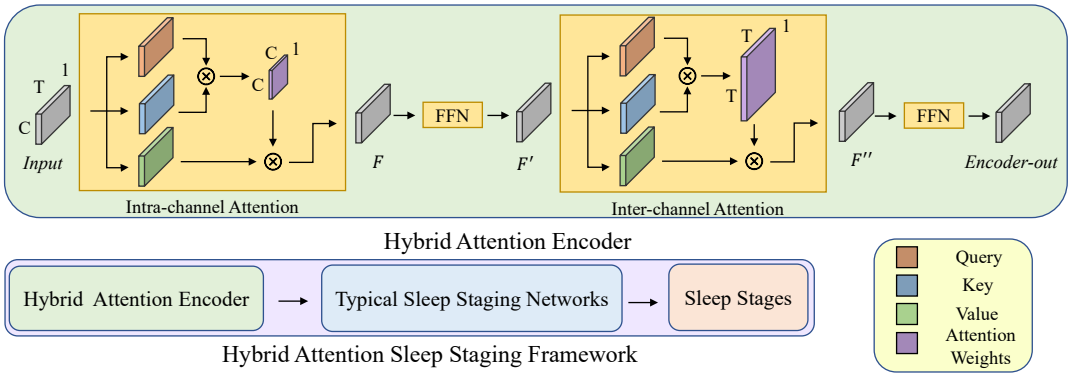

FIGURE 3.1: Overall of the Hybrid Attention Sleep Staging Framework

3.3 Methodology

3.3.1 Description of Matrix Q, K and V

The Query (Q), Key (K), and Value (V) matrices play an essential role in the self-attention mechanism, which is designed to capture long-range dependencies and complex relationships in input time sequences. These matrices are derived from the input representations through linear transformations using three different weight matrices, namely W_Q, W_K, and W_V.

The Q matrix represents the transformed state of the current input element, enabling the model to search for relevant context within the input sequence. In essence, it serves as a basis for comparison against other input elements to determine their relevance. The K matrix encompasses the transformed context representations of the other input elements, allowing for comparing the query and each context element. This comparison is crucial for calculating attention scores, which indicate the relative importance of each element in the time sequence. The V matrix retains the original input element representations used to compute the final attention-weighted output. This output serves as a context-aware representation of

the current input element, considering the relationships between the element and its surroundings.

3.3.2 Hybrid Attention Framework

The hybrid attention framework, as shown in FIGURE 3.1, contains two parts. The first is the novel hybrid attention encoder, and the second is the typical sleep staging network. The hybrid attention encoder captures the spatial and temporal relationships in the EEG signals, and the typical sleep staging network is utilized to achieve the classification. Specially inside the hybrid attention encoder, the intra-channel attention, inter-channel attention, and FFN model are described respectively in the following.

3.3.2.1 Intra-channel Attention

The intra-channel attention encoder can capture the spatial relationship between the channels. We set $Input$ as I, where $I \in \mathbb{R}^{C\times T\times 1}$, the encoder process the $Input$ into $F' \in \mathbb{R}^{C\times T\times 1}$ as follows:

$$F = \mathrm{LN}\left(I + \mathrm{DA}\left(I_{\mathrm{Q}}\,, I_{\mathrm{K}}, I_{\mathrm{V}}\,\right); \Theta, \mathbf{\Phi}\right), \tag{3.1}$$

$$F' = \mathrm{LN}\left(F + \mathrm{FFN}\left(F; \mathbf{\Psi}\right)\right). \tag{3.2}$$

3.3.2.2 Inter-channel Attention

Unlike the intra-channel attention encoder, the inter-channel one makes the DA inside each channel, which can capture the temporal relationship from the sleep signals. The calculations are conducted as follows:

$$F'' = \mathrm{LN}\left(F' + \mathrm{DA}\left(F'_{\mathrm{Q}}, F'_{\mathrm{K}}, F'_{\mathrm{V}}\right); \Theta, \mathbf{\Phi}\right), \tag{3.3}$$

$$Encoder\_out = \mathrm{LN}\left(F'' + \mathrm{FFN}\left(F''; \mathbf{\Psi}\right)\right), \tag{3.4}$$

where LN denotes the layer normalization [165] and DA denotes dot-product attention. To be specific, given query $Q \in \mathbb{R}^{d_k \times N}$, key $K \in \mathbb{R}^{d_k \times N}$, and d_v-dimensional value $V \in \mathbb{R}^{d_v \times N}$ inputs, DA is calculated as:

$$Q^{(n)} = W_Q^{(n)} Q + \boldsymbol{b}_Q^{(n)} \mathbf{1}^\top \in \mathbb{R}^{\frac{d_k}{m} \times N}, \tag{3.5}$$

$$K^{(n)} = W_K^{(n)} K + \boldsymbol{b}_K^{(n)} \mathbf{1}^\top \in \mathbb{R}^{\frac{d_k}{m} \times N}, \tag{3.6}$$

$$V^{(n)} = W_V^{(n)} V + \boldsymbol{b}_V^{(n)} \mathbf{1}^\top \in \mathbb{R}^{\frac{d_v}{m} \times N}, \tag{3.7}$$

$$\mathrm{DA}(Q, K, V; \boldsymbol{\Theta}, \boldsymbol{\Phi}) = W_O \begin{bmatrix} V^{(1)} A^{(1)\top} \\ \vdots \\ V^{(m)} A^{(m)\top} \end{bmatrix} + \boldsymbol{b}_O \mathbf{1}^\top \in \mathbb{R}^{d_v \times N}, \tag{3.8}$$

$$A^{(n)} = \mathrm{softmax}\left(\frac{Q^{(n)\top} K^{(n)}}{\sqrt{d_k / m}} \right) \in (0,1)^{N \times N}, \tag{3.9}$$

where m is the number of heads, $n \in \{1, \ldots, m\}$ is the index the head. The set of parameters Θ and Φ are defined as:

$$\boldsymbol{\Theta} := \bigcup_{1 \le i \le m} \left\{ W_Q^{(n)}, \boldsymbol{b}_Q^{(n)}, W_K^{(n)}, \boldsymbol{b}_K^{(n)} \right\}, \tag{3.10}$$

$$\boldsymbol{\Phi} := \{W_O, \boldsymbol{b}_O\} \cup \bigcup_{1 \le i \le m} \left\{ W_V^{(n)}, \boldsymbol{b}_V^{(n)} \right\}. \tag{3.11}$$

3.3.3 Feed-forward Network

There are two FFN models in the hybrid attention sleep staging framework. Each FFN model consists of two dense layers and can be calculated as:

$$\begin{aligned} \mathrm{FFN}_1(F; \boldsymbol{\Psi}) &= \left(W_2 \left[W_1 F + \boldsymbol{b}_1 \mathbf{1}^\top \right]_+ + \boldsymbol{b}_2 \mathbf{1}^\top \right), \\ \boldsymbol{\Psi} &:= \{W_1, \boldsymbol{b}_1, W_2, \boldsymbol{b}_2\}. \end{aligned} \tag{3.12}$$

$$\begin{aligned} \mathrm{FFN}_2(F'; \boldsymbol{\Psi}) &= \left(W_4 \left[W_3 F' + \boldsymbol{b}_3 \mathbf{1}^\top \right]_+ + \boldsymbol{b}_4 \mathbf{1}^\top \right), \\ \boldsymbol{\Psi} &:= \{W_3, \boldsymbol{b}_3, W_4, \boldsymbol{b}_4\}. \end{aligned} \tag{3.13}$$

where $W_1, W_3 \in \mathbb{R}^{d_f \times D}$ and $W_2, W_4 \in \mathbb{R}^{D \times d_f}$ are mapping matrices, $\boldsymbol{b}_1, \boldsymbol{b}_3 \in \mathbb{R}^{d_f}$ and $\boldsymbol{b}_2, \boldsymbol{b}_4 \in \mathbb{R}^D$ are biases, and $[\cdot]_+$ is the unit slope function.

3.4 Experiments and Results

3.4.1 Dataset

To thoroughly evaluate the efficacy of the HASS framework, we conduct experiments using two well-established datasets: the Institute of Systems and Robotics of the University of Coimbra (ISRUC) and the Montreal Sleep Study Archive-SS3 (MASS) datasets. The ISRUC dataset comprises PSG recordings from 100 adult subjects, each with 6 EEG channels. Meanwhile, the MASS dataset includes PSG records from 62 adult subjects, each with 20 EEG channels. To standardize the data, we divide the recordings into time slices corresponding to sleep epochs, each representing 30 seconds of sleep. To ensure the accuracy of our evaluations, we enlist the expertise of sleep specialists who manually classify each time slice into one of five different sleep stages: Wake (W), N1, N2, N3, and REM, following the AASM criteria [166].

3.4.2 Settings

In our experiment, the input is denoted as $I \in \mathbb{R}^{C\times T\times 1}$. Specifically, C denotes the number of EEG channels, with 6 channels in the ISRUC dataset and 20 channels in the MASS dataset. T signifies the time slices corresponding to sleep epochs, each denoting 30 seconds of sleep. Regarding intra-channel attention, the DA is calculated on channel dimension. Conversely, the DA is computed along the time slices dimension for inter-channel attention.

3.4.3 Result and Analysis

Table 3.2 and Table 3.1 present the comparison of the performance of four typical sleep staging networks, namely TinySleepNet (TSN) [167], DeepSleepNet (DSN) [163], MCNN [160], and U-Time [168], on the MASS and ISRUC datasets, before and after applying the proposed HASS framework. The performance of the networks is evaluated using overall F1 scores, accuracies, and F1 scores for each sleep stage.

TABLE 3.1: Performance comparison (F1 Score and Accuracy) of ISRUC dataset between the original four sleep staging networks and after applying HASS.

Baselines	HASS	ISRUC Dataset						
		F1	Acc	W	N1	N2	N3	REM
TSN [167]	Yes	**0.774**	**0.794**	**0.891**	**0.576**	**0.874**	**0.884**	**0.847**
	No	0.751	0.769	0.868	0.545	0.855	0.878	0.828
DSN [163]	Yes	**0.755**	**0.781**	**0.883**	**0.535**	**0.861**	**0.889**	**0.831**
	No	0.731	0.756	0.852	0.527	0.829	0.879	0.796
MCNN [160]	Yes	**0.782**	**0.801**	**0.902**	**0.603**	**0.861**	0.893	**0.842**
	No	0.764	0.782	0.885	0.571	0.851	**0.894**	0.833
U-Time [168]	Yes	**0.727**	**0.769**	**0.869**	0.524	**0.813**	**0.885**	**0.769**
	No	0.713	0.755	0.844	**0.525**	0.793	0.861	0.755

TABLE 3.2: Performance comparison (F1 Score and Accuracy) of MASS dataset between the original four sleep staging networks and after applying HASS.

Baselines	HASS	MASS Dataset						
		F1	Acc	W	N1	N2	N3	REM
TSN [167]	Yes	**0.811**	**0.881**	**0.885**	**0.565**	**0.892**	0.847	**0.921**
	No	0.798	0.858	0.873	0.547	0.888	**0.848**	0.889
DSN [163]	Yes	**0.791**	**0.855**	**0.873**	**0.528**	**0.888**	0.821	**0.883**
	No	0.773	0.834	0.861	0.514	0.861	**0.825**	0.864
MCNN [160]	Yes	**0.818**	**0.875**	**0.895**	**0.553**	**0.918**	**0.836**	**0.917**
	No	0.804	0.862	0.884	0.548	0.898	0.825	0.897
U-Time [168]	Yes	**0.794**	**0.873**	**0.885**	**0.528**	**0.891**	**0.813**	**0.883**
	No	0.782	0.854	0.865	0.517	0.882	0.801	0.875

The results demonstrate that HASS can significantly improve the performance of the typical sleep staging networks. Specifically, for the MASS dataset, all four networks achieved higher overall F1 scores and accuracies when using HASS. The improvement in F1 scores for each stage is consistent across all networks. After applying HASS, the performance of the networks for all stages is improved. Notably, the most significant improvement was observed for the W stage, with all four networks achieving significantly higher F1 scores with HASS over the original networks. For instance, TSN and DSN achieved F1 scores of 0.885 and 0.873, respectively, with HASS, compared to 0.873 and 0.861 without HASS. Similarly, MCNN and U-Time achieved F1 scores of 0.895 and 0.885, respectively, with HASS, compared to 0.884 and 0.865 without HASS.

Likewise, for the ISRUC dataset, HASS improved the performance of all four networks, with higher overall F1 scores and accuracies over the original networks.

However, the improvement in F1 scores is less significant compared to the MASS dataset, and three out of four networks achieved higher F1 scores for the W stage with HASS. Nevertheless, the improvement in F1 scores for all stages is consistent across all networks.

The overall F1 scores and accuracy improvement are significant and consistent across all networks and for all sleep stages. These results suggest that the HASS framework can be highly effective in enhancing the accuracy of sleep staging networks, which is crucial for accurately diagnosing and treating sleep disorders. With these improvements, the HASS framework has the potential to significantly enhance the quality of sleep staging in clinical settings, thereby improving the overall quality of patient diagnosis and care.

3.4.4 Ablation Studies

To evaluate the contributions of individual components within the HASS framework, we conducted ablation studies focused on isolating the impact of the spatial and temporal attention mechanisms.

- **Spatial-only Model**: A variant of HASS with only the spatial attention mechanism enabled.
- **Temporal-only Model**: A variant of HASS with only the temporal attention mechanism enabled.
- **Baseline Models**: Including TSN, DSN, MCNN, and U-Time architectures for performance benchmarking.

The ISRUC dataset results, as shown in Figure 3.2, demonstrate the effectiveness of the HASS framework compared to its ablated variants and baseline models. Specifically: The original HASS framework achieved the highest accuracy of **80.1%**, outperforming both the spatial-only (78.7%) and temporal-only (78.5%) variants. Among the baseline models, MCNN achieved 77.4%, while U-Time reached 76.5%, both significantly lower than the full HASS framework. The results highlight that the integration of both spatial and temporal attention mechanisms contributes significantly to the overall performance.

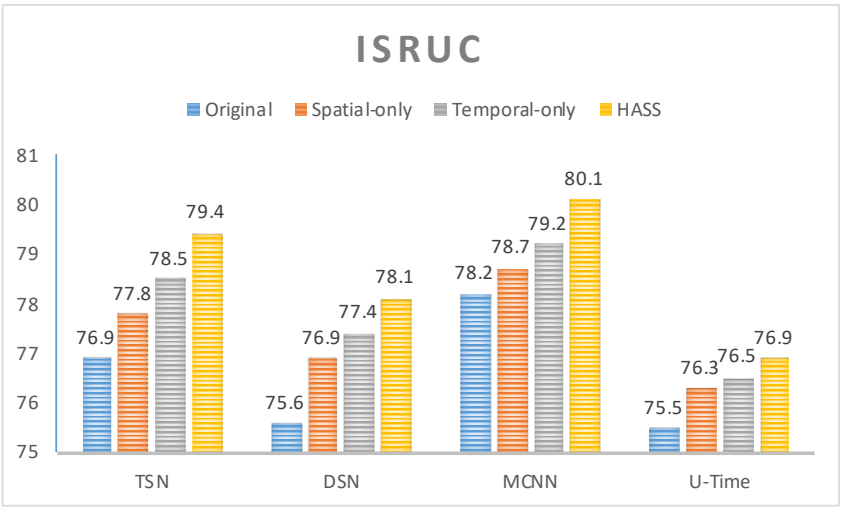

FIGURE 3.2: Overall of the Hybrid Attention Sleep Staging Framework

The MASS dataset results, summarized in Figure 3.3, further validate the robustness of the HASS framework. Key findings include: The full HASS model achieved the best performance with an accuracy of **88.1%**, surpassing the spatial-only (86.5%) and temporal-only (86.8%) configurations. Among the baseline models, MCNN achieved 86.2%, while U-Time recorded 85.5%, both trailing behind the HASS framework. These results emphasize the importance of jointly leveraging spatial and temporal attention for improved sleep staging performance.

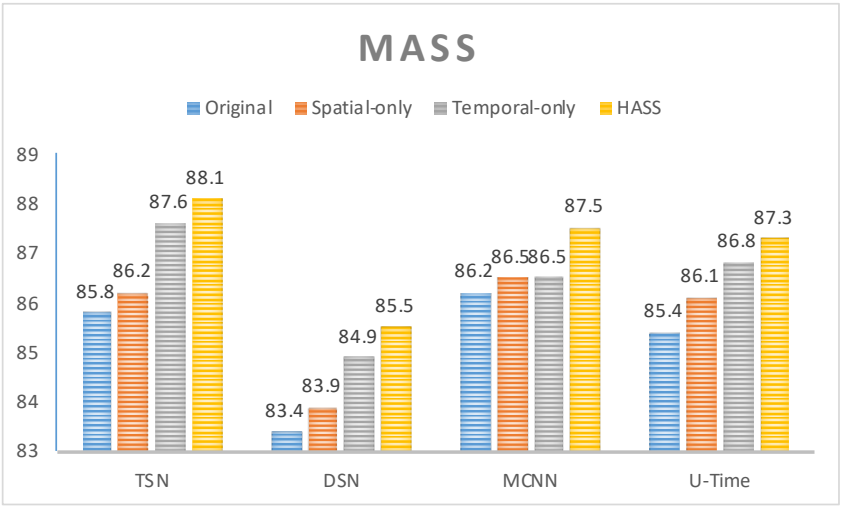

FIGURE 3.3: Overall of the Hybrid Attention Sleep Staging Framework

The ablation studies confirm that the HASS framework's superior performance

stems from the combined contributions of its spatial and temporal attention mechanisms. While the spatial-only and temporal-only models exhibit competitive performance, their integration within the HASS framework significantly enhances accuracy, demonstrating the synergistic benefits of jointly modeling spatial and temporal dependencies in EEG sleep staging.

3.5 Summary

The proposed HASS framework effectively improves the performance of typical sleep staging networks on two different datasets: MASS and ISRUC. The experiment results demonstrate that HASS can significantly enhance the overall F1 scores and accuracies of the networks and improve the performance of the networks for all sleep stages. Specifically, the most remarkable improvement was observed for the W stage on the MASS dataset. These findings suggest that the HASS method has the potential to enhance the accuracy of sleep staging networks, which is critical for the accurate diagnosis and treatment of sleep disorders.

Chapter 4

Learning Robust Global-Local Representation from EEG for Neural Epilepsy Detection

4.1 Introduction

Epilepsy is one of the most prevalent neurological disorders worldwide, affecting individuals across all age groups. This chronic condition, characterized by recurrent seizures, can significantly impair cognitive and mental functions and, in severe instances, may be life-threatening. Consequently, epilepsy is a significant concern for healthcare professionals globally, underscoring the imperative for precise diagnostic methods. Clinical detection and diagnosis of epilepsy typically employ various imaging modalities, including Electroencephalography (EEG), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Magnetoencephalography (MEG). While MRI, CT, and MEG are instrumental in identifying secondary epilepsy resulting from intracranial lesions or injuries, they offer limited diagnostic value for primary epilepsy in the absence of solid intracranial lesions. Conversely, EEG, as a noninvasive, readily accessible, and cost-effective method, has emerged as the principal tool for the clinical diagnosis of epilepsy. However, the extensive volume of EEG data presents a significant challenge to neurologists and clinicians, necessitating meticulous analysis that is both time-consuming and subjective. This

highlights the critical need for advanced automated EEG classification and detection algorithms. Developing and implementing these adaptive algorithms have become a focal point of contemporary research, aiming to alleviate the diagnostic burden on healthcare professionals and enhance the accuracy and efficiency of epilepsy diagnosis.

With the help of deep learning, automatic epilepsy detection with EEG has earned more and more attention [169–171]. Specifically, neural network models drawing upon the architectures of the attention mechanism and convolutional neural networks (CNN) have been widely adopted in epilepsy detection tasks that utilize EEG. To be specific, the attention mechanism was first renowned for its good performance in capturing long-range dependencies in context and sequence data [172, 173]. It has been widely applied to the field of EEG epilepsy detection [174, 175], extracting the spatio-temporal relationship features associated with epilepsy seizure. In contrast, CNN-based neural networks present a different approach to analyzing time series. These networks prioritize capturing local features and can pay particular attention to signal waveforms of different frequencies in EEG signals [176, 177]. It is facilitated through layered convolution operations, allowing CNN to delve into the intricacies of the EEG signals and extract high-dimensional representations [178]. These representations play an important role in detecting and classifying epilepsy. Both attention-based networks and CNN show good capability in achieving EEG epilepsy detection, respectively [179, 180]. However, epilepsy signals have complex patterns, and only mining either global or local dependencies could suffer limited performance.

Specifically, in detecting epilepsy via time-series EEG signals, both global and local information play a non-negligible role. Global feature usually reflects long-distance dependencies in the brain signal segment, whereas local feature focuses more on subtle changes. Regarding global information, it has been found that the spike-wave is different in the pre-ictal and inter-ictal periods and that there is an increasing trend in the spike-wave in the period close to the epilepsy seizure [181]. Based on this phenomenon, it is possible to capture the number of spikes detected in the global EEG signal to predict seizures. There are also methods to monitor the dynamics of the EEG signal by analyzing the similarity between the current period of the EEG of an epilepsy patient and two reference periods [182], which compares the difference between the patient's current EEG state and

the reference state of that patient's historical interictal and pre-seizure periods respectively. Turning to local information, some localized abnormal waveforms and sudden changes in amplitude can help detect epilepsy waves. For instance, areas with a lot of waveform jitter in EEG time-domain signals may be able to serve as biomarkers for the seizure onset zone (SOZ), which is made up of short-lived, violent oscillations of bioelectrical activity seen in the intracranial electroencephalogram (iEEG) [183]. Initially, this localized feature could be detected using a root mean square (RMS) detector [184].

Therefore, to comprehensively leverage the interleaved global and local EEG signals' epilepsy features to more precisely and robust epilepsy detection, we propose th End-to-End Neural Epilepsy Detection (EENED)/ Global-Local Neural Epilepsy Detection Network (GlepNet). The proposed method is based on an encoder with interleaved convolution and attention mechanism structure, as graphically represented in FIGURE 4.5. The architecture of EENED/GlepNet draws inspiration from previous works [185], and the encoder blocks within the model employ the self-attention mechanism [186, 187]. It is applied to discern the complex and interleaved temporal relationships within epilepsy EEG signals, improving the model's performance in epilepsy detection tasks. To be specific, we have eliminated the traditional multi-head attention mechanism's positional encoding [188, 189]. Additionally, we have partitioned the traditional feed-forward layer into two sections, creating a unique sandwich structure. Meanwhile, we have incorporated a hybrid convolution module within the encoder blocks to extract local features that might otherwise be overlooked by the multi-head attention model and thus enhance the model's performance. The designed hybrid convolution module comprises one-dimensional depth-wise and one-dimensional point-wise convolution. This combination allows for more nuanced analyses of local features in EEG signals that may be linked to epilepsy, such as abnormal waveforms and abrupt changes in frequency and amplitude. Furthermore, to verify the effectiveness and interpretability of EENED/GlepNet, we utilized the Grad-CAM algorithm for visualization [124]. This visually intuitive method validates the dependability of our method by highlighting its heightened capability to detect time segments related to epilepsy episodes precisely [190]. Specifically, our visualization demonstrates that the proposed method can capture the abnormalities, ranging from the sudden onset of spikes to the dispersion of sharp and slow waves, which are hallmark indicators of epileptic manifestations, which have been proven in neuroscience [191]. This form

of visualization offers clinicians a transparent and interpretable view of the EEG signals [192].

The experiments demonstrate EENED/GlepNet's superior performance. Regarding accuracy and F1 score, EENED/GlepNet surpasses the performance of both traditional CNN and attention based networks. The increased performance is attributed to EENED/GlepNet's ability to not only capture the long-range dependencies in temporal EEG signals but also combine with extracting local signal features through the convolution module. Furthermore, we make the ablation study and comparative analysis to discuss the influence of the convolutional blocks, the multi-head attention block, as well as the number of encoder blocks. The given result demonstrates that the above blocks are all necessary and that the setting of the three encoder blocks is suitable. EENED/GlepNet not only exhibits reliability in epilepsy detection tasks but also displays significant potential for further advancements and applications in this field. By offering a more comprehensive utilization of the global and local dependencies of the EEG signals, EENED/GlepNet can substantially enhance diagnostic accuracy.

The contributions of this section are summarized as follows.

- To the best of our knowledge, this is the first attempt to interleave multi-head attention mechanism and convolution techniques within a framework for capturing robust global-local representation, thereby enhancing EEG neural epilepsy detection.
- The proposed EENED/GlepNet achieve the state-of-the-art performance on four open-source EEG epilepsy datasets in terms of binary epilepsy classification.
- We employ the interpretable technique of Grad-CAM to visually substantiate that EENED/GlepNet are adept at discerning the global-local representations essential for robust epilepsy detection.

4.2 Related Work

Neural epilepsy detection and management are crucial research fields, with EEG established as a potent diagnostic tool [193, 194]. Given the voluminous nature of

EEG data, it becomes impractical for neurologists to analyze it efficiently [195]. This has spurred the development of automated epilepsy detection algorithms to alleviate the workload on physicians, marking a significant research direction [196]. The innovation in wearable EEG technologies has been particularly noted by Casson *et al.*, who explored their use in real-time monitoring and emphasized their translational value from laboratory to clinical settings [197].

Epilepsy detection methods broadly fall into two categories: manual feature engineering and deep learning approaches. The former relies on signal processing techniques such as Fourier transform, wavelet transform [198, 199], local mean decomposition (LMD) [200, 201], and empirical mode decomposition (EMD) [202] to extract features. These features are then classified using methods like SVMs and random forests [203, 204]. Although these techniques are comprehensive, they can be complex and not sufficiently effective.

In contrast, deep learning techniques, particularly CNNs, have proven effective in utilizing EEG data due to their ability to capture localized patterns [205]. Recent advancements have introduced attention-based networks, such as transformers, which were initially designed for NLP tasks but have since adapted for time-series analysis [1]. These networks excel in recognizing long-range dependencies within the spatio-temporal data of EEG signals. Attention to the design of such models, including the role of positional encoding, has been critical in improving performance [188]. Hybrid models, combining CNNs and transformer architectures, like those developed by Gulati *et al.* and Song *et al.*, integrate localized feature extraction with global contextual understanding, offering a promising direction for robust epilepsy detection [89, 185].

From a broader perspective, understanding both the global and local aspects of EEG data is essential. Globally, studies have noted variations in spike-wave patterns during different phases of epileptic activity, with frequency increases preceding seizures [181]. Locally, identifying specific abnormal waveforms and amplitude changes can pinpoint the seizure onset zones (SOZ), where intense oscillations in bioelectrical activity occur, as seen in intracranial EEGs [183]. Therefore, leveraging both global and local EEG features could lead to more accurate and comprehensive epilepsy detection systems.

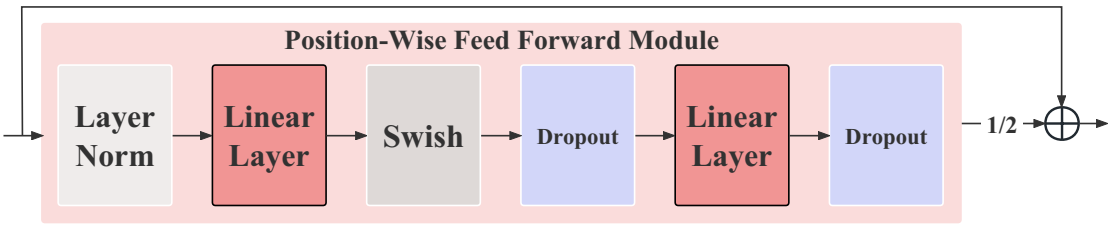

FIGURE 4.1: **Position-wise feed-forward model.** Two linear layers increase and decrease the dimensionality of the data, respectively.

4.3 EENED

4.3.1 Methodology

4.3.1.1 Encoder Blocks

Inspired by Macaron-Net [206], the encoder blocks of EENED adopt a sandwich structure, which divides the feedforward layer in the Transformer encoder into two half-step residual feedforward units. Mathematically, for given input $(f^t_{(e-1)}|t = 1, ..., T)$ to the e^{th} encoder block, the output $(f^t_{(e)}|t = 1, ..., T)$ of the block is:

$$f^{(e)}_{FF_1} = (f^t_{(e-1)}|t = 1, ..., T) + \frac{1}{2}\,\mathrm{PWFF}((f^t_{(e-1)}|t = 1, ..., T)) \tag{4.1}$$

$$f^{(e)}_{MHSA} = f^{(e)}_{PWFF_1} + \mathrm{MHSA}(f^{(e)}_{PWFF_1}) \tag{4.2}$$

$$f^{(e)}_{Conv} = f^{(e)}_{MHSA} + \mathrm{Conv}(f^{(e)}_{MHSA}) \tag{4.3}$$

$$(f^t_{(e)}|t = 1, ..., T) = \mathrm{LayerNorm}(f^{(e)}_{Conv} + \frac{1}{2}\,\mathrm{PWFF}(f^{(e)}_{Conv})) \tag{4.4}$$

Where PWFF() denotes the position-wise feed-forward module, MHSA() denotes to the multi-head self-attention module, and Conv() denotes to the convolution module.

4.3.1.2 Position-wise feed-forward module

The Transformer architecture [207] employs a feed-forward module before and after the MHSA module, consisting of two linear layers and activation as shown

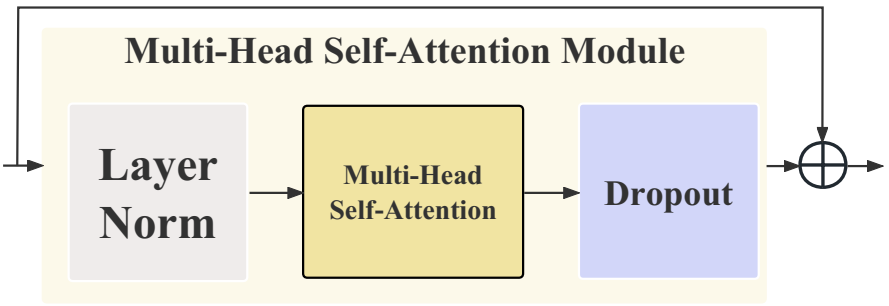

FIGURE 4.2: **Multi-head self-attention module.** We use a multi-head self-attention similar to the transformer encoder but remove the relative positional embedding in this pre-norm residual unit.

in FIGURE 4.1. The e^{th} PWFF module transforms a sequence of input vectors $(f_{(e-1)}^{t}|t = 1, ..., T)$ from the previous encoder as follows:

$$F^{(e)} = \text{LayerNorm}([f_t^{(e-1)} \cdots f_T^{(e-1)}]) \in \mathbb{R}^{T \times D} \tag{4.5}$$

$$\bar{F}^{(e)} = \text{Swish}(F^{(e)} W_1^{(e)} + 1 b_1^{(e)^\top}) W_2^{(e)} + 1 b_2^{(e)^\top} \in \mathbb{R}^{T \times D} \tag{4.6}$$

Where T is the length of time and D is the feature dimension. $W_1^{(e)} \in \mathbb{R}^{D \times d_{pwff}}$ and $b_1^{(e)} \in \mathbb{R}^{d_{pwff}}$ are the projection matrix and bias of the first linear layer, $1 \in \mathbb{R}^T$ is an all-one vector. d_{pwff} is the dimension of hidden units. We also apply Swish activation [208] Swish$(\cdot)$ and dropout [209] to help regularize the module. $W_2^{(e)} \in \mathbb{R}^{D \times d_{pwff}}$ and $b_2^{(e)} \in \mathbb{R}^{d_{pwff}}$ are the second linear projection matrix and bias. Following the pre-norm residual units [210], the final output of the PWFF module is computed as follows:

$$F_{PWFF}^{(e)} = F^{(e)} + \frac{\text{Dropout}(\bar{F}^{(e)})}{2} \in \mathbb{R}^{T \times D} \tag{4.7}$$

4.3.1.3 Multi-head self-attention module

The structure of the multi-head self-attention module is shown in FIGUE 4.2. The MHSA module in the e^{th} encoder processes features from the PWFF module. The input features $F_{PWFF}^{(e)}$ is converted by layer normalization:

$$\bar{F}^{(e)} = \text{LayerNorm}(F_{PWFF}^{(e)}) \in \mathbb{R}^{T \times D} \tag{4.8}$$

Then, in the multi-head self-attention block, each attention head computes a pair-wise similarity matrix $S_h^{(e)}$ applying the dot products of query vectors $\bar{F}^{(e)}Q_h^{(e)} \in \mathbb{R}^{T\times d}$ and key vectors $\bar{F}^{(e)}K_h^{(e)} \in \mathbb{R}^{T\times d}$

$$S_h^{(e)} = \bar{F}^{(e)}Q_h^{(e)}(\bar{F}^{(e)}K_h^{(e)})^\top \in \mathbb{R}^{T\times T}(1 \le h \le H) \tag{4.9}$$

where H denoted the number of heads. $Q_h^{(e)}$,$K_h^{(e)} \in \mathbb{R}^{D\times d}$ are the h^{th} head's query and key projection matrices, respectively. Scaled by $1/\sqrt{D/H}$, $S_h^{(e)}$, the matrix of similarity and softmax are applied to produce the attention weight matrix $A_h^{(e)}$:

$$A_h^{(e)} = \text{Softmax}\left(\frac{S_h^{(e)}}{\sqrt{D/H}}\right) \in \mathbb{R}^{T\times T} \tag{4.10}$$

Then the attention weight matrix $A_h^{(e)}$ is used to compute context vectors $C_h^{(e)}$ with the value vectors $\bar{F}^{(e)}V_h^{(e)} \in \mathbb{R}^{T\times d}$:

$$C_h^{(e)} = A_h^{(e)}(\bar{F}^{(e)}V_h^{(e)}) \in \mathbb{R}^{T\times d} \tag{4.11}$$

where $V_h^{(e)} \in \mathbb{R}^{D\times d}$ represents the matrix of value projection. The output features of the multi-head self-attention module are obtained by concatenating the context vectors from all heads and then applying an output projection matrix $O^{(e)} \in \mathbb{R}^{D\times D}$:

$$\bar{F}_{MHSA}^{(e)} = [C_1^{(e)} \cdots C_H^{(e)}]O^{(e)} \in \mathbb{R}^{T\times D} \tag{4.12}$$

$$F_{MHSA}^{(e)} = \text{LayerNorm}(\bar{F}^{(e)} + \text{DropOut}(\bar{F}_{MHSA}^{(e)})) \in \mathbb{R}^{T\times D} \tag{4.13}$$

4.3.1.4 Convolution module

Similar to [211], the convolution module, as shown in FIGURE 4.3, takes $F_{MHSA}^{(e)}$ as input and starts with a point-wise convolution and a gated linear unit (GLU) [212]:

$$\bar{F}^{(e)} = \text{PWConv}(\text{LayerNorm}(F_{MHSA}^{(e)})) \in \mathbb{R}^{T\times 2D} \tag{4.14}$$

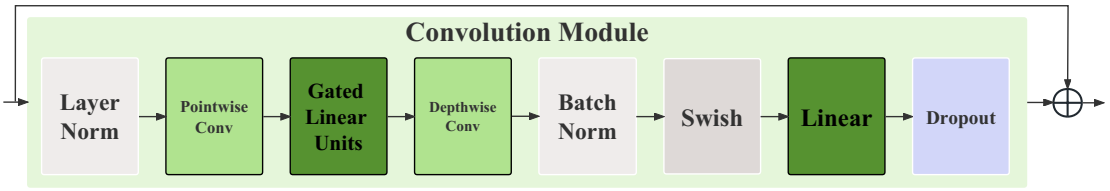

FIGURE 4.3: **Convolution module.** The convolution module contains two convolutional layers of different scales, with Gated linear units used as the activation layer in the middle. The Swish activation function is used, followed by a linear layer.

$$\bar{F}_{glu}^{(e)} = (\hat{F}^{(e)} W_1^{(e)} + b_1^{(e)}) \bigotimes \sigma(\check{F}^{(e)} W_2^{(e)} + b_2^{(e)}) \in \mathbb{R}^{T \times D} \tag{4.15}$$

where PWConv$(\cdot)$ is a one-dimensional point-wise convolutional layer with kernel size (1 X 1) and stride of 1. The output $\bar{F}^{(e)} \in \mathbb{R}^{T \times 2D}$ could be divided into $\hat{F}^{(e)} \in \mathbb{R}^{T \times D}$ and $\check{F}^{(e)} \in \mathbb{R}^{T \times D}$, which are the first half and the second half of the output, respectively. $W_1^{(e)} \in \mathbb{R}^{D \times D}$, $b_1^{(e)} \in \mathbb{R}^D$,$W_2^{(e)} \in \mathbb{R}^{D \times D}$, $b_1^{(e)} \in \mathbb{R}^D$ are learned parameters, σ is the sigmoid function and $\bigotimes$ is the element-wise product. The output of GLU $\bar{F}_{glu}^{(e)}$ is processed with a DWConv$(\cdot)$:

$$\bar{F}_{DWConv}^{(e)} = W_{conv}^{(e)} \, \text{Swish}(\text{DWConv}(\bar{F}_{glu}^{(e)})) + b_{Conv}^{(e)} \in \mathbb{R}^{T \times D} \tag{4.16}$$

where $W_{Conv}^{(e)} \in \mathbb{R}^{d_{pwconv} \times D}$, $b_{Conv}^{(e)} \in \mathbb{R}^D$ are learned linear parameters of Convolution module. Swish$(\cdot)$ is the activation function. DWConv$(\cdot)$ is a one-dimensional depth-wise convolutional layer with a kernel size of 15. d_{pwconv} is the dimension of depth-wise convolution layer output. The final output features $F_{Conv}^{(e)}$ of the convolution module are as follows:

$$F_{Conv}^{(e)} = F_{MHSA}^{(e)} + \text{Dropout}(\bar{F}_{DWConv}^{(e)}) \in \mathbb{R}^{T \times D} \tag{4.17}$$

4.3.2 Experiments and Results

4.3.2.1 Dataset

The Epileptic Seizure Recognition dataset [59] contains EEG recordings from 500 subjects. The brain activity of each subject was recorded for 23.6 seconds. Since

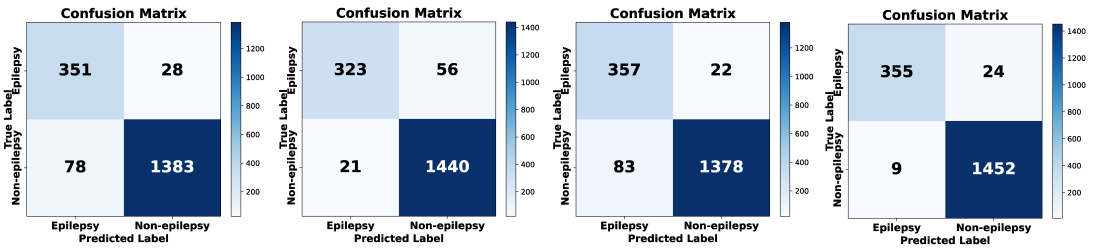

FIGURE 4.4: Confusion matrix of the predicted results of the four models. From left to right: Dense-CNN, CNN-LSTM, Transformer, and EENED.

four of the five categories are unrelated to epilepsy, we reduced the labels to one category. The training set contains 7360 segments of EEG signal data, and the test set contains 1840 segments of EEG signal data, of which 1461 are non-epileptic EEG signals. We followed the data processing in [213].

4.3.2.2 Model configuration

- **EENED**

  EENED contains three Encoder blocks, and each Encoder layer contains a self-attention module and a convolution module. The attention mechanism consists of 8 attention heads, each with a dimension of 64. The convolution module uses one-dimensional convolution layers and GLU activation function; the convolution kernel size is 15, the step size is 1, the padding is 7, and the number of groups is 512. The model also contains two feed-forward layers and residual connections, and finally, it applies two linear layers and a sigmoid activation function to classify the features.

- **Dense-CNN**

  Dense-CNN [214] is comprised of multiple convolutional layers, with each layer having a different set of convolutional filters. The first is a linear layer that outputs a tensor of size 512. The subsequent layers use an architecture called Dense-Inception which is made up of several Inception modules. Each Inception module consists of multiple branches that process the input in parallel using different convolutional filters of different kernel sizes. The output of each branch is then concatenated along the channel dimension and fed into a 1x1 convolutional layer to reduce the number of channels. The output of one Inception module is then fed into the next Inception module,

and the process is repeated until the final layer, which outputs a tensor of size 18.

- **Transformer**

  This network architecture is a Transformer-based classifier consisting of an upsample layer, two linear layers, Transformer-Encoder layers with 3 Transformer encoders, a fully connected layer, and a sigmoid activation function. The upsample layer maps the input sequence to a hidden state of size 512. The first linear layer maps the hidden state to a single value, which is then used to scale the output of the Transformer-Encoder layer. The Transformer-Encoder layer contains a self-attention mechanism and two linear and dropout layers. The self-attention mechanism uses 8 attention heads and an output projection matrix of size 512. The dropout probability is set to 0.1 for the attention and linear layers. The fully connected layer maps the output of the Transformer-Encoder to a single value, which is then passed through the sigmoid activation function to produce the final classification output.

- **CNN-LSTM**

  CNN-LSTM [215] comprises a convolutional neural network and a long and short-term memory (LSTM) layer. The convolutional neural network consists of two convolutional layers and a maximum pooling layer, and the output size of the fully connected layer is 512. The input of the LSTM layer is the output of the fully connected layer and contains two LSTM layers with a hidden state size of 128.

4.3.3 Result and Analysis

As shown in FIGURE 4.4, from the prediction result confusion matrix of the four models, the accuracy rates of Dense-CNN and Transformer are 0.942 and 0.943, respectively. The accuracy rates of CNN-LSTM and EENED are 0.958 and 0.982, respectively. Among them, EENED achieved the highest accuracy rate. When Dense-CNN is used to process time-domain signals, it cannot perform better due to the lack of a mechanism for processing sequence data. Transformer is good at global modeling of time-domain data. However, it is challenging to capture short sequence signal features related to epilepsy in EEG signals, such as the appearance

TABLE 4.1: Comparison of performance (F1 Score and Accuracy) of four neural networks on The Epileptic Seizure Recognition dataset.

Network	**Accuracy**	**F1 Score**
Dense-CNN[214]	0.942	0.963
Transformer	0.943	0.963
CNN-LSTM[215]	0.958	0.974
EENED	**0.982**	**0.989**

of spikes and the sprinkling of sharp and slow waves. In contrast, CNN-LSTM integrates local and global feature extraction to a certain extent, which can effectively capture the temporal dependence of time series data. EENED has the highest accuracy rate in the experimental results. It combines the characteristics of CNN and Transformer. While learning the global temporal dependence of epilepsy-related EEG signals, it captures local signal features through the convolution module.

4.3.4 Summary

EENED achieves higher accuracy in epilepsy detection tasks than other transformer and CNN-based neural networks. The experimental results show that EENED can combine the ability of the self-attention mechanism to build the long-term dependence of EEG signals and the characteristics of the convolution module to extract local EEG signal features, which is crucial for using EEG to detect epilepsy as a reference for medical diagnosis.

4.4 GlepNet

4.4.1 Methodology

4.4.1.1 Overall GlepNet

The proposed GlepNet architecture integrates traditional attention-based and CNN structures into a cohesive framework for neural epilepsy detection as shown in FIGURE. 4.5. At the core of this architecture are several encoder blocks, each combining interleaved multi-headed self-attention with convolution modules. This design allows for the extraction of robust, interleaved global-local EEG epilepsy

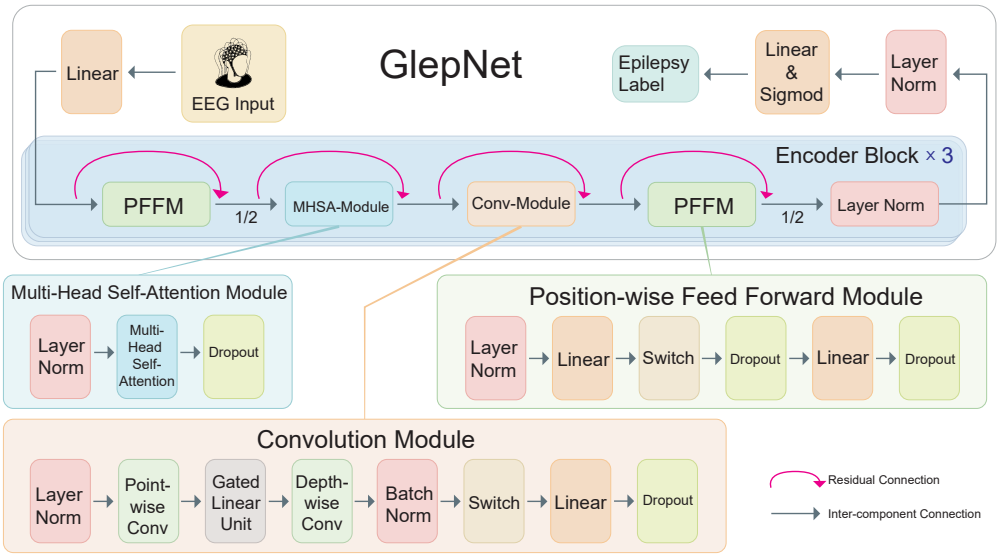

FIGURE 4.5: Global-Local Neural Epilepsy Detection Network Architecture. GlepNet incorporates a linear layer for feature embedding, followed by specialized encoder blocks that enhance the detection of global-local epilepsy patterns. Each encoder block includes multi-head self-attention and convolution modules, positioned between two macaroon-like, position-wise feed-forward layers with a half-step residual connection designed to capture interleaved EEG epilepsy features effectively. A normalization layer addresses the gradient vanishing issue, while the final linear and sigmoid layers are dedicated to accurate epilepsy detection.

features. Each encoder block is fortified with macaroon-style feedforward layers that enhance the processing of these representations.

GlepNet utilizes half-step residual connections within these blocks to stabilize gradients and facilitate learning across the stacked encoders. The attention modules in each encoder are crucial for capturing long-range dependencies in the EEG signals, while the convolutional layers focus on extracting local spatial relationships. This integration enables GlepNet to harness multi-scale contextual and positional information effectively.

The architecture further employs feedforward layers to refine features between encoders, optimizing the preparation of inputs and outputs for subsequent processing stages. The fusion of attention mechanisms and convolutional layers within this encoder-decoder framework allows GlepNet to construct hierarchical representations, addressing both local and global patterns essential for accurate epilepsy detection. To ensure comprehensive capturing of the EEG epilepsy feature, the encoder block is multiplying three times within the GlepNet architecture. This

Algorithm 1 GlepNet for EEG-based Epilepsy Detection

Require: EEG Input
Ensure: Epilepsy Label
Input: EEG Signal
Output: Predicted Epilepsy Label
function GLEPNET(EEG)
    $x \leftarrow \text{Linear}(EEG)$
    for $i = 1$ to 3 **do**
        $x \leftarrow \text{EncoderBlock}(x)$
    end for
    $x \leftarrow \text{Linear}(x)$
    $label \leftarrow \text{Sigmoid}(x)$
    return *label*
end function

configuration is designed to capture a broad spectrum of features across various levels of abstraction, employing a serialization of encoder blocks as a multi-scale feature extraction strategy that is effective for EEG epilepsy detection.

To provide an intuitive demonstration of GlepNet and its components, the pseudocode for the following elements is presented below in the methodology section: GlepNet (see Algorithm 1), Encoder Block (see Algorithm 2), Position-wise Feedforward Module (see Algorithm 3), Multi-Head Self-Attention Module (see Algorithm 4) and Convolution Module (see Algorithm 5).

4.4.1.2 Encoder Blocks

Inspired by Macaroon-Net [206], the encoder blocks of GlepNet adopt a sandwich structure, which divides the feedforward layer in the Transformer encoder into two half-step residual feedforward units. A multi-head self-attention and convolution modules are included between the two feedforward modules. Mathematically, for given input $(f^t_{(e-1)}|t = 1, ..., T)$ to the e^{th} encoder block, the output $(f^t_{(e)}|t = 1, ..., T)$ of the block is formulated as

$$\begin{aligned} f^{(e)}_{FF_1} &= (f^t_{(e-1)}|t = 1, ..., T) + \\ &\frac{1}{2}\,\text{PFFM}((f^t_{(e-1)}|t = 1, ..., T)) \end{aligned} \tag{4.18}$$

$$f^{(e)}_{MHSA} = f^{(e)}_{PFFM_1} + \text{MHSA}(f^{(e)}_{PFFM_1}) \tag{4.19}$$

Algorithm 2 GlepNet - Encoder Block

function EncoderBlock(x)
  $x_{residual} \leftarrow x$
  $x \leftarrow$ LayerNorm(x)
  $x \leftarrow$ MHSA-Module(x) + $\frac{1}{2}x_{residual}$
  $x_{residual} \leftarrow x$
  $x \leftarrow$ LayerNorm(x)
  $x \leftarrow$ Conv-Module(x)
  $x_{residual} \leftarrow x$
  $x \leftarrow$ LayerNorm(x)
  $x \leftarrow$ PFFM(x) + $\frac{1}{2}x_{residual}$
  return x
end function

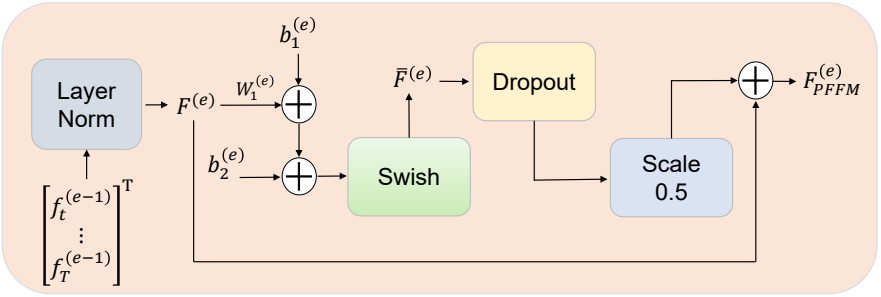

FIGURE 4.6: Position-wise feed-forward model. Two linear layers increase and decrease the dimensionality of the data, respectively.

$$f_{Conv}^{(e)} = f_{MHSA}^{(e)} + \text{Conv}(f_{MHSA}^{(e)}) \tag{4.20}$$

$$\begin{gathered}(f_{(e)}^{t} | t = 1, ..., T) = \\ \text{LayerNorm}(f_{Conv}^{(e)} + \frac{1}{2}\,\text{PFFM}(f_{Conv}^{(e)}))\end{gathered} \tag{4.21}$$

where PFFM($\cdot$) denotes the position-wise feed-forward module, MHSA($\cdot$) signifies the multi-head self-attention module, and Conv($\cdot$) represents the convolution module.

4.4.1.3 Position-wise feed-forward module

The GlepNet architecture employs a feed-forward module before and after the MHSA module, consisting of two linear layers and activation as shown in FIGURE. 4.6. The e^{th} PFFM module transforms a sequence of input vectors $(f_{(e-1)}^{t} | t = 1, ..., T)$ from the previous encoder as follows:

Algorithm 3 GlepNet - Positionwise Feed-Forward Module (PFFM)

function PFFM(x)
    $x \leftarrow$ LayerNorm(x)
    $x \leftarrow$ Linear(x)
    $x \leftarrow$ Switch(x)
    $x \leftarrow$ Dropout(x)
    $x \leftarrow$ Linear(x)
    $x \leftarrow$ Dropout(x)
    return x
end function

$$F^{(e)} = \text{LayerNorm}([f_t^{(e-1)} \cdots f_T^{(e-1)}]) \in \mathbb{R}^{T \times D} \tag{4.22}$$

$$\begin{aligned} \bar{F}^{(e)} = \text{Swish}(F^{(e)} W_1^{(e)} + b_1^{(e)^\top}) W_2^{(e)} + \\ b_2^{(e)^\top} \in \mathbb{R}^{T \times D} \end{aligned} \tag{4.23}$$

where T denoted the time length and D is the feature dimension. $W_1^{(e)} \in \mathbb{R}^{D \times d_{PFFM}}$ and $b_1^{(e)} \in \mathbb{R}^{d_{PFFM}}$ are the projection matrix and bias of the first linear layer, $1 \in \mathbb{R}^T$ is an all-one vector. d_{PFFM} is the dimension of hidden units. We also apply Swish activation [208] Swish($\cdot$) and dropout [209] to help regularize the module. $W_2^{(e)} \in \mathbb{R}^{D \times d_{PFFM}}$ and $b_2^{(e)} \in \mathbb{R}^{d_{PFFM}}$ are the second linear projection matrix and bias. Following the pre-norm residual units [210], the final output of the PFFM module is computed as follows:

$$F_{PFFM}^{(e)} = F^{(e)} + \frac{\text{Dropout}(\bar{F}^{(e)})}{2} \in \mathbb{R}^{T \times D} \tag{4.24}$$

4.4.1.4 Multi-head self-attention module

The structure of the multi-head self-attention module is shown in FIGURE. 4.7. The MHSA module in the e^{th} encoder processes features from the PFFM module. The input features $F_{PFFM}^{(e)}$ is converted by layer normalization

$$\bar{F}^{(e)} = \text{LayerNorm}(F_{PFFM}^{(e)}) \in \mathbb{R}^{T \times D} \tag{4.25}$$

Algorithm 4 GlepNet - Multi-Head Self-Attention Module (MHSA)

function MHSA-MODULE(x)
    $x \leftarrow \text{LayerNorm}(x)$
    $x \leftarrow \text{MultiHeadSelfAttention}(x)$
    $x \leftarrow \text{Dropout}(x)$
    return x
end function

Then, in the multi-head self-attention block, each attention head computes a pair-wise similarity matrix $S_h^{(e)}$ using the dot products of query vectors $\bar{F}^{(e)}Q_h^{(e)} \in \mathbb{R}^{T\times d}$ and key vectors $\bar{F}^{(e)}K_h^{(e)} \in \mathbb{R}^{T\times d}$

$$S_h^{(e)} = \bar{F}^{(e)}Q_h^{(e)}(\bar{F}^{(e)}K_h^{(e)})^\top \in \mathbb{R}^{T\times T}(1 \le h \le H) \tag{4.26}$$

where H denoted the number of heads. $Q_h^{(e)}$,$K_h^{(e)} \in \mathbb{R}^{D\times d}$ are the h^{th} head's query and key matrices of projection. Scaled by $1/\sqrt{D/H}$, $S_h^{(e)}$, denotes the matrix of similarity, together with softmax to produce the attention weight matrix

$$A_h^{(e)} = \text{Softmax}\left(\frac{S_h^{(e)}}{\sqrt{D/H}}\right) \in \mathbb{R}^{T\times T} \tag{4.27}$$

Then the attention weight matrix $A_h^{(e)}$ is used to compute context vectors $C_h^{(e)}$ with the value vectors $\bar{F}^{(e)}V_h^{(e)} \in \mathbb{R}^{T\times d}$:

$$C_h^{(e)} = A_h^{(e)}(\bar{F}^{(e)}V_h^{(e)}) \in \mathbb{R}^{T\times d} \tag{4.28}$$

where $V_h^{(e)} \in \mathbb{R}^{D\times d}$ is the matrix of value projection. The final output feature of the multi-head self-attention module is computed by the concatenation of all heads' context vectors and an output projection matrix $O^{(e)} \in \mathbb{R}^{D\times D}$:

$$\bar{F}_{MHSA}^{(e)} = [C_1^{(e)} \cdots C_H^{(e)}]O^{(e)} \in \mathbb{R}^{T\times D} \tag{4.29}$$

$$\begin{aligned} F_{MHSA}^{(e)} = \text{LayerNorm}(\bar{F}^{(e)} + \\ \text{DropOut}(\bar{F}_{MHSA}^{(e)})) \in \mathbb{R}^{T\times D} \end{aligned} \tag{4.30}$$

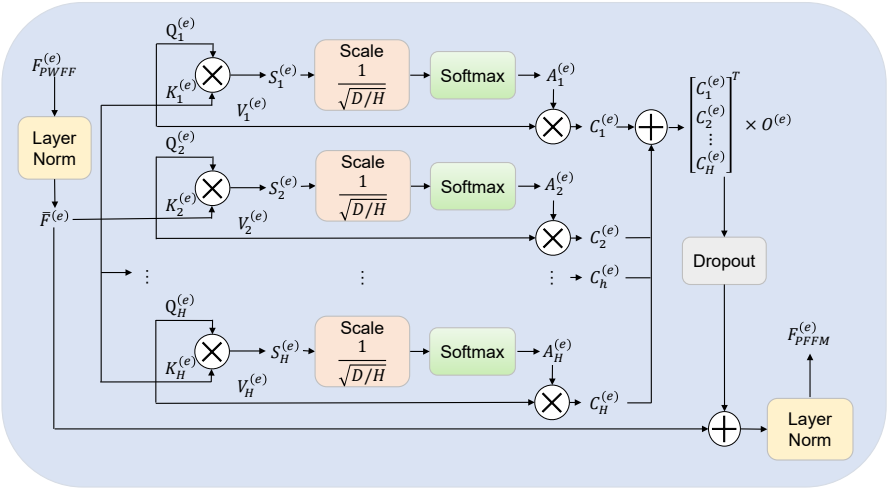

FIGURE 4.7: Multi-head self-attention module. We apply a multi-head self-attention without the relative positional embedding in this residual unit.

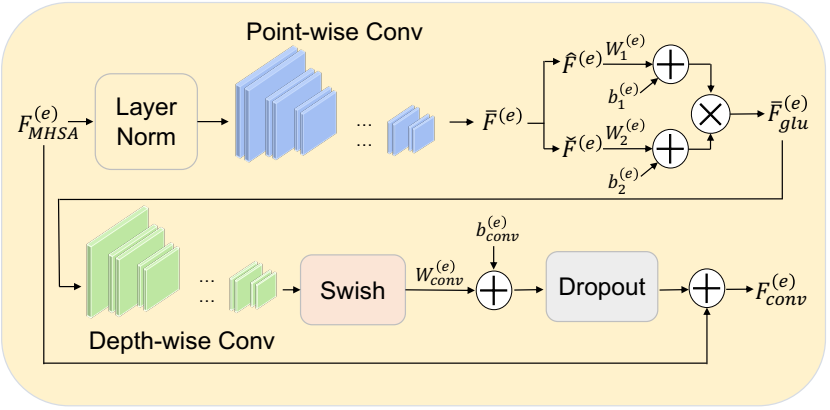

FIGURE 4.8: Convolution module. The convolution module contains two convolutional layers of different scales, with Gated linear units used as the activation layer in the middle. The Swish activation function is used, followed by a linear layer.

4.4.1.5 Convolution module

Similar to [211], the convolution module, as shown in FIGURE 4.8, takes $F^{(e)}_{MHSA}$ as input and starts with a point-wise convolution and a gated linear unit (GLU) [212]:

$$\bar{F}^{(e)} = \text{PWConv}(\text{LayerNorm}(F^{(e)}_{MHSA})) \in \mathbb{R}^{T\times 2D} \tag{4.31}$$

$$\bar{F}^{(e)}_{glu} = (\hat{F}^{(e)}W^{(e)}_1 + b^{(e)}_1) \otimes \sigma(\check{F}^{(e)}W^{(e)}_2 + b^{(e)}_2) \in \mathbb{R}^{T\times D} \tag{4.32}$$

Algorithm 5 GlepNet - Convolution Module

function CONV-MODULE(x)
    $x \leftarrow \text{LayerNorm}(x)$
    $x \leftarrow \text{PointwiseConv}(x)$
    $x \leftarrow \text{GatedLinearUnit}(x)$
    $x \leftarrow \text{DepthwiseConv}(x)$
    $x \leftarrow \text{BatchNorm}(x)$
    $x \leftarrow \text{Switch}(x)$
    $x \leftarrow \text{Linear}(x)$
    $x \leftarrow \text{Dropout}(x)$
    return x
end function

where PWConv$(\cdot)$ is a one-dimensional point-wise convolutional layer with kernel size (1×1) and stride of 1. The output $\bar{F}^{(e)} \in \mathbb{R}^{T \times 2D}$ could be divided into $\hat{F}^{(e)} \in \mathbb{R}^{T \times D}$ and $\check{F}^{(e)} \in \mathbb{R}^{T \times D}$, which are the first half and the second half of the output, respectively. $W_1^{(e)} \in \mathbb{R}^{D \times D}$, $b_1^{(e)} \in \mathbb{R}^D$,$W_2^{(e)} \in \mathbb{R}^{D \times D}$, $b_1^{(e)} \in \mathbb{R}^D$ are learned parameters, σ is the sigmoid function and $\otimes$ is the element-wise product. The output of GLU $\bar{F}_{glu}^{(e)}$ is processed with a DWConv$(\cdot)$:

$$\bar{F}_{DWConv}^{(e)} = W_{conv}^{(e)} \,\text{Swish}(\text{DWConv}(\bar{F}_{glu}^{(e)})) + b_{Conv}^{(e)} \in \mathbb{R}^{T \times D} \tag{4.33}$$

where $W_{Conv}^{(e)} \in \mathbb{R}^{d_{pwconv} \times D}$, $b_{Conv}^{(e)} \in \mathbb{R}^D$ are learned linear parameters of Convolution module. Swish$(\cdot)$ is the activation function. DWConv$(\cdot)$ is a one-dimensional depth-wise convolutional layer with a kernel size of 15. d_{pwconv} is the dimension of depth-wise convolution layer output. The final output features $F_{Conv}^{(e)}$ of the convolution module are as follows:

$$F_{Conv}^{(e)} = F_{MHSA}^{(e)} + \text{Dropout}(\bar{F}_{DWConv}^{(e)}) \in \mathbb{R}^{T \times D}. \tag{4.34}$$

4.4.1.6 Linear Layer, LayerNorm and Sigmoid

Linear layers, layer normalization, and the sigmoid activation function are key components in constructing deep neural networks. In GlepNet, the two linear layers are applied to help the network map the input to an intermediate valuable representation for the task. Layer normalization is a technique that normalizes the activations across the features to stabilize network training. By subtracting

the mean and dividing by the standard deviation across the neurons, the layer norm helps the network generalize better. Finally, the sigmoid activation takes a real-valued input and squashes it to a probability value between 0 and 1, which represents outputs as probabilities and allows the network to predict the epilepsy class.

4.4.1.7 Grad-CAM for Neural Epilepsy Detection

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique that can be used to visualize and understand which regions in medical images are most important for neural network-based models in making predictions. Recently, Grad-CAM has been applied in research for detecting epilepsy from EEG and MRI scans. By highlighting the critical regions that contain discriminative information in the EEG data, Grad-CAM provides insight into which part of neural signals the models rely on to identify epilepsy.

For example, studies have used Grad-CAM on convolutional neural networks analyzing EEG data to indicate the key time segments and channels that contain epileptiform discharges and brain oscillations associated with seizures [216]. Other work has applied Grad-CAM on MRI analysis models to discover the importance of temporal and frontal lobe abnormalities [217]. Incorporating Grad-CAM interpretation enables better trust and transparency around automated neural epilepsy detection, as doctors can visually see what drives the AI model's outputs. Going forward, Grad-CAM could be used to improve and refine neural networks to focus learning on clinically relevant biomarkers and spatial patterns for robust epilepsy risk stratification and care.

The Grad-CAM method in this chapter is utilized for single-channel EEG signals for binary classification task. The basic overview is to input data into the proposed model as forward propagation and obtain raw scores before sending them into fully connected Layers. Then the back propagation is applied to obtain the weights. Eventually, we multiply the weights with each feature score, respectively. Through a ReLU activation function, we obtain the Grad-CAM values and visualize them to obtain the heatmaps. The basic step can be elaborated as Eq. (4.35), where parameter W represents output feature values after the last convolutional layer, t denotes types (0 or 1), r means the rth channel of a specific feature layer (here we

set r as time sampling number). W^r symbolizes rth channel of last feature layer W and w_r^t represents weight of rth channel in categories t. Finally, G_t^r means the Grad-CAM values of rth channel in t classification and is given by

$$G_t^r \leftarrow \max(\sum_r w_r^t W^r, 0). \tag{4.35}$$

The total heatdata is a matrix of $1 \times r$ combining with r channel of Grad-CAM values of feature layer W, which is used to draw heatmaps with the same sampling points as that of raw input data.

4.4.2 Experiments and Discussions

4.4.2.1 Datasets

The experiments were conducted using four single-channel EEG datasets. These datasets encompass: 1) The Epileptic Seizure Recognition dataset (ESR) [218]. 2) EEG Epilepsy Datasets sourced from the Neurology & Sleep Centre (NSC) in New Delhi [219]. 3) Bonn EEG Datasets [59]. 4) Single electrode EEG data collected from healthy and epileptic patients [220].

1) Epileptic Seizure Recognition Dataset (ESR)

Each EEG data spans 23.6 seconds. For our study, we consolidated the original 5 labels into 2 categories. Specifically, we grouped four out of the five labels that were unrelated to epileptic activity into a singular category. We adopted the data pre-processing methodology outlined in [213].

2) EEG Epilepsy Datasets from Neurology & Sleep Centre, New Delhi (NSC)

The dataset originates from the Neurology & Sleep Centre (NSC) in Hauz Khas, New Delhi. It encompasses EEG time series segments from 10 patients with epilepsy. Each dataset file contains 1024 sample points, spanning 5.12 seconds, with a sampling rate of 200 Hz. The dataset is categorized into pre-ictal, interictal, and ictal three categories. For our study, we streamlined the original three labels into two by amalgamating the interictal and ictal categories into a singular label.

3) Bonn EEG dataset (Bonn)

This dataset is from Universitätsklinikum Bonn in Germany and contains 100 single-channel EEG time series. Each datum consists of 4097 sample points recorded over 23.6 seconds (sampling rate: 173.6 Hz). The data are categorized into five groups: Z, O (healthy subjects), and N, F, S (epileptic subjects). For our experiment, we focus only on categories N, F, and S, ignoring Z and O. Since categories N and F are collected from epileptic intervals, and S is collected from the epileptic phase, we combine N and F into a single label. Thus, the dataset is simplified into a dichotomous dataset.

4) Single electrode EEG data of healthy and epileptic patients (SEHE)

This dataset is produced in a similar method of Bonn EEG dataset that mentioned before, for instance, the sampling rate, time period of each file are the same. However, this dataset has only 2 labels ("H" for healthy subjects and "E" for epilepsy subjects). It also overcomes many limitation of Bonn EEG dataset like inconsistent sensor position during acquisition [221]. All data are collected from 15 healthy and epileptic people with surface EEG electrodes. And we pre-process this dataset in a same method as that of Bonn EEG dataset after considering the similarity between them.

4.4.2.2 Baseline Model and Configuration

The proposed model is compared with following baseline model, including SVM [222], ConvNet [196], TS-SENet [223], Dense-CNN [214], CNN-LSTM [215], Transformer.

1) SVM

Empirical mode decomposition (EMD) has been applied efficiently as a prevalent temporal and spectral feature extractor technique in analysing time series. A support vector machine (SVM) is frequently incorporated as the classifier and makes the prediction. This method is a foundational baseline for accurately identifying and analyzing epileptic EEGs.

2) ConvNet

The ConvNet comprises four convolution-maxpooling blocks. The initial block is uniquely structured to manage input efficiently, followed by three standard convolution-maxpooling blocks, culminating in a dense softmax layer. In the primary layer, filters carry out convolutions across time. Each filter executes spatial filtering in the subsequent layer, using weights for every potential electrode pair based on the filters from the preceding temporal convolution.

3) TS-SENet

TS-SENet is a comprehensive end-to-end framework designed for EEG seizure detection. At its outset, TS-SENet integrates both multi-level spectral and multi-scale temporal analyses concurrently. This process results in capturing hierarchical multi-domain representations using a modified squeeze-and-excitation block. To conclude, a classification network recognizes epileptic EEG patterns based on the features harvested from the earlier subnetworks.

5) Dense-CNN

The Dense-CNN architecture consists of multiple convolutional layers, whereby each layer is equipped with a distinct set of convolutional filters. The first component consists of a linear layer. The subsequent layers employ an architectural framework known as Dense-Inception, including several Inception modules. The Inception module is composed of numerous branches that concurrently process the input by employing distinct convolutional filters with different kernel sizes. After that, the results obtained from each branch are combined by concatenating them along the channel dimension. This combined output is then sent through a convolutional layer in order to decrease the number of channels. The output tensor produced by one Inception module is then given as input to the subsequent Inception module, and this sequence of operations is iterated until reaching the final layer.

6) Transformer

The network architecture is a classifier based on the Transformer model. It comprises an upsampling layer, two linear layers, Transformer-Encoder layers, a fully connected layer, and a sigmoid activation function. The upsampling layer transforms the input sequence into a higher hidden state. The first linear layer is responsible for mapping the hidden state into a single value, which

subsequently serves for scaling the output of the Transformer-Encoder layer. The Transformer-Encoder layer is comprised of a self-attention mechanism, as well as two linear layers and dropout layers. The self-attention mechanism employs eight attention heads and a projection matrix for its output. The dropout are applied to the attention and linear layers. The output of the Transformer-Encoder is mapped to one value by the fully connected layer. The sigmoid activation function further processes this value to get the final classification result.

7) CNN-LSTM

CNN-LSTM [215] consists of a convolutional neural network and a long and short-term memory (LSTM) layer. The convolutional neural network comprises of two convolutional layers and a maximum pooling layer. The fully connected layer yields an output size of 512. The input to the LSTM layer consists of the output from the fully connected layer. It has two LSTM layers, each having a hidden state size of 128.

4.4.2.3 Training Paradigm

The k-fold cross-validation paradigm is applied to divide the training and testing sets. For each dataset, the data is divided into 5 folds, where one fold is the testing set, and the remaining are used to train. The data used for training is divided into 80% and 20%, respectively, where the 20% is the validation set, and the remaining is the training set.

4.4.2.4 Cropping Strategy

To expand the data volume, our experiments used a cropped strategy [196], which crops the data of each trail into $0.5s$ non-overlapping segments. The cropping is done strictly after splitting the training, validation, and testing sets, which prevents data leakage. We make all baselines follow the same cropping strategy to prevent abnormally high accuracy and F1 scores.

4.4.2.5 Hyperparameters and Environments

The proposed GlepNet architecture is composed of three Encoder blocks, with each Encoder layer consisting of a self-attention module and a convolution module. The attention mechanism has a total of 8 attention heads, with each head having a dimensionality of 64. The convolution module employs one-dimensional convolution layers and the gated linear unit (GLU) activation function. The convolution kernel has a size of 15, a step size of 1, a padding of 7, and is divided into 512 groups. The model is comprised of two feed-forward layers and incorporates residual connections. It concludes with the application of two linear layers and a sigmoid activation function to categorise the extracted features.

The model is trained on an NVIDIA Tesla A100. The batch size is set to 128, and we use learning rate decay by setting an initial learning rate of $1e-4$ and a decay factor of 0.1. The minimum learning rate is set as $1e-8$. Additionally, The total training epoch is default 500 and the early stopping strategy was implemented. We retain the best model obtained to do the model inference, where we set the batch size the same as training process.

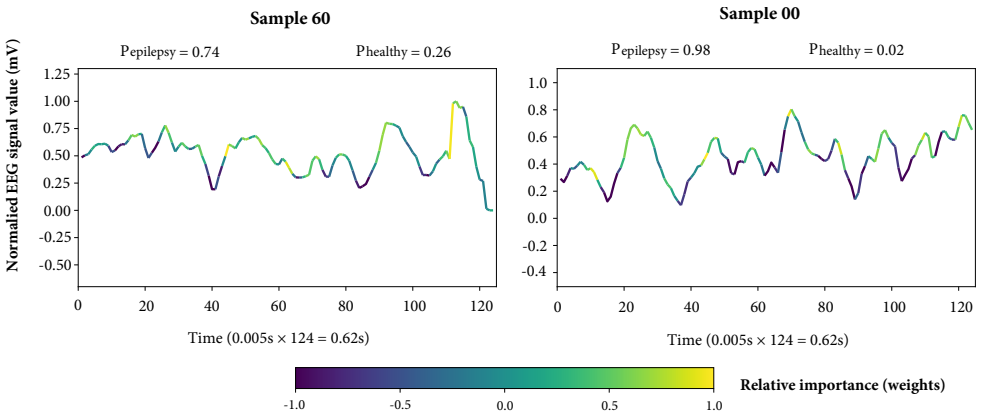

FIGURE 4.9: Interpretable Visualization. For those samples have a label "Epilepsy", GlepNet model focuses more on the part of the value that has a large gradient and changes drastically over time. As seen in sample 60 (left), our GlepNet model exhibits the highest degree of attention towards the segments located approximately at 25, 45-60, 90-95, and 110-120, respectively. Notably, these segments have comparatively greater amplitudes in comparison to other regions. As shown in sample 00 (right), it is revealed by the Grad-CAM method that the peaks of the signal are lighter that the other area.

TABLE 4.2: Comparison (in %) of Accuracy (ACC) of our proposed GlepNet and six Baselines on the Epileptic Seizure Recognition (ESR), EEG Epilepsy Datasets from Neurology & Sleep Centre in New Delhi (NSC), Bonn EEG dataset (Bonn), and Single electrode EEG data of healthy and epileptic patient (SEHE). The best and highest results are in **bold**.

Method	ESR (ACC)	NSC (ACC)	Bonn (ACC)	SEHE (ACC)
SVM [222]	75.23±7.13	72.81±8.22	75.25±7.22	68.82±6.92
ConvNet [196]	86.10±4.52	78.88±6.01	80.94±6.22	74.88±7.16
TS-SENet [223]	89.72±4.57	86.02±5.01	92.80±4.25	87.28±3.56
DenseCNN [214]	91.13±4.55	85.70±4.98	88.54±6.22	95.47±2.92
CNN-LSTM [215]	92.82±4.92	79.91±5.43	94.74±3.17	80.71±6.53
Transformer	92.30±3.73	71.51±8.68	85.31±6.55	65.61±8.01
GlepNet (ours)	**96.91±1.30**	**93.30±2.48**	**97.02±2.22**	**98.17±1.79**

TABLE 4.3: Comparison (in %) of F1 Score (F1) of our proposed GlepNet and six Baselines on the Epileptic Seizure Recognition (ESR), EEG Epilepsy Datasets from Neurology & Sleep Centre in New Delhi (NSC), Bonn EEG dataset (Bonn), and Single electrode EEG data of healthy and epileptic patient (SEHE). The best and highest results are in **bold**.

Method	ESR (F1)	NSC (F1)	Bonn (F1)	SEHE (F1)
SVM [222]	77.21±7.42	74.13±9.33	81.81±5.12	67.61±6.22
ConvNet [196]	89.34±4.98	80.15±3.50	87.01±4.28	70.92±4.82
TS-SENet [223]	84.63±6.47	84.55±5.42	85.88±6.07	85.32±4.22
DenseCNN [214]	95.31±3.01	80.01±5.61	85.17±6.62	95.11±3.66
CNN-LSTM [215]	93.14±5.02	69.32±8.02	93.62±2.62	82.73±4.72
Transformer	92.33±3.69	64.03±8.12	77.12±6.77	62.82±8.42
GlepNet (ours)	**98.12±0.80**	**90.63±1.52**	**97.24±2.01**	**99.11±0.52**

4.4.3 Performance Comparison

4.4.3.1 Baseline comparison

In this section, we evaluate the performance and robustness of our proposed GlepNet alongside six baseline models across four datasets. The results are presented in Table 4.2 and Table 4.3. Overall, GlepNet consistently outperforms the baseline models on these datasets, demonstrating significant advantages in terms of accuracy and robustness.

To be specific, while SVMs are effective for general classification tasks, they struggle with the sequential nature of EEG data without extensive kernel modifications and feature engineering. Dense-CNN and ConvNet, which primarily focus on spatial feature extraction, do not perform optimally with time series data due to their limitations in handling sequential information effectively.

The Transformer, known for its proficiency in global modeling of time-domain data, excels in capturing long-range dependencies, which are crucial for identifying complex epileptic patterns such as spikes and the interleaving of spikes with slow waves. However, despite its strengths, the Transformer is computationally intensive and can suffer from inefficiencies as well as overfitting when scaling to larger datasets or in cases where local context is vital.

CNN-LSTM improves upon these models by effectively capturing the temporal dependencies inherent in time-series data. Nevertheless, it is somewhat limited in integrating these features with global context. TS-SENet, which integrates squeeze-and-excitation (SE) blocks with convolutional networks, aims to enhance the network's sensitivity to temporal dynamics by adaptively recalibrating channel-wise feature responses. Despite these advancements, TS-SENet often faces challenges in effectively scaling these features across more complex datasets, potentially limiting its practical utility in real-world EEG analysis where variability across data is high.

In comparison, GlepNet demonstrates the highest accuracy and robustness by combining the strengths of attention mechanisms and convolutional features. Through its convolution module, GlepNet detects local signal characteristics, while its attention mechanism facilitates the learning of overall time dependencies in epilepsy-related EEG representations. This comprehensive approach allows GlepNet to outperform other models, leveraging detailed local insights alongside broader temporal patterns to enhance EEG epilepsy detection.

4.4.3.2 Ablation Study

In this part, we conduct two ablation studies in terms of discussing the influence of the convolution module, multi-head self-attention module as well as the number of the encoder blocks.

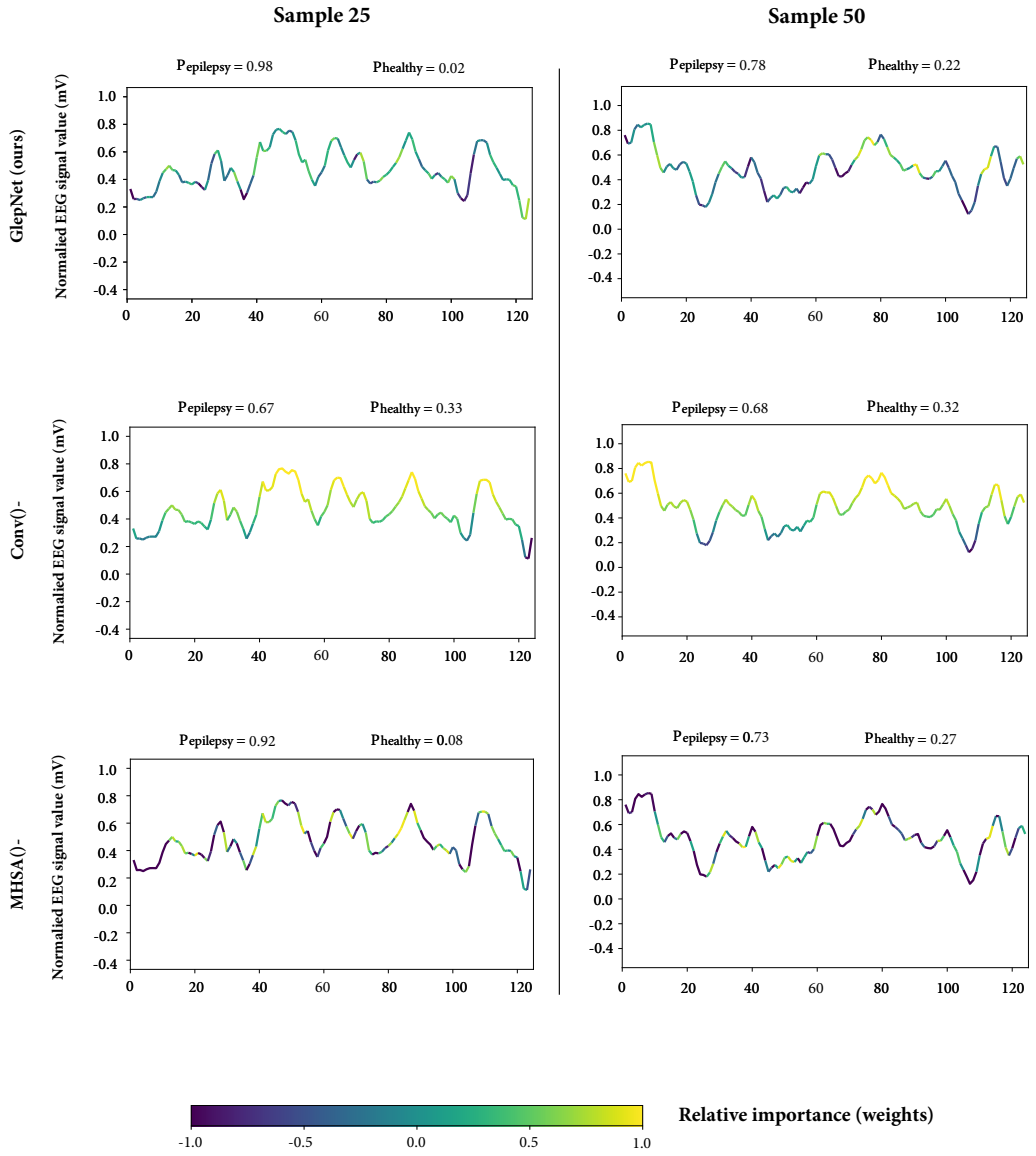

FIGURE 4.10: Comparison of the attention of the GlepNet model with convolutional layers, the GlepNet model with Convolution Module removed and the GlepNet model with Multi-Head Self-Attention Module removed for two different samples (index 25 on the left and 50 on the right) visualized with the Grad-CAM method. The RELATIVE IMPORTANCE can be represented by different colors, with brighter colors denoting more pronounced attention of the model: the two images on the first row are generated by our proposed GlepNet model, where the model pays attention to a proper number of localities and global information; the images on the second row are produced by the GlepNet model with the Convolution Modules removed, where the model pays too much attention to global information and ignores local characteristics; the images on the third row represent the attention of GlepNet model without Multi-Head Self-Attention Modules, where the model pays too much local attention but overlooks the global features. The classifying effect can be indicated by probability, and it is evident that the first row images - which has a higher probability of epilepsy - have better effect of classifying than that of the others.

Convolution Module & Multi-head Self-attention Module: We conducted an ablation study to assess the impact of the convolution module and the multi-head self-attention module on the network's performance. This study involved comparing the test accuracy of our proposed model, GlepNet, against two modified versions: one where the convolutional layers were removed (Conv −) and another where the multi-head self-attention modules were eliminated (MHSA −). The results are presented in Table 4.4. From the table, it is evident that removing the convolutional modules resulted in a reduction in accuracy as well as robustness across all datasets, with the most pronounced reduction observed in the NSC dataset. On the other hand, eliminating the multi-head self-attention modules led to a slight performance improvement in the ESR and Bonn datasets but a noticeable decline in the NSC and SEHE datasets.

The complete GlepNet model, incorporating both the convolution and multi-head self-attention modules, consistently outperformed the other versions across all datasets. The GlepNet model's superiority underscores both modules' complementary roles in enhancing the network's predictive capacity. Thus, integrating convolutional layers with multi-head self-attention mechanisms provides a robust architecture that effectively leverages spatial and temporal features, leading to improved performance in epilepsy dataset classification.

Moreover, we make a interpretable visualization as shown in FIGURE 4.9 and an ablation visualization aims to elucidate the relative importance and areas of attention the model allocates when diagnosing the presence of epilepsy as shown in FIGURE 4.10 using the Grad-CAM method.

Three distinct configurations of the GlepNet model were examined: the complete GlepNet model, the model with the convolutional layers removed (Conv −), and the model with the Multi-Head Self-Attention Module eliminated (MHSA −). The color-coded weights represent the relative importance, with brighter colors signifying areas where the model pays heightened attention.

The visualizations show that the complete GlepNet model demonstrates a nuanced allocation of attention across the time series, as evidenced by the balanced distribution of weights. This balanced approach allows the model to capture both global and localized features effectively, leading to a higher probability of accurately classifying the presence of epilepsy (as denoted by the probabilities in each

TABLE 4.4: Ablation study (Accuracy in %) on four Epilepsy Datasets: As shown, "Conv −" means removing the convolution modules from the proposed GlepNet model, whereas "MHSA −" means eliminating Multi-Head Self-Attention modules from GlepNet. The best and highest results are in **bold**.

Method	ESR	NSC	Bonn	SEHE
Conv −	93.18±3.63	68.83±4.52	83.18±5.01	67.28±4.27
MHSA −	92.28±4.83	89.63±2.81	92.77±3.35	94.02±3.51
GlepNet (ours)	**96.91±1.30**	**93.30±2.48**	**97.02±2.22**	**98.17±1.79**

chart). Conversely, the Conv − configuration appears to spread its attention more broadly but lacks depth in certain critical regions. This potentially results in a decreased capacity to discern nuanced features. Similarly, the MHSA − configuration focuses on particular time segments but misses broader patterns. This comparative analysis underscores the complementarity of convolutional layers and multi-head self-attention mechanisms in the GlepNet model. Fusing both components allows the model to achieve a fine balance between localized detail and holistic understanding, optimizing its diagnostic accuracy for epilepsy.

Number of Encoder Blocks: The number of encoder blocks in GlepNet was adjusted to find the optimal performance. The results for one fold are illustrated in FIGURE 4.11. We observe that the best performance across the four datasets is achieved with 3 encoder blocks. There is a clear trend where both accuracy and F1 score generally peak at 3 encoder blocks before declining with additional blocks. Specifically, for the ESR dataset, accuracy and F1 score increase up to three encoder blocks, reaching nearly 100% and 95%, respectively. Beyond this point, further increases in the number of encoder blocks lead to diminishing returns. This pattern is consistent across the Bonn and SEHE datasets, with optimal performance similarly occurring at 3 encoder blocks. For the Bonn dataset, both accuracy and F1 score peak just below 100%. In the SEHE dataset, the peak values are slightly lower. The NSC dataset also shows a similar pattern. These results suggest that three encoder blocks are optimal, maximizing accuracy and F1 score across most datasets. Using fewer or more encoder blocks compromises performance.

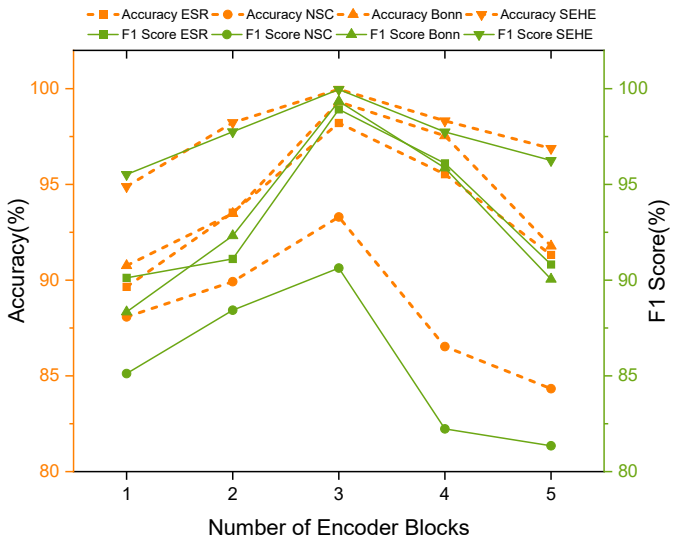

FIGURE 4.11: Comparative analysis of the impact of varying encoder blocks on accuracy and F1 score across four datasets under one fold. Optimal performance is generally observed with 3 encoder blocks, with deviations in accuracy and F1 score emerging beyond this point.

4.4.4 Interpretability Analysis

In this section, we consider investigating and analysing the features learned by our model. Inspired by [224], we decided to conduct our experiment on the dataset of EEG Epilepsy Datasets from NSC via Gradient-weighted Class Activation Mapping (Grad-CAM) [124] to visualize and analyse the patterns acquired by the model for interpretation. To interpret the common patterns acquired by the model, we analysed the results and displayed representative samples from a variety of subjects.

To determine the increased efficacy of the model resulting from the addition of convolution layers, we examine the attention of select test samples on Transformer (Multi-Head Self-Attention Modules only), Convolution Neural Network (Convolution Modules only) and our proposed GlepNet model (Multi-Head Self-Attention Modules + Convolution Modules). Since EEG signals contain potent one-dimensional structural information and are closely linked with local attributes [225], the visualization analysis effectively showcases the benefits of each type of added layers. To capture long-range dependencies, the Transformer employs the self-attention mechanism to model global contextual knowledge while disregarding some local detailed features [207]. As shown in the FIGURE 4.10, it is evident that the model utilizing only the Transformer has a wider and less focused attention

in comparison to GlepNet, and allocates less attention to some local detailed features. In extreme contrast, for the Convolution Neural Network, the model pays more attention to detailed localities, but ignores latent interrelated features, and the bright regions in the Grad-CAM attention representation map are intermittent and irregular, and the captured features are more difficult to be learnt by the model [207]. Conversely, our model pays attention not only to global features properly but also to local detailed features.

We also consider investigating shared EEG features across the various samples, which are learned by our proposed model as epilepsy or healthy category indicators. The presence of significant waveform jitter in EEG time-domain data suggests that these specific locations might serve as valuable biomarkers for the identification of epileptogenic areas in individuals with epilepsy. They are dense areas of peaks and troughs in EEG, usually recorded by intracranial electrodes [183]. They are continuous oscillatory waves in the time domain after filtering, with uniform waveforms and amplitudes significantly higher than those of the background wave [226].

For those samples attached to a high likelihood of epilepsy, we have observed that our proposed models share commonalities in focusing on the region with intense waveform jitter of epileptic EEG signals. As shown in FIGURE 4.9, the parts of the signal whose values change drastically over time (more significant gradient) take more weight in calculating the probability values for the final classification. Furthermore, for the peak portion of the signal that is classified as epilepsy, our model will pay more attention to it.

4.4.5 Epilepsy Detection Future Direction

4.4.5.1 Integration of Prior Human Knowledge

A significant limitation of current EEG epilepsy detection systems is their inability to incorporate established physiological principles as prior information. In practice, these systems tend to generate correlations between features and diagnostic outcomes without discerning which features are clinically relevant. This often leads to models focusing on irrelevant or misleading features. To address this, future EEG epilepsy detection systems should be designed to leverage prior human knowledge

effectively. By integrating physiological principles directly into the learning process, these systems can ensure that the models prioritize features that are genuinely indicative of epilepsy, enhancing both the accuracy and trustworthiness of the diagnoses.

4.4.5.2 Real-Time and Personalized Monitoring

The next frontier in EEG monitoring involves developing systems capable of providing real-time, personalized insights. This means creating algorithms that not only detect epileptic events as they occur but also adapt to the unique physiological patterns of individual patients. Such systems would continuously learn from each patient's data, refining their diagnostic capabilities and potentially predicting seizures before they happen. By focusing on personalized models, we can ensure that each patient receives the most accurate and timely intervention, tailored to their specific needs.

4.5 Summary

This thesis presents GlepNet, an innovative method for neural epilepsy detection, which melds the global interdependency-capturing capabilities of the attention-based mechanism with the local feature extraction strength of the convolution blocks. The core of the GlepNet's design is an interleaved temporal convolution model together with the multi-head attention mechanism within the GlepNet's encoder blocks, allowing for comprehensive learning of the temporal relationships and global-local epilepsy-related abnormalities in EEG signals, which are critical indicators of manifestations. Incorporating the Grad-CAM algorithm further attests to GlepNet's robustness and efficacy, presenting visual interpretability of immense value to clinicians. Through transparently learning the robust global-local epilepsy representation, GlepNet not only improves epilepsy detection but also contributes significantly to machine learning in healthcare.

Chapter 5

Interpretability Guided EEG Channel Selection Framework for Driver Drowsiness Detection

5.1 Introduction

Driver drowsiness has been a leading risk factor for traffic accidents and poses a major threat to roadway safety [227]. Therefore, developing high-performance drowsiness detection systems that continuously monitor the driver's state and alert the driver at the onset of drowsiness is quite crucial for safe driving.

Considerable efforts have been made to driver drowsiness detection based on various bio-traits, including eye blinking [228], heart rate [229], respiration [230], and facial expression [231]. Among these bio-traits, the EEG, which directly measures brain activities, has gained more attention owing to its essential relationship with fatigue [232]. Current prevailing EEG-based driver drowsiness detection methods mainly utilize DNNs to learn feature representations directly from raw EEG data for binary classification (drowsiness or not). The raw EEG data is obtained by multiple electrodes positioned on specific areas of the scalp and continuously capture the brain's electrical activity. Thus the EEG data is high-dimensional and usually contains noise, artifacts, and task-irrelevant information. The existing DNNs for drowsiness detection either use single-channel EEG data [233] or full-head channel EEG data [24] as input data, resulting in limited performance due to insufficient

learning (from single-channel with incomplete information) or biased learning (from the full-head channel with noises). Since data is the basis of training DNN models, it is highly demanded to study how the different channels of EEG data affect the DNNs and how to select suitable EEG channels for effective feature learning.

Although some EEG channel selection methods exist, they are all task-specific (designed for a particular task and cannot be successfully applied to drowsiness detection) and subject-specific (the selected channels vary with each subject and may not be suitable for other subjects). To make the selected channels more general to the driver drowsiness detection task, we propose a two-stage training strategy that contains a teacher network and a student network to refine the feature learning with channel selection in a coarse-to-fine fashion. For the teacher network, we first use an interpretability method to analyze each channel's contribution to the drowsiness detection with training data and then design a voting scheme to select the top N contributing channels according to the performance of the teacher network. The selected channels of EEG data are finally fed to the student network for further training. Experiments conducted on a public dataset demonstrate that our method is highly applicable to driver drowsiness detection and can significantly improve the detection accuracy on cross-subjects.

5.2 Related Work

EEG-based Drowsiness Detection. EEG signals directly measure brain activity and reflect the fatigue state more straightforwardly than other types of signals (e.g., eye blinking, driver expression, and heart rate). In the pre-deep learning era, EEG-based drowsiness detection tasks focused more on extracting effective features. Chai *et al.* [234] utilized independent components by entropy rate bound minimization analysis (ERBM-ICA) for the source separation and autoregressive modeling for extracting the features. Gao *et al.* [8] proposed a multileveled feature generator that combines a one-dimensional binary pattern and statistical features. Schwendeman *et al.* [235] applied Welch's method to extract the power spectral density (PSD) as drowsiness features from the dry electrode in-ear EEG. Recently, deep learning has been adopted prevalently for EEG-based drowsiness detection. Nissimagoudar *et al.* [236] and Ding *et al.* [237] applied DNNs to single-channel EEG data to detect driver drowsiness. Furthermore, Cui *et al.* [24] proposed an

InterpretableCNN with two convolutional layers, one point-wise convolution layer, and one depth-wise convolution layer to process multi-channel EEG data and construct a more robust detection system. Different from these methods that rely on insufficient or noisy data, we select more informative data from raw data for model training.

Channel Selection. In previous studies, some scholars formulated the EEG channel selection into mathematical and statistical problems. With the help of the non-dominated sorting genetic algorithm (NSGA), Luis Alfredo Moctezuma *et al.* presented a channel optimation method for epileptic-seizure classification. Alyasseri *et al.* [238] proposed hybrid optimization techniques based on the binary flower pollination algorithm (FPA) and β-hill climbing (called FPA β-hc). Varsehi *et al.* [239] assumpted that there are causal interactions between the channels of EEG signals, and applied Granger Causality (GC) analysis to verify this assumption. Besides, reducing the number of channels can effectively minimize the computational complexity and training time. To this end, Yang *et al.* [240] proposed a grouped dynamic EEG channel selection method, GDCSBR, which provided flexibility for subsequent dynamic channel selection. Differently, Zhang *et al.* [241] utilized a sparse squeeze-and-excitation module, and Gaur *et al.* [242] proposed using the Pearson Correlation Coefficient (PCC), achieving automatic subject-specific selection, respectively. These methods are based on learning schemes or statistical schemes but are unsuitable for driver drowsiness detection. In contrast, we use interpretability to guide the channel selection for drowsiness detection.

5.3 Methodology

5.3.1 Framework overview

Our interpretability-guided channel selection (ICS) framework for EEG data is shown in Fig. 5.1. To select useful EEG channels, the ICS uses a two-stage training strategy. In the first training stage, the teacher network takes as input all EEG channels and is optimized for the drowsiness detection task. Then the class activation mapping (CAM) [123] is applied to each training sample to show the importance of different EEG channels, following which the channel voting is performed to select the contributing channels. In the second stage, the student

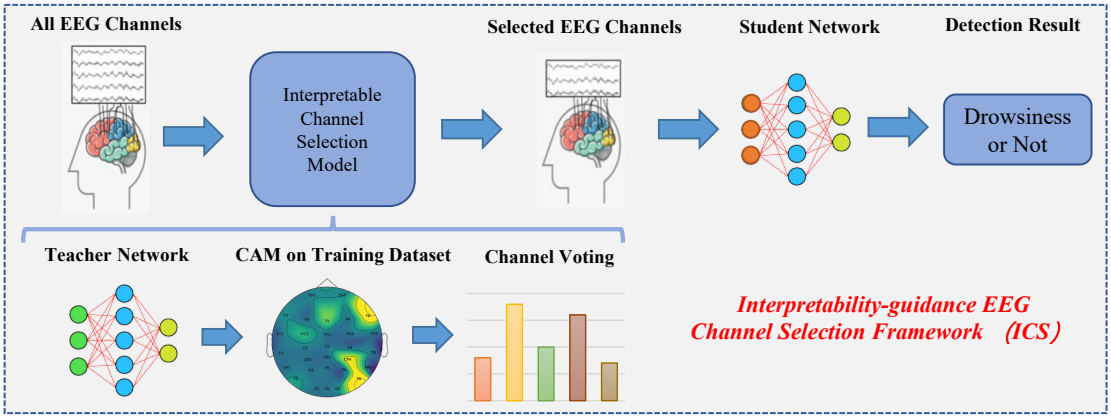

FIGURE 5.1: The structure of the proposed ICS framework.

network is trained with the selected EEG channels and is used as the final detection model. The teacher network and the student network have similar architectures, except that the convolution channels in the first layer of the student network are adjusted with the number of selected EEG channels.

5.3.2 Interpretability Guidance

Besides the two-stage networks, the interpretation technique, CAM method, plays a crucial role in the ICS framework. The CAM method is an effective interpretation technique that can pinpoint the discriminative regions of each input sample for a trained Convolutional Neural Networks (CNN) model. In particular, a heatmap is generated from the activations following the final convolutional layer for each input sample. The map is then interpolated to the size of the input sample to reveal what extent the input sample's local regions contribute to the classification.

However, the CAM method cannot be used directly for time-series EEG data and our model, because it was initially only intended for 2-D image data and deep CNN structure with a global average pooling (GAP) layer. Inspired by Cui et al.'s work [24], we modify the original CNN by changing the final dense layer to the GAP layer and applying the CAM method to visualize the EEG classification. The visualized heatmap can verify the contribution of each channel to the classification. After modification, the two-stage network architecture integrated with the CAM technique can outperform the original detection system's performance.

5.3.3 Voting Scheme

In this section, we detail how the voting scheme works. To select the top N channels that contribute significantly to the system's classification result, we apply CAM on the training set to calculate the contribution of each channel to the classification. We pick up the input sample points with high confidence levels. Specifically, the probability of the model's prediction for both categories (drowsiness or not) is obtained by the softmax layer

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad , \text{ for } i = 1, \ldots, K, \tag{5.1}$$

where $\mathbf{z}$ is the intermediate tensor inside the model before feeding into the softmax layer, e^{z_i} donates the standard exponential function for the input tensor, and K is the number of prediction classes in the teacher network ($K = 2$ in our task). Hence, we can obtain the prediction probabilities of two classes.

We consider the sample points in channels as high-contribution for the model classification when and only when the model predicts correctly and the prediction softmax probability is greater than or equal to 0.90. We conduct a prediction with the teacher model for each sample point and then apply the CAM method to the entire training dataset to visualize the EEG classification during the training process. The visualized heatmap, as shown in Fig. 5.2, demonstrates the contribution of each channel. We select the high-confidence sample points in the next step based on the aforementioned selection rule.

After obtaining the high-contribution sample points, we need to set them to make voting trace back to the top N contribution channels. To use the global information, we sum up the heatmap scores of each channel for each sample and then rank all the channels in descending order. As such, we obtain the ranking of all channels in terms of their contribution to the correct model classification. Next, we select the top N channels according to the ranking. With high-confidence samples on the training dataset and voting scheme, our approach can select the top N channels that are most closely associated. Finally, the student network is retrained using the the top N channels and thus can improve the model performance.

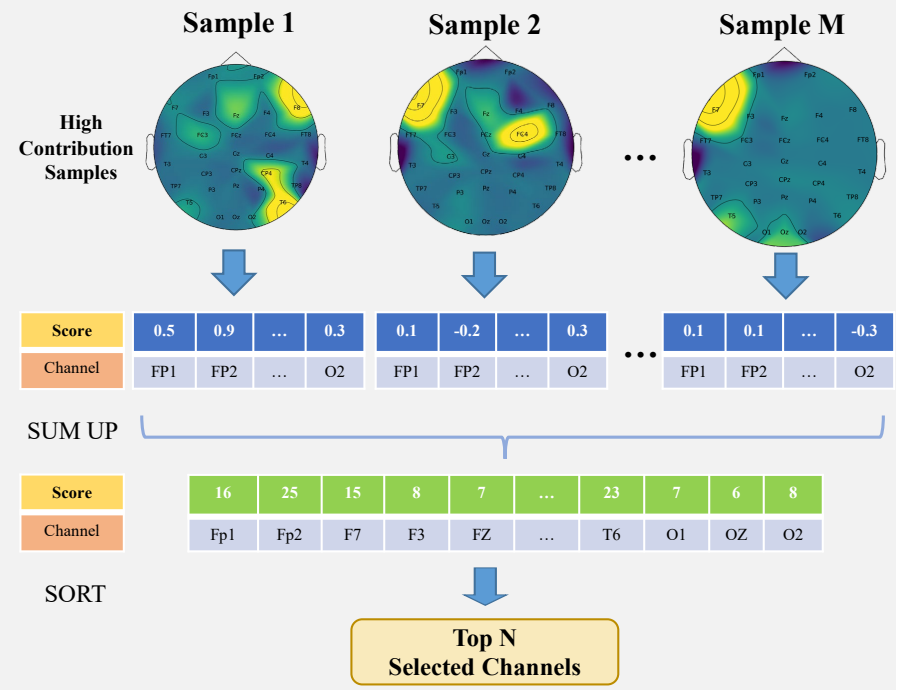

FIGURE 5.2: Visualization of CAM and voting scheme. The normalized heatmap score represents the level of contribution. The higher the score (indicated by the lighter color), the more significant the contribution. The voting can be divided into two processes, sum up and sort. Sum and sort high-contribution samples' heatmap scores, then we can obtain the channel contributions.

5.4 Experiments

5.4.1 Experimental Setup

5.4.1.1 Dataset

We perform the drowsiness detection experiments on a public EEG dataset [55]. The data was collected from 27 subjects who completed a lengthy driving task in a virtual reality simulator. The participants were required to immediately return the vehicle to the center lane in response to random lane departure events. The label (drowsiness or not) can be determined based on the subjects' reaction times. We follow the previous works [24, 243] and build a mini version of the dataset containing 11 subjects and 2952 samples, and we call it a mini driver drowsiness dataset (MDDD). Each sample, $X \in R^{30\times384}$, where 30 and 384 are the number of channels and sample points, respectively.

TABLE 5.1: Mean accuracies on MDDD with the changes in the numbers of top channels selected.

Nums of Channels	$N = 5$	$N = 10$	$N = 15$	$N = 20$	$N = 25$	$N = 30$
EEGNet 4,2 [83]	0.6742	**0.7296**	0.7283	0.7121	0.7022	0.6831
EEGNet 8,2 [83]	0.6798	**0.7369**	0.7342	0.7155	0.6977	0.6842
ShallowConvNet [196]	0.7741	**0.7944**	0.7881	0.7821	0.7725	0.7665
InterpretableCNN [24]	0.7785	**0.8138**	0.8051	0.8012	0.789	0.7826

5.4.1.2 Baselines

To demonstrate the superiority of the proposed ICS framework, we compare with four typical CNN-based methods in EEG classifications, including EEGNet [83] (both EEGNet 4,2 and EEGNet 8,2), ShallowConvNet [196] and InterpretableCNN [24]. All the baselines are implemented using their default settings.

5.4.1.3 Training and Testing

We conduct the training and testing with the Leave-one-subject-out analysis. In particular, the classifiers are trained using the EEG data from all other subjects and tested using the data from just one subject. Every subject is used as a test subject once during each iteration. The performance of drowsiness detection is measured by detection accuracy, which is the ratio of the number of correctly predicted testing samples to the number of all testing samples.

5.4.2 Ablation Study

The values of N in the ICS framework are adjusted to select the most suitable N value for the performance. The results are shown in Table 5.1. We can see that the best performance of the four classifiers is achieved with $N = 10$. When the number of selected channels is too small, the classifiers can not perform very well because they do not have enough information. As the value of N increases (from 10 to 30, with a step size of 5), the classifier's performance decreases since there are too many noises in the newly added channels, and the classifier incorrectly learns the features, hence decreasing the model performance.

TABLE 5.2: The comparison of the performance(Mean Accuracies and Standard Deviation) of MDDD between the original four classifiers and them after applying ICS for selecting the top 10 channels.

Method	**EEGNet 4,2**		**EEGNet 8,2**		**ShallowConvNet**		**InterpretableCNN**	
Sub ID	Orig. Acc	ICS Acc	Orig. Acc	ICS Acc	Orig. Acc	ICS Acc	Orig. Acc	ICS Acc
0	0.8352	**0.8455**	0.8251	**0.8312**	0.8228	**0.8422**	0.8457	**0.8617**
1	0.3125	**0.5333**	0.4525	**0.6033**	0.8018	**0.8226**	0.7121	**0.8939**
2	0.5675	**0.6021**	0.5528	**0.7021**	0.6446	**0.7612**	**0.8333**	0.8133
3	0.8521	**0.8666**	0.8024	**0.8228**	0.7026	**0.7643**	0.7770	**0.8041**
4	0.4571	**0.6051**	0.4563	**0.6017**	**0.8308**	0.8128	0.8661	**0.8929**
5	0.4776	**0.6188**	**0.5076**	0.4522	0.7912	**0.8225**	**0.8916**	0.8313
6	**0.8663**	0.8443	**0.8218**	0.8011	0.7313	**0.7757**	0.6176	**0.6373**
7	**0.7665**	0.7227	0.7435	**0.7882**	**0.7122**	0.6883	**0.7765**	0.6970
8	0.6998	**0.7127**	0.7323	**0.7943**	**0.8988**	0.8873	**0.8694**	0.8631
9	0.8781	**0.8521**	0.8328	**0.8844**	**0.8308**	0.8067	0.6759	**0.8519**
10	0.8011	**0.8225**	0.7994	**0.8476**	0.6643	**0.7548**	0.7434	**0.8053**
Mean	0.6831	**0.7296**	0.6842	**0.7390**	0.7665	**0.7944**	0.7826	**0.8138**
Std	(0.1971)	(0.1232)	(0.1575)	(0.1324)	(0.080)	(0.053)	(0.027)	(0.008)

5.4.3 Comparison with Previous Methods

We apply our ICS framework to existing driver drowsiness detection methods and show the results in Table 5.2. Note that we use $N = 10$ for channel selection for the following experiments. We can see from Table 5.2 that the proposed ICS framework provides a significant contribution to improving the previous models' performance in driver drowsiness detection. Specifically, the original EEGNet 4,2 and EEGNet 8,2 have a similar performance in terms of average accuracy, reaching around 68.3% and 68.4%. After making the channel selection with ICS, their average accuracies improved significantly, from 68.31% to 72.96% and 68.42% to 73.69%, respectively. As for ShallowConvNet and InterpretableCNN, their average accuracies increase 2.79% and 3.12%, respectively. In addition to improving the average accuracy, ICS also greatly boosts the accuracy of each subject. In terms of the Leave-one-subject-out cross-validation on the four classifiers, 70.45% (31 out of 44) of the subjects' detection accuracy has been improved).

Apart from the accuracy improvement, the ICS framework can also effectively improve the stability of the classifier. Specifically, for the two EEGNet classifiers, the application of ICS decreases their standard deviation of detection accuracy for each test subject from 0.1971 to 0.1232 and from 0.1575 to 0.1324, respectively. In contrast, the original ShallowConvNet and InterpretableCNN already have higher stability in detection accuracy, with standard deviations of 0.080 and 0.088, respectively. The improvement of their stability via ICS is limited, with a decrease

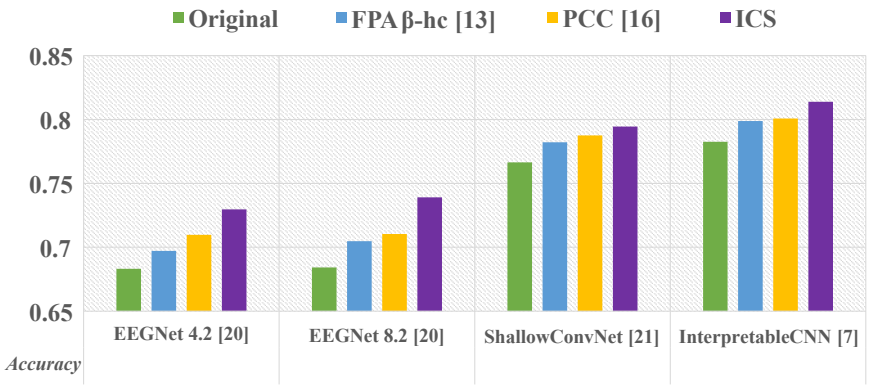

FIGURE 5.3: Comparison of ICS and other typical channel selection methods.

of 0.027 and 0.008, respectively. Hence, we can conclude that ICS can help the model to learn more informative features and thus improve the robustness of driver drowsiness detection.

5.4.4 Comparison with other Channel Selection Schemes

In this section, we compare our ICS with other typical channel selection methods to see how channel selection contributes to the drowsiness detection task. The results are shown in Fig. 5.3, which demonstrates that previous typical methods, *i.e.*, FPA β-hc, and PCC, can only slightly improve the detection accuracy by 1.5%-2.0%. In contrast, the ICS significantly outperforms the previous methods, achieving an average improvement of around 4.0%.

5.5 Summary

This thesis proposes an effective channel selection framework for driver drowsiness detection with the guidance of interpretability. Based on the interpretable CAM algorithm and the voting scheme, the proposed ICS can effectively select the most relevant channels and remove the channels that do not contribute positively to drowsiness detection. The ICS is an interpretability-guidance approach to cross-subject channel selection, taking advantage of interpretability and selecting the channels intuitively and transparently. The experiment on a public dataset demonstrates that our framework is universally applicable and can significantly improve the performance of previous models.

As for the limitations, the ICS is not an end-to-end framework, which means that it requires two training stages and is thus time-consuming, which may limited under high-dimensional scenarios. In the future, we intend to study the efficient optimization of teacher network and student network in a unified manner. We will further apply our ICS to other EEG scenarios, such as EEG epilepsy detection.

Chapter 6

Learning Interpretable Relational Probabilistic Graphs for EEG-based Emotion Recognition

6.1 Introduction

Emotion recognition has wide-ranging applications in human-computer interaction (HCI) as well as intelligent applications [244–246]. In social contexts, detecting emotions allows systems to attune better to human feelings and needs, improving experiences in customer service, virtual communication and entertainment by adapting to individuals' affective states. For healthcare, emotion recognition assists in the early diagnosis and monitoring of mental health conditions by analyzing subtle emotional changes to identify early warning signs of disorders like depression and anxiety [247–249]. Furthermore, emotion recognition can be widely applied to enhance users' cognitive performance [250] and provide interactive tools [251, 252].

A variety of emotion recognition methods have been proposed, which mainly identify emotions based on behavioral cues such as facial expressions [253–256], vocal intonations [257–259], body gestures [260] and textual information [261]. Despite much progress, these methods have inherent limitations that constrain recognition accuracy. First, behavioral cues may not precisely reflect the true emotional state due to the dissimulation of emotions by individuals, leading to misinterpretation. Second, they are susceptible to individual differences. Expressions, voice

and gestures can vary greatly across cultures, genders, ages, and so on, demanding extensive training data for emotion recognition models to account for these disparities. Third, these methods are context-dependent. Facial and gestural cues are hard to interpret without contextual understanding, resulting in the same behaviors conveying different emotions.

All the above behavior-based methods do not take into account brain signals, which can provide more accurate and objective measurements of emotional responses. Hence, EEG signals that record the electrical activity of the brain have been gradually utilized for emotion recognition [11, 40, 262–267], promoting the development of personalized and intelligent applications. However, these EEG-based methods just treat the EEG signals as time-series data, and usually employ convolutional neural networks (CNNs) [268, 269], recurrent neural networks (RNNs) [270] or graph neural networks (GNNs) [271] to learn discriminative features for emotion state classification.

Neuroscience research has elucidated that human emotion is mediated by intricate neural pathways and circuits connecting diverse brain regions, and different emotions involve different patterns of activity and connectivity across brain regions [272]. This suggests that the relationships among brain regions, especially the significance of particular regions, are crucial for emotion recognition. However, the EEG channels corresponding to different brain regions do not explicitly convey such relationships. Existing EEG-based methods treat all EEG channels equally, even learning from irrelevant channels, which weakens the learning capacity of models and thus degrades their emotion recognition performance. Therefore, considering and modelling the relationships in EEG channels can provide deeper insight into brain activities and help for emotion recognition.

Moreover, the traditional CNN, RNN and GNN models on EEG data operate as black boxes, only outputting emotion labels without explaining the relation between specific emotions and brain regions (EEG channels). This hinders their practical application in high-stake domains like healthcare, where model transparency and explainability are indispensable for informed decision-making and user acceptance.

In this chapter, we address the problems of channel relations and interpretability in EEG-based emotion recognition in a unified manner. We propose a deep learning

solution via a relational probabilistic graph convolutional network (RPGCN). The RPGCN involves a two-stage learning process. To model the relations among EEG channels, we adopt the relational thinking theory [273, 274] in the first stage to investigate the relationships among EEG channels and train a relational probabilistic graph (RPG) model. Specifically, we first leverage the relational thinking theory to learn a summary probabilistic graph that captures the overall emotional relation and variant probabilistic graphs that represent the infinite variations of the emotional relations. For each probabilistic graph, the node denotes EEG channels, and the edge indicates the relationships between nodes. The two types of probabilistic graphs are then extracted to generate an emotion relation graph, which measures the importance of different EEG channels relation. In the second stage, we apply a graph convolutional network (GCN) to learn the feature representations of emotional states from EEG data guided by the emotion relation graph. After training the RPG model, the edge weights that are post-processed can be directly used to explain the relations among EEG channels, without any additional interpretability tools required.

We conducted extensive experiments to validate the effectiveness of RPGCN. The experimental results show that our RPGCN achieves superior emotion recognition accuracy over previous methods, and also obtains intrinsic model interpretability that previous methods do not provide. The contributions of this chapter are summarized as follows:

- We propose a two-stage learning strategy for EEG-based emotion recognition. The strategy facilitates model training and model interpretability in a one-shot fashion.
- We deduce the relationships among EEG channels via relational thinking theory, and present an emotion relation graph generation method that can guide feature learning with established EEG channel relationships.
- We provide a model interpretability method that can validate the relationships between EEG channels while confirming the emotion recognition results.

The rest of this chapter is organized as follows. Section 6.2 reviews the related works from two aspects. Section 6.3 briefly introduces the basic concepts of valence, arousal and multi-regional brain interactions. The detailed description of the

proposed RGPCN is provided in Section 6.4. In Section 6.5, experiments under various conditions and analyses are presented. Finally, section 6.6 concludes this work.

6.2 Related Work

6.2.1 Behavior-based Emotion Recognition

Behavior-based emotion recognition relies on the visual cues (for instance, facial expressions, posture and textual communication) or auditory cues (for instance, vocal patterns and speech rate) of individuals, usually using some computer vision or signal processing techniques. For example, Yeh *et al.* [257] and Xie *et al.* [258] recognize emotions by analyzing the acoustic features in speech signals. Besides, Li *et al.* [275] classify different facial emotions by using CNNs with adaptive pooling maps. Similarly, Pranav *et al.* [254] follow up by constructing a deep CNN to classify five different facial emotions. Moreover, Piana *et al.* [276] extract features from 3D motion data of full-body movements to perform real-time automatic emotion recognition. There are also methods for analyzing emotions in multimodal data, specifically textual content and speech information. For instance, Xu *et al.* [277] propose to learn the alignment between speech frames and text words using attention mechanism, in order to produce more accurate emotion classification's multimodal feature representations.

A common drawback of the above behavior-based methods is that they do not directly take into account the brain signals, as the brain is the direct reflection of emotions. This causes behavior-based methods to fail to provide accurate and objective measurements of emotional responses, thus limiting their usage for intelligent applications.

6.2.2 EEG-based Emotion Recognition

Compared with behavior-based emotion recognition, applying EEG signals for direct emotion recognition has raised more and more attention recently [278–280]. Manual extraction of emotion-centric features from EEG data is difficult and runs

the risk of omitting essential details from the original signal. Therefore, directly inputting the raw EEG signals into the CNNs and RNNs, and leveraging their automatic feature extraction capabilities for emotion recognition is a better choice [281, 282]. Meanwhile, EEG signals are inherently unstable, often showing sharp contrasts between individuals. Applying domain discriminators can address the domain changes between training and test sets [263]. This technique assists in discerning more domain-agnostic features, narrowing the disparity between source and target domains.

Besides, neuroscience research underscores the profound connection between human emotions and specific cortical zones [283]. Given the multichannel composition of EEG signals, it is imperative to consider the relations between the signals when exploring the link between electrodes and the emotional influences of brain regions. An adjacency matrix, reflecting electrode topology, and a dynamical graph convolutional neural network (DGCNN) can discern intrinsic connections between nodes [265]. Another way is by categorizing electrodes based on spatial placements, fusing features, and feeding them into a feature extraction component. Such fusion facilitates the understanding of the correlations within and between brain sectors [40]. Beyond leveraging spatial relations for emotion classification, integrating spatio-temporal information in emotion signals has gained attraction among researchers. Scholars have engineered separate modules for handling spatial and temporal data, harnessing capsule neural networks [266] and demographic networks [267] for emotion recognition.

Although the above methods achieve good emotion recognition results, their black-box property limits their application to intelligent applications. For instance, neither a commercial customer nor a patient would pay for an application lacking in interpretability. Therefore, developing interpretable EEG-based emotion recognition techniques is highly essential.

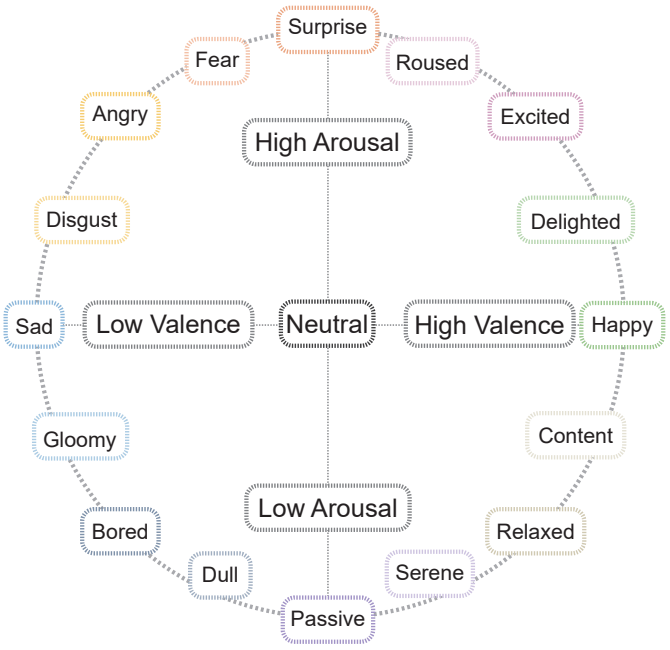

FIGURE 6.1: Two-dimensional map of valence and arousal.

6.3 Preliminary

6.3.1 Valence & Arousal

Valence and arousal are two fundamental dimensions in characterizing emotional states [284–286]. Valence describes the inherent appeal or averseness of an emotional event, object or situation. It determines whether an emotion is positive, such as the high valence associated with happiness, or negative, as seen with the low valence of sadness. Arousal pertains to an individual's physiological and psychological reaction to stimuli, indicating emotional intensity. This spectrum ranges from the high arousal experienced during excitement to the subdued arousal associated with calmness.

The combination of valence and arousal creates a two-dimensional emotional space that allows for a more nuanced understanding of emotions. Emotions can be positioned within this space based on their valence and arousal levels (low to high), as shown in Fig. 6.1.

6.3.2 Multi-regional Brain Interactions

Neuroscience models often conceptualize the brain as a hierarchy of interconnected nodes, where each region forms part of a more extensive neural network. In particular, they suggest that any brain region receives significant input from many other regions rather than operating in isolation [287]. In addition, brain-related tasks are not limited to specific brain regions but rather to whole-brain patterns [288]. This framework of broadly distributed neural circuits facilitates the studying of complex cognitive functions emerging from interactions across large populations of neurons spanning multiple brain areas, such as the generation of emotions.

6.4 Relational Probabilistic Graph Convolutional Network

6.4.1 Motivation and Problem Formulation

Relational thinking refers to the cognitive process of understanding and analyzing relationships between different elements or entities [273, 289]. It involves identifying connections, patterns and dependencies among various pieces of information or concepts. This motivates us to use the relational thinking to reveal the hidden mechanisms underlying the emotional activities in various brain regions, and further analyze the relationships among EEG channels to enable learning feature representations of emotional states.

To apply relational thinking to EEG signals, we represent the EEG signals as a graph structure. The reasons are as follows. First, relational thinking focuses on relationships, specifically how entities connect. The graph structure explicitly captures these relations as connections between nodes, aligning closely with the core focus of relational thinking. Second, EEG measures electrical activity from neuronal interactions in the brain. These interactions can be effectively modeled as a graph, with nodes denoting EEG channels (for instance, brain regions) and edges denoting functional connections, as demonstrated in [265, 271, 290].

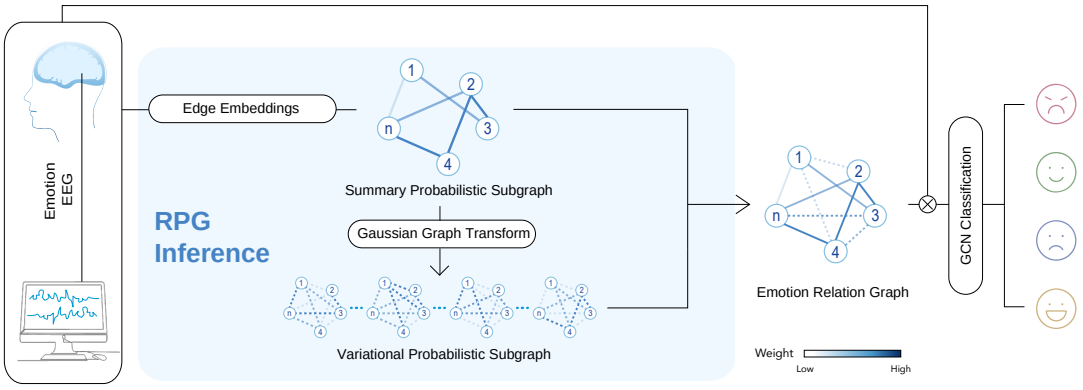

FIGURE 6.2: Architecture of RPGCN. First, RPG inference extracts edge embeddings between the signal nodes and generates the summary probabilistic subgraph. Second, the variant probabilistic subgraphs represent the variable emotion states transformed from the summary probabilistics graph via Gaussian graph transforms. Next, an emotion relation graph representing the emotion relations is obtained by merging the two subgraphs. Finally, the emotion relation graph and EEG signals are fed into a graph convolutional network to extract emotional feature representations, which are further fed into a classifier to predict emotion labels.

Previous works directly learned feature representations from raw EEG data or graph-structured EEG data for emotion recognition. They have difficulty in capturing intricate relationships in EEG channels, and the learned graphs also fail to explain recognition results. In contrast, we formulate EEG-based emotion recognition as a two-stage recognition task, and propose a two-stage learning strategy to learn emotional features progressively.

The overall architecture of the relational probabilistic graph convolutional network (RPGCN) is shown in Fig. 6.2.

The first stage is the RPG inference task, which obtains the emotion relation graphs that represent the relationships of EEG channels. Let $X = \{X_t \in \mathbb{R}^{C\times N} | t \in T\}$ denotes the EEG signals in T time segments and $G^{Emo} = \{G_t^{Emo} | t \in T\}$ denotes a set of emotion relation graphs obtained from X, in which C is the number of EEG channels and N is the number of sampling points within a time segment. This task can be formalized as

$$G^{Emo} = \mathrm{RPG}(X), \tag{6.1}$$

where RPG denotes the RPG inference function. See Section 6.4.2 for details of the RPG inference process.

The second stage is the classification task, which uses a GCN to map the input EEG signals to the corresponding emotion categories with the guidance of emotion relation graphs. This task is formalized as

$$\hat{y} = \mathrm{GCN}(X, G^{Emo}), \tag{6.2}$$

where GCN denotes a GCN model and $\hat{y}$ denotes the label of the emotion category. See the GCN details in Section 6.4.3.

Finally, the two stages are progressively learned with carefully designed loss functions, which are detailed in Section 6.4.4.

6.4.2 RPG Inference

As shown in Fig. 6.2, the RPG inference aims to model the emotion relation as the posterior graphs' edge distribution and obtain a emotion relation graph.

For this purpose, we first transform the EEG signals into initial graph structures, which are used for further emotional relational learning. To better model the intricate emotional activities within graph-structured EEG, we introduce two types of graphs based on relational thinking: the summary probabilistic graph and the variant probabilistic graph. The former is designed to capture the general emotional relations, while the latter, generated from the former, can represent the possible variations of the emotional relations. Then we extract the two types of graphs to generate an emotion relation graph, which contains the overall emotional activities and variant emotional activities, thus simulating the emotional state of the human in reality.

6.4.2.1 Node Embedding and Edge Embedding

The input EEG signals are first converted into graphs by forming nodes from EEG channels and extracting edge embeddings between EEG channels. Specifically, for the t-th EEG signal segment $X_t \in \mathbb{R}^{C\times N}$, we adopt a linear layer $\boldsymbol{f}_{node}(\cdot)$ to convert the EEG data from each channel into individual node embeddings of the graph, and then concatenate every two nodes to generate edge embeddings with a

convolutional network $\boldsymbol{f}_{edge}(\cdot)$. The node and edge embedding generation processes are as follows:

$$v_{i,t} = \boldsymbol{f}_{node}(x_{i,t}), \tag{6.3}$$

$$\boldsymbol{e}_{i,j,t} = \boldsymbol{f}_{edge}(v_{i,t}, v_{j,t}), \tag{6.4}$$

where $i, j \in [1, C]$ and $i \neq j$ are the channel/node indexes, $x_{i,t}$ denotes the EEG data in channel i of X_t, $v_{i,t} \in \mathbb{R}^{1\times D_n}$ denotes the node embedding corresponding to $x_{i,t}$, and $\boldsymbol{e}_{i,j,t} \in \mathbb{R}^{1\times D_e}$ represents the edge embedding for nodes i and j. D_n and D_e represent the dimensions of the node embedding and edge embedding, respectively.

6.4.2.2 Summary Probabilistic Graph

To accurately recognize the emotion state, it is necessary to reveal the global dependency among the brain regions. To this end, we seek to construct a graph that represents general emotion relations. We refer to such a graph as a summary probabilistic graph, which is denoted by G^{Sum}.

The summary probabilistic graph G^{Sum} is built upon the phenomenon of multi-regional brain interactions, as introduced in Section 6.3.2. The connection between two neurons is affected not only by the nodes directly linked to it, but also by other neurons whose connections pass through it. The associated neuron connections may be innumerable. That is to say, the edges actually depend on innumerable nodes.

Let $\lambda_{i,j} = \Pr(e_{i,j})$ denote the probability of edge existence, and $\alpha_{i,j}$ denote the edge that follows a specific distribution. Here, we use a distinct symbol $\alpha_{i,j}$ to differentiate the deduced edge and the initial edge $e_{i,j,t}$, and also omit the subscript t for notational convenience.

The basic neural interaction involves the propagation of action potentials (spikes) between neurons. This spike transmission can be modeled as a binary event — either a spike is transmitted or not. Therefore, we assume the edge $\alpha_{i,j}$ follows a Bernoulli distribution:

$$\alpha_{i,j}^{n} \sim \mathbf{Bern}(\lambda_{i,j}), \tag{6.5}$$

where **Bern** denotes the Bernoulli distribution, superscript $n \in [0, +\infty]$ is the index of additional node n other than node i and j, and $\lambda_{i,j} \to 0$ indicates that single neural activity is physiologically slight [291].

Then the edge $\alpha_{i,j}^{Sum}$ in G^{Sum} can be constructed as a combination of all $\alpha_{i,j}^{n}$, and is given by

$$\alpha_{i,j}^{Sum} = \sum_{n=1}^{\infty} \alpha_{i,j}^{n}. \tag{6.6}$$

Since it is computationally intractable to compute all $\alpha_{i,j}^{n}$ in Eq. (6.6) directly, an equivalent binomial distribution is employed to denote the edge distribution in G^{Sum}:

$$\alpha_{i,j}^{Sum} \sim \lim_{n\to\infty,\lambda_{i,j}\to 0} \mathbf{Bin}(n, \lambda_{i,j}), \tag{6.7}$$

where **Bin** denotes the binomial distribution, and $n \to \infty$ means that the $\alpha_{i,j}^{Sum}$ is affected by innumerable nodes.

According to VRNN [292], the parameters of the approximate posterior distribution can be estimated via an RNN encoder. However, the edge distribution (for instance, $\mathbf{Bin}(n, \lambda_{i,j})$) of the summary probabilistic subgraph has an infinity parameter n, which causes the inference and sampling cannot be computed directly. To solve the problem of the infinity parameter n and near-zero parameter $\lambda_{i,j}$, we apply the De Moivre-Laplace theorem [293] that the binomial distribution $\mathbf{Bin}(n, \lambda_{i,j})$ can be approximated by a Gaussian distribution $\mathcal{N}(n\lambda_{i,j}, n\lambda_{i,j}(1-\lambda_{i,j})$ when n tends to infinity:

$$\alpha_{i,j}^{Sum} \sim \mathcal{N}(\mu_{i,j}^{Sum}, \mu_{i,j}^{Sum}(1-\mu_{i,j}^{Sum})) \tag{6.8}$$

where $\mu_{i,j}^{Sum}$ denotes the mean of the Gaussian approximation of the binomial distribution and is defined as

$$\mu_{i,j}^{Sum} = \frac{(1-2\mu_{i,j})+2\sigma_{i,j}^{2}}{1-2\mu_{i,j}} - \frac{\left|1-2\mu_{i,j}+\sqrt{(1-2\mu_{i,j})^{2}+4\sigma_{i,j}^{2}}\right|}{1-2\mu_{i,j}} \tag{6.9}$$

$$q_{i,j} = \frac{1}{1-2\mu_{i,j}} \sim \text{Softplus}(\mu_{i,j}) + \epsilon_{i,j} \tag{6.10}$$

$$\mu_{i,j}^{Sum} \sim \frac{1+2q_{i,j}\sigma_{i,j}^{2}-\sqrt{1+4q_{i,j}^{2}\sigma_{i,j}^{4}}}{2} \tag{6.11}$$

A Gaussian distribution $\mathcal{N}(\mu_{i,j}, \sigma_{i,j})$ with $\mu_{i,j} < \frac{1}{2}$ is required before computing the posterior approximation variable $\alpha_{i,j}^{Sum}$ for the binomial edge variables. Hence, in

Eq. (6.9), $\mu_{i,j}$ and $\sigma_{i,j}$ are computed from the input nodes with two convolutional layers f_{mean} and f_{std} respectively. In Eq. (6.10), to avoid the explosion of $\frac{1}{1-2\mu_{i,j}}$, we define it as an approximation $q_{i,j}$ with a very small hyperparameter $\epsilon_{i,j}$ to constrain its lower bound. As a result, $\alpha_{i,j}^{Sum}$ is further simplified as Eq. (6.11). This allows the re-parametrization to draw samples from the posterior distribution $Q(\alpha_{i,j}^{Sum})$ from this Gaussian approximation $\mathcal{N}(\mu_{i,j}^{Sum}, \mu_{i,j}^{Sum}(1-\mu_{i,j}^{Sum}))$ [294].

6.4.2.3 Variant Probabilistic Graph

The conduction of variant probabilistic subgraph information relies upon the presence of the summary probabilistic subgraph. Specifically, based on the sample $z_{i,j}^{Sum}$ from the distribution of summary probabilistic subgraph, we introduce Gaussian variables $\mathcal{N}(\tilde{\mu}_{i,j,t}, \tilde{\sigma}_{i,j,t}^2)$ to define the variant probabilistic subgraph:

$$z_{i,j}^{Sum} = \sqrt{\mu_{i,j}^{Sum}(1-\mu_{i,j}^{Sum})}\,\varepsilon_{i,j}^{Sum} + \mu_{i,j}^{Sum} \tag{6.12}$$

$$\alpha_{i,j,t}^{Var} \sim \mathcal{N}(z_{i,j}^{Sum}\,\tilde{\mu}_{i,j,t},\ z_{i,j}^{Sum}\,\tilde{\sigma}_{i,j,t}^2) \tag{6.13}$$

where $z_{i,j}^{Sum}$ denotes the re-parametrization sampling result from the posterior distribution $Q(\alpha_{i,j}^{Sum})$ of the summary probabilistic subgraph. The Gaussian distribution $\mathcal{N}(\tilde{\mu}_{i,j,t}, \tilde{\sigma}_{i,j,t}^2)$ is generated by a linear layer as the Gaussian variable for the Gaussian graph transforms [295] to obtain the conditional probability distribution $Q(\alpha_{i,j,t}^{Var})$.

$$z_{i,j,t}^{Var} = \sqrt{z_{i,j}^{Sum}}\,\sigma_{i,j,t}^{Var}\,\varepsilon_{i,j,t}^{Var} + z_{i,j}^{Sum}\,\mu_{i,j,t}^{Var} \tag{6.14}$$

where $z_{i,j,t}^{Var}$ represents edge sampling from the variant probabilistic subgraph.

6.4.2.4 Emotion Relation Graph

In the final step of RPG inference, to extract emotional feature representation, based on two subgraphs generated in the summary probabilistic graph and the variant probabilistic graph, we define the edge of the emotion relation graph as follows:

$$\alpha_{i,j,t}^{Emo} = z_{i,j,t}^{Var} \, z_{i,j}^{Sum} \tag{6.15}$$

$\alpha_{i,j,t}^{Emo}$ describes the edges of the final emotion relation graph generated from the summary probabilistic subgraph and variant probabilistic subgraph.

6.4.3 Graph Convolutional Network Emotion Classification

Given a graph $G = (V, E)$ with $V = \{v_{i,t}\}$ and $E = \{e_{i,j,t}\}$, GCNs aim to derive a function f that updates node representations by aggregating information from their neighbors. The core operation of a GCN is defined by:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{6.16}$$

where $H^{(l)}$ denotes the node feature matrix at layer l, $\tilde{A}$ denotes the adjacency matrix with self-loops, $\tilde{D}$ denotes the diagonal node degree matrix of $\tilde{A}$, $W^{(l)}$ denotes the layer l weight matrix, and σ denotes the ReLU activation function.

Following the RPG inference, we apply GCN for emotion recognition. For classification, the representation $H^{(l+1)}$ after applying the propagation rule is subjected to a Sigmoid operation to predict the class probabilities Z for each emotion:

$$Z = \text{Sigmoid}\left(H^{(l+1)}W^{(l+1)}\right) \tag{6.17}$$

$$Z = \text{Sigmoid}\left(\sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)W^{(l+1)}\right) \tag{6.18}$$

6.4.4 Learning of Relational Probabilistic Graphs

We apply variant inference [296] to optimize modeling probability distributions over RPG inference. We define our specific optimization target as the following evidence lower bound (ELBO):

$$ELBO = \min\{\text{KL}\left[Q(G^{Emo} \mid X)||P(G^{Emo} \mid X)\right] + CE(Y, Y'|G^{Emo}, X)\} \tag{6.19}$$

$$\begin{aligned}&\mathrm{KL}\left[Q(G^{Emo} \mid X)||P(G^{Emo} \mid X)\right] = \\ &\sum_{(i,j)\in E}\{KL[Q(\alpha_{i,j}^{Sum})||P(\alpha_{i,j}^{Sum})) + KL(Q(\alpha_{i,j}^{Var})||P(\alpha_{i,j}^{Var})]\}\end{aligned} \tag{6.20}$$

In Eq. (6.19), the ELBO for RPG inference contains three terms.
$\mathrm{KL}\left[Q(G^{Emo} \mid X)||P(G^{Emo} \mid X)\right]$ is the combination of two KL terms. Since the RPG inference can be represented as two types of random variables: binomial variables related to the edges of the summary probabilistic subgraph $\alpha_{i,j}^{Sum}$ and Gaussian variables related to the edges of the variant probabilistic subgraph $\alpha_{i,j,t}^{Emo}$, this KL term can be further divided into summary KL term (Sum KL term) and emotion KL term (Emo KL term) in Eq. (6.20). Here $\min\{CE(Y, Y'|G^{Emo}, X)$ is the Emo Inference Term. It represents the subject's emotional inference generated from RPG inference via input EEG signals.

6.4.4.1 Term 1: Summary KL Term.

We define the summary probabilistic subgraph as the basis of RPG inference across time segments to construct the dynamic change of the variant probabilistic subgraph between time segments. RPG inference defines the computation of the summary probabilistic subgraph as follows:

$$E_{i,j}^{Sum} = f_{edge}(\sum_{t=1}^{T}(v_{i,t}), \sum_{t=1}^{T}(v_{j,t})(i, j \in [1, C]\&\&i \neq j) \tag{6.21}$$

$$\lambda_{i,j} = \mathrm{Softplus}(f_{\lambda}(E_{i,j}^{Sum})) \tag{6.22}$$

$$\mu_{i,j} = f_{\mu}(E_{i,j}^{Sum}) \tag{6.23}$$

$$\sigma_{i,j} = \mathrm{Softplus}(f_{\sigma}(E_{i,j}^{Sum})) \tag{6.24}$$

where $f_{edge}(\cdot)$, $f_{\lambda}(\cdot)$, $f_{\mu}(\cdot)$, and $f_{\sigma}(\cdot)$ denote the four different networks used to compute edge embeddings of a summary probabilistic subgraph $E_{i,j}^{Sum}$ and the variables of the two distributions. $\mu_{i,j}$ and $\sigma_{i,j}$ are variables of the Gaussian distributions $\mathcal{N}(\mu_{i,j}, \sigma_{i,j})$ referred to in Eq. (6.9), which are used to compute Gaussian approximations to the posterior binomial distribution. The $\lambda_{i,j}$ denotes the probability of the prior binomial distribution in Eq. (6.7). However, the Summary

KL term is computationally intractable since the posterior probability distribution of the summary probabilistic subgraph edge $Q(\alpha_{i,j}^{Sum}) = \mathbf{BIN}(n, \lambda_{i,j})$ has an infinite parameter n. We transform the incalculable Summary KL term into an approximation as a closed-form solution that is irrelevant to the infinite parameter n:

$$KL[Q(\alpha_{i,j}^{Sum})||P(\alpha_{i,j}^{Sum})] \sim \sum_{(i,j)\in E} \{\mu_{i,j}^{Sum} \log \frac{\mu_{i,j}^{Sum} + \epsilon}{\lambda_{i,j} + \epsilon} + (1 - \mu_{i,j}^{Sum}) \log \frac{1 - \mu_{i,j}^{Sum} + {\mu_{i,j}^{Sum}}^2/2 + \epsilon}{1 - \lambda_{i,j} + {\lambda_{i,j}}^2/2 + \epsilon}\} \tag{6.25}$$

where ϵ is a tiny hyperparameter to prevent explosions.

6.4.4.2 Term 2: Variant KL Term.

In Step 3, we further define the calculation of the posterior distribution of variant probabilistic subgraph edges $Q(\alpha_{i,j,t}^{Var})$ in Eq. (6.13) and the corresponding prior distribution $P(\alpha_{i,j,t}^{Var})$ as follows:

$$Q(\alpha_{i,j,t}^{Var}) \sim \mathcal{N}(z_{i,j}^{Sum} \times f_{\tilde{\mu}}(E_{i,j,t}), z_{i,j}^{Sum} \times \zeta(f_{\tilde{\sigma}}(E_{i,j,t}))^2) \tag{6.26}$$

$$P(\alpha_{i,j,t}^{Var}) \sim \mathcal{N}(z_{i,j}^{Sum} \times f_{\bar{\mu}}(E_{i,j,:t}), z_{i,j}^{Sum} \times \zeta(f_{\bar{\sigma}}(E_{i,j,:t}))^2) \tag{6.27}$$

where $f_{\tilde{\mu}}$, $f_{\tilde{\sigma}}$, $f_{\bar{\mu}}$, $f_{\bar{\sigma}}$, and $f_{\bar{\sigma}}$ are four different networks used to generate the Gaussian variable to compute the prior and posterior distributions, respectively. $E_{i,j,t}$ and $E_{i,j,:t}$ represent the variant probabilistic subgraph edge of the t^{th} time segment and the variant probabilistic subgraph edge of the segments before the t^{th} time segment. We abbreviate the two distributions as $Q(\alpha_{i,j,t}^{Var}) \sim \mathcal{N}(\mu_{i,j,t}^{q}, {\sigma_{i,j,t}^{q}}^2)$ and $P(\alpha_{i,j,t}^{Var}) \sim \mathcal{N}(\mu_{i,j,t}^{p}, {\sigma_{i,j,t}^{p}}^2)$. Therefore, the variant KL term can be computed as:

$$KL\left(Q\left(\alpha_{i,j}^{Var}\right) \| P\left(\alpha_{i,j}^{Var}\right)\right) = \sum_{t\in T,(i,j)\in E^t} \left\{ 2\log\left(\frac{\sigma_{i,j|t}^{p} + \epsilon}{\sigma_{i,j|t}^{q} + \epsilon}\right) + \frac{{\sigma_{i,j|t}^{q}}^2 + \left(\mu_{i,j|t}^{q} - \mu_{i,j|t}^{p}\right)^2}{\left(\sigma_{i,j|t}^{p} + \epsilon\right)^2} - 1 \right\} \tag{6.28}$$

where ϵ is a tiny hyperparameter to prevent explosions.

6.4.4.3 Term 3: Emotion Relation Term.

According to Eq. (6.15), we define an adjacency matrix $A_t^{Emo} = [\alpha_{i,j,t}^{Emo}]$, and we define the Emotion Relation Inference process as follows:

$$\bar{V}_t = GCN(V_t, A_t^{Emo}) \tag{6.29}$$

$$\bar{Y} = f_{n2y}(\sum_{t \in T}(\bar{V}_t)) \tag{6.30}$$

where $GCN(\cdot)$ denotes a GCN, $V_t = \{v_{c,t} | c \in C\}$ denotes the node in the t^{th} time segment, and $\bar{V}_t$ denotes the node based on the emotion relation graph after updating the weights by the GCN. Finally, we combine the signals of the nodes in different time segments and apply a classification network f_{n2y} to obtain the input signal's emotion classification result $\bar{Y}$. This result will be used with the ground truth Y to compute the CE loss in Eq. (6.19) as the emotion relation term in the RPG optimization objective.

6.4.5 Model Interpretability of RPGCN

The interpretability of the RPGCN comes from its model's self-interpretability. In the first stage, the RPG inference uses relational thinking theory to model the relationships among EEG channels. Via this method, the variant probabilistic graphs reflect the complex emotional connections and diverse variations through nodes symbolizing EEG channels and edges illustrating the connections between these channels. Combining these probabilistic graphs to generate an emotion relation graph highlights the importance of various EEG channels and establishes a basis for the model's interpretability. During the second phase, the GCN utilizes feature representations from the emotion relation graph and achieves robust emotion classification. The emotion relation graph plays a key role in guiding the GCN's learning process by providing interpretable relational information among EEG channels, which is essential for comprehending emotional states.

The RPGCN's interpretability is crucial because it can offer direct insights into the relationships among EEG channels after training. After training, the edge weights

in the emotion relation graph are adjusted to clarify the connections between the channels. This straightforward interpretation strategy eliminates the requirement for external interpretability tools, therefore simplifying the process of comprehending how the model identifies and utilizes the connections among EEG channels to detect emotional states.

6.5 Experiment and Discussion

6.5.1 Datasets

6.5.1.1 Deap

The Database for Emotion Analysis using Physiological Signals (Deap) dataset [45] is a multimodal physiological database tailored for emotion analysis. It was constructed with the primary objective of exploring the emotional responses of individuals when subjected to visual stimuli, notably through measurements of EEG and peripheral physiological signals such as Electrocardiography (ECG), Galvanic Skin Response (GSR), and others. The dataset comprises 32 participants exposed to 40 different one-minute music video clips. These clips were carefully curated to elicit varying emotional responses, characterized along dimensions of valence (high/low) and arousal (high/low). After each video clip, participants provided ratings regarding the emotions they experienced. To be specific, the dataset includes 1) EEG data across 32 channels. 2) Other peripheral physiological signals: electrocardiography (ECG), galvanic skin response (GSR), respiration, and skin temperature.

6.5.1.2 Dreamer

The Dreamer dataset [46] stands as a significant benchmark in biologically driven emotion recognition. Designed to further our understanding of the interconnections between physiological signals and emotional experiences, Dreamer offers an in-depth exploration of emotional states through both EEG and ECG measurements. This dataset encompasses 23 participants with movie clips tailored to elicit specific emotional reactions across multiple dimensions. The key among these is valence

(positive/negative) and arousal (high/low), facilitating a nuanced analysis of the spectrum of emotional responses. Specifically, the Dreamer dataset includes: 1) EEG recordings across 14 channels, capturing the nuances of brain activity in response to emotional stimuli. 2) ECG recording across 14 channels offers insight into the cardiac responses associated with different emotions.

6.5.2 Implementation Details

6.5.2.1 Experiment Setting

We applied k-fold cross-validation and trained a specific model for each subject, where k was set to 10. In each step of 10-fold cross-validation, one fold was selected as the test set, and the rest were randomly divided into $8:2$, where the 80% were utilized for training and the 20% were applied for validation. Then, we crop each data into $4s$ non-overlapping segments to ensure that cropped experiments [196] are conducted without data leakage. We apply the Adam optimizer with a learning rate of $1e-5*3$ and a weight decay of $1e-3$. The network was trained for 500 epochs, and the five models with the highest validation accuracy were selected to be averaged as the best model for testing. The above training process was repeated 10 times for each subject, and the average accuracy and average F1 score of all folds were finally reported as the final results.

6.5.2.2 Operating Environment

For our experiments, we utilized a computational environment fortified with specific software and hardware configurations. On the software front, we used Python 3.8.11, PyTorch (version 1.8.0) and NumPy (version 1.20.2). Hardware-wise, our system was powered by an Intel (R) Xeon (R) CPU E5-2620 v4 clocked at 2.10 GHz, buttressed by a substantial 256 GB of RAM, and accelerated by a GeForce Tesla P40 GPU. As for the configuration of the training hyper-parameters, the details are listed in Table 6.1.

Hyper-parameter	Value
HiddenDim	128
GraphDim	128
Sequence Dim	512
Node Embedding Dim	128
Edge Embedding Dim	128
Epoch	500
Heads	4
Time Series	32
Weight Decay	$1 \times \mathbf{e} - 3$
Learning Rate	$3 \times \mathbf{e} - 5$
eps (ϵ)	$1 \times \mathbf{e} - 6$
$\epsilon_{i,j}$	$1 \times \mathbf{e} - 5$
Dropout	0.5
Conv Channel	8

TABLE 6.1: The optimal values of hyper-parameters.

6.5.3 Experiment Results & Comparison with Prior Art

Table 6.2 demonstrates the emotion recognition performance of our proposed RPGCN on the Deap and Dreamer datasets, respectively. The results of the experiments include I) accuracy and F1 scores for per-subject on the Deap and Dreamer dataset (see Fig. 6.3 and Fig. 6.4), II) the overall accuracy and F1 score on the Deap and Dreamer dataset, and the comparison against the results from the existing ten other baselines (see Table. 6.2).

Per-Subject Result: For the per-subject result on the Deap dataset, RPGCN achieves an accuracy of 80.46% for arousal and 79.28% for valence, and an F1 score of 79.13% for arousal and 77.21% for valence. The standard deviations in accuracy for arousal and valence are 4.77 and 5.48. For the F1 score, they are 3.46 for arousal and 3.74 for valence. For the per-subject result on the Dreamer dataset, the accuracies for arousal and valence are 80.68% and 77.28%. The F1 scores for arousal and valence on this dataset are 79.88% and 77.73%. The standard deviations in accuracy for arousal and valence are 4.77 and 5.48, respectively. The standard deviations in accuracy on the Dreamer dataset are 4.12 for arousal and

Method	Deap				Dreamer			
	Arou Acc	Arou F1	Vale Acc	Vale F1	Arou Acc	Arou F1	Vale Acc	Vale F1
DGCNN [265]	75.23	75.63	70.38	73.45	74.28	73.88	75.30	75.82
RGNN [271]	70.56	69.18	69.94	68.11	68.82	67.45	66.51	67.77
EEGNet [83]	75.65	74.88	70.61	66.08	70.65	71.17	69.43	67.11
TSception [297]	65.34	64.91	67.39	52.18	66.32	65.43	67.27	65.48
SpikingNN [298]	63.15	60.28	66.23	66.21	62.72	64.39	64.29	63.85
3DCANN [299]	72.15	69.81	65.23	64.18	70.23	71.81	61.67	62.89
MSDTT [300]	60.23	59.94	61.08	62.84	61.12	59.03	59.88	60.37
SS-STANN [301]	75.56	76.18	72.94	76.11	72.82	71.45	71.51	70.77
MTGNN [302]	74.76	73.54	72.84	74.13	70.28	70.55	72.71	73.12
ASTG-LSTM[303]	75.72	74.15	76.34	76.22	74.85	75.47	72.82	72.33
ST-GCLSTM [304]	76.47	76.53	75.83	75.17	74.27	74.78	74.01	73.26
RPGCN (Ours)	**80.46**	**79.13**	**79.28**	**77.21**	**80.68**	**79.88**	**77.28**	**77.73**

TABLE 6.2: Comparison (in %) of RPGCN and Baselines on the Deap and Dreamer Datasets.

Method	Deap				Dreamer			
	Arou Acc	Arou F1	Vale Acc	Vale F1	Arou Acc	Arou F1	Vale Acc	Vale F1
CNN-Spa	68.61	67.82	67.61	66.28	60.12	60.08	61.28	61.96
CNN-Temp	68.37	66.91	67.89	64.88	60.74	60.91	60.19	62.81
TF-Spa	61.23	60.11	62.18	56.21	61.61	60.22	60.23	60.89
TF-Temp	62.33	59.17	60.12	60.01	61.33	61.01	61.88	62.21
CNN-Joint	75.66	73.82	74.71	72.13	71.01	67.92	70.08	69.92
TF-Joint	76.92	77.18	75.04	73.41	70.56	70.11	69.94	68.11
RPG Inference	**80.46**	**79.13**	**79.28**	**77.21**	**80.68**	**79.88**	**77.28**	**77.73**

TABLE 6.3: Comparison of RPG Inference and other Feature Extractors on the Deap and Dreamer Datasets.

3.78 for valence. The corresponding F1 score standard deviations are 2.65 for arousal and 2.97 for valence. Intriguingly, we also note that for both datasets, the difficulty of predicting the two sentiment dimensions is not consistent. Considering the tradeoff between accuracy and F1 score, we find that valence is harder to predict for Deap, while arousal is harder to predict for Dreamer.

Comparison Results: The comparison results contain a comparison of RPGCN with other baselines that have different characteristics. Among them, EEGNet [83] has the relational inference capability by capturing features in the time and

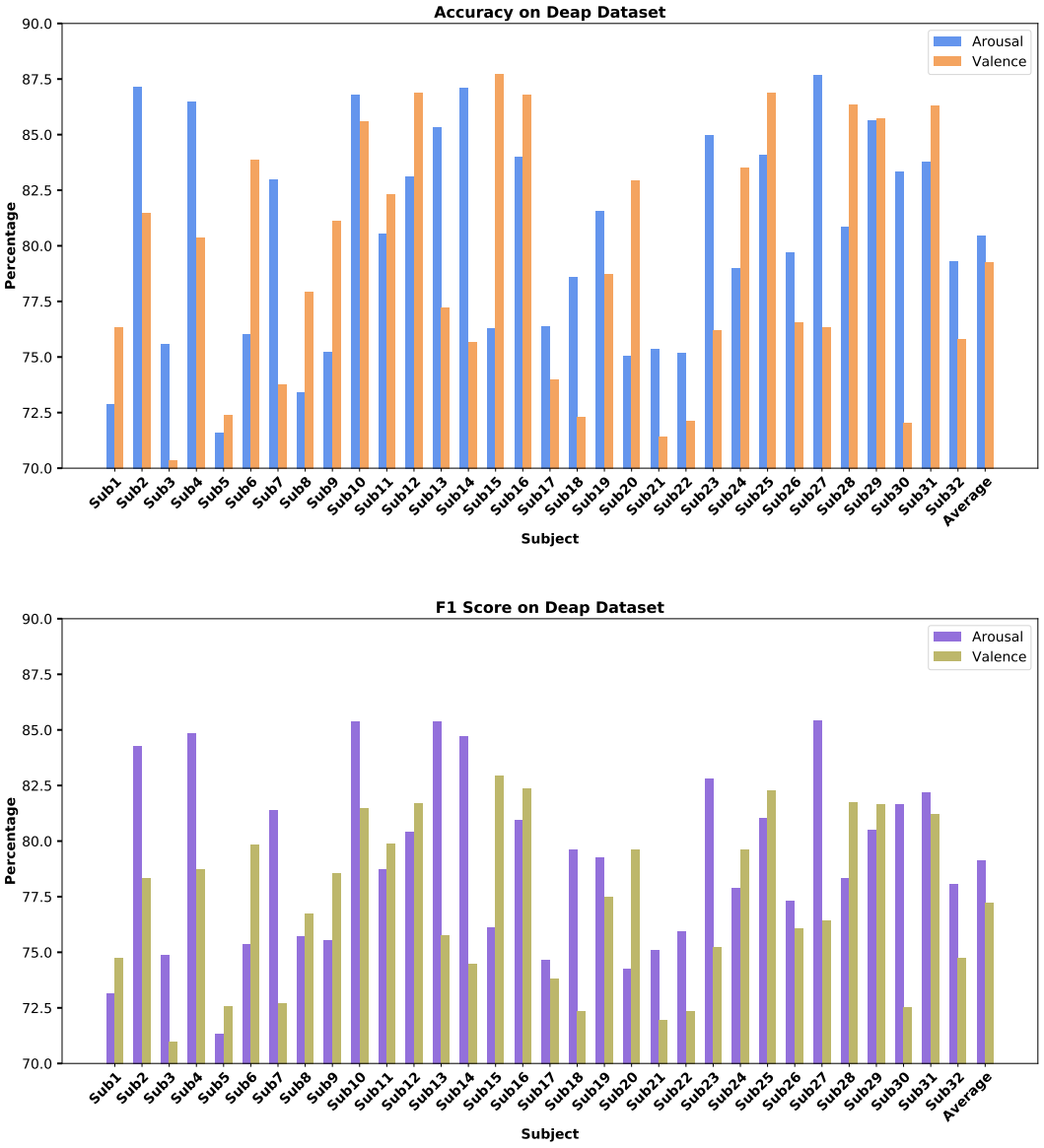

FIGURE 6.3: **Accuracy and F1 Score of RPGCN on Deap Dataset**

frequency domains through convolutional and pooling layers. However, the scope of convolutional attention is usually localized and cannot fully explain its spatio-temporal relations.

The impulse transfer in Spiking NN (SNN) [298] can infer the signal implicit relation. Information transfer between impulse neurons is based on the temporal differences and patterns of the impulses, but it cannot take into account the spatial information of the signal. TSception [297] and MSDTT [300] introduce independent temporal and spatial inference components in the model to infer spatio-temporal relations. TSception employs two convolutional networks in two temporal and

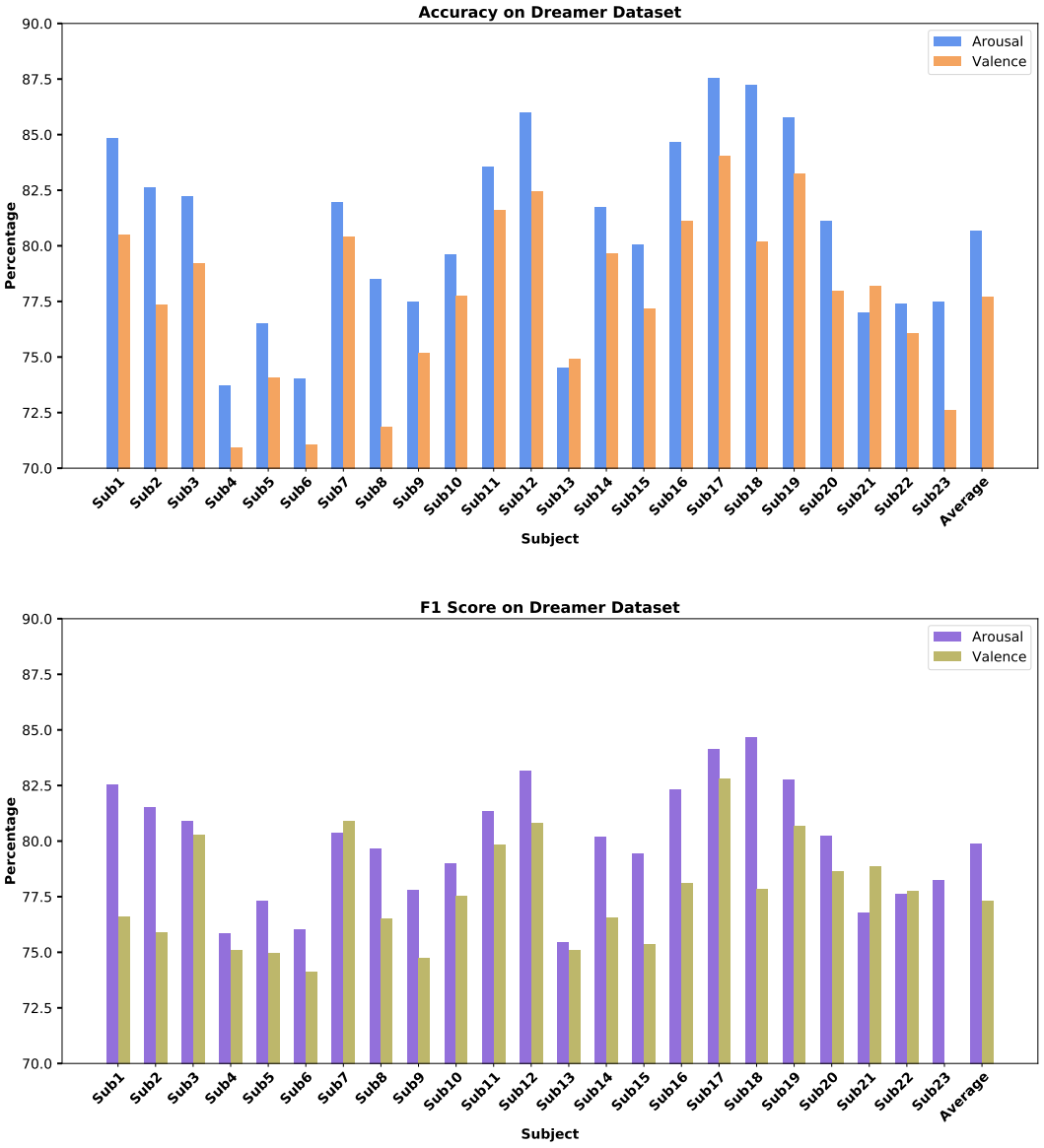

FIGURE 6.4: **Accuracy and F1 Score of RPGCN on Dreamer Dataset**

spatial dimensions. MSDTT contains the multi-domain spatial transformer (MST) module and a dynamic temporal transformer (DTT). Specifically, when performing temporal inference, the temporal relation built by the temporal component is only a correlation between different time segments. In contrast, the complex spatial relations within the time segments are ignored. Therefore, it limits the inference capability of models.

3DCANN [299] and SS-STANN [301] apply a spatio-temporal encoder as the core for feature extraction, representing the spatio-temporal relation of EEG signals as an abstract high-latitude feature through a convolutional network. However, this

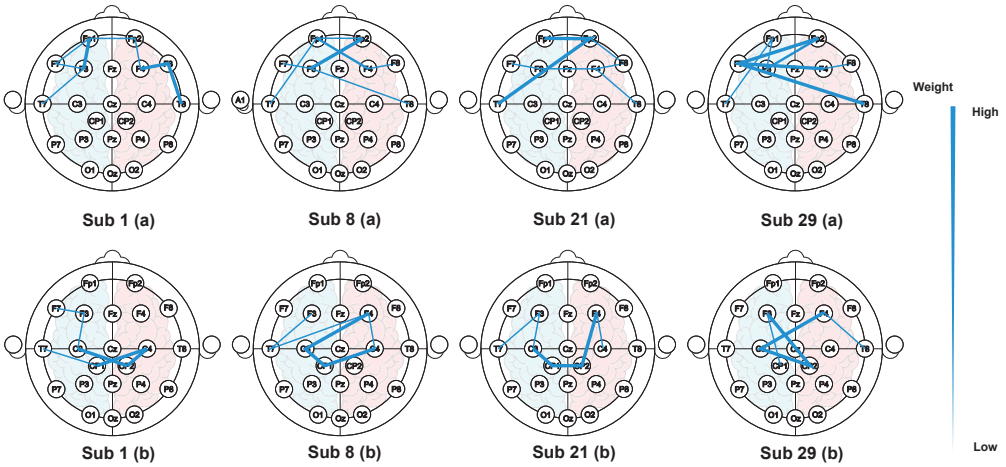

FIGURE 6.5: **Interpretabe Emotion Relation Feature Visualization on Deap Dataset.** Subfigures (a) indicate that the emotion relationship of positive emotions varies in different subjects. In contrast, the emotion relationship is located in the prefrontal cortex and anterior insula regions. Subfigures (b) indicate that the emotion relationship of negative emotions varies in different subjects, while the emotion relationship is located in the postcentral gyrus region.

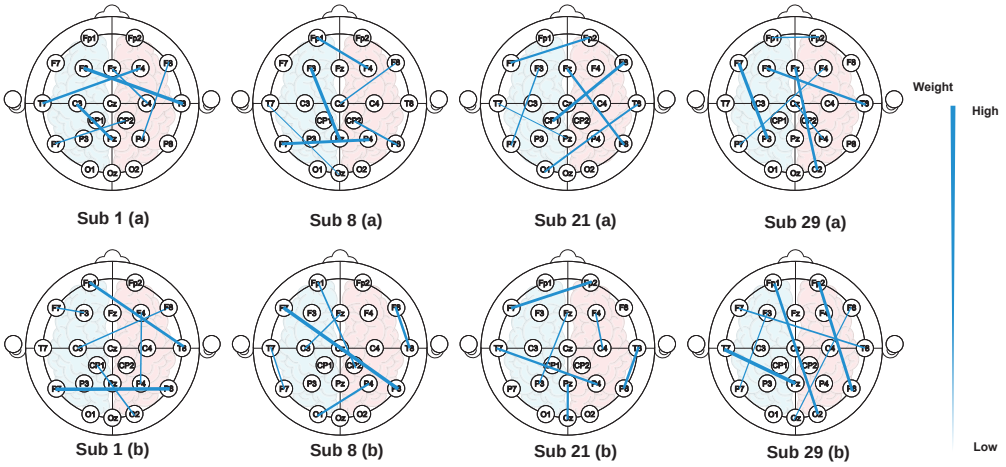

FIGURE 6.6: **Comparsion Emotion Feature Visualization on Deap Dataset.** The visualization illustrates that traditional GCN with a random-initiated adjacency matrix cannot capture the relation between specific brain regions (EEG channels) and emotion.

spatio-temporal relation is not interpretable and lacks theoretical support from brain physiology. It is only built at the feature level, ignoring the brain's topology, resulting in limited performance. In a direct comparison with 3DCANN and SS-STANN, the prowess of RPGCN becomes more evident.

Similarly, for the graph-based methods, DGCNN [265], MTGNN [302], ASTG-LSTM [303], and ST-GCLSTM [304] parameterize the spatial or temporal relation as part of the model, which is updated by loss. The above methods transform the EEG signals into spatial or frequency domain features and do not include the brain's topology, resulting in limited performance. To compare, RGNN [271] introduces specific prior knowledge as the spatial dependency. The prior knowledge used by RGNN to define spatial dependency is that correlations between brain regions show regular decay with physical distance. However, leveraging only prior knowledge cannot adequately represent the robust emotion classification due to missing other critical features. Leveraging the model's self-inference to infer brain physiology-aligned features would be more convincing.

6.5.4 Relation Feature Visualization Comparison

To further validate the interpretability of the RPG inferred emotion relation and have an intuitive comparison with the traditional GCN method (without RPG inference), we model the weight of the GCN's input adjacency matrix as the emotion relation. Specifically, the difference between the RPGCN and the traditional GCN is that the adjacency matrix of the traditional GCN is randomly initialized without the RPG inference guided. We visualize the emotion relation of the EEG signals of selected subjects under high-arousal & high-valence (positive emotion) and low-arousal & low-valence (negative emotion) in the Deap dataset. We have retained only the most salient parts to clear the range of emotion relations. Row (a) in Fig. 6.5 illustrates the emotion relation of positive emotion and their most related brain area. Negative emotion relations are shown in Fig. 6.5 row (b).

In Fig. 6.5 row (a), the universal emotion relation of the signals is mainly concentrated in the first half of the brain while varying in different subjects, which shows that similar emotions are highly variable. [305] has shown that positive emotions significantly elevate the correlations between the prefrontal cortex and the anterior insula region, where the prefrontal cortex corresponds to Fp1, Fp2, F3, F4, F7, and

F8, and the anterior insula corresponds to T7 and T8. The emotion relations associated with negative emotions in Fig. 6.5 row (b) are centered on the mid-posterior sides of the brain and also vary in different subjects. This is also consistent with the conclusion in [305] that the brain under negative emotions is associated with a strong connection in the postcentral gyrus. The postcentral gyrus corresponds to C3, C4, CP1, and CP2. Our experimental result proves that the RPG inference can infer not only the highly variable emotion relation but also the universal emotion relation, which is consistent with the physiological phenomenon of the brain.

To compare with, in Fig. 6.6, without the RPG inference guidance, the visualized adjacency matrix's weight of the traditional GCN cannot illustrate the correlations between EEG channels and either positive emotions or negative emotions. The random-initiated adjacency matrix cannot effectively model the emotion relations with the specific brain region, which shows the interpretable superiority of our RPG inference.

6.5.5 Ablation Study

To verify the advantages of RPG inference, we replace it with other traditional feature extractors and do the ablation studies on the Deap and Dreamer datasets, respectively. The results of the ablation study (Table 6.3) show that our RPG inference has an advantage over the typical feature extractors. They are the CNN spatial feature extractor (CNN-Spa), CNN temporal feature extractor (CNN-Temp), Transformer spatial feature extractor (TF-Spa), Transformer temporal feature extractor (TF-Temp), CNN spatio-temporal feature extractor (CNN-Joint), and Transformer spatio-temporal feature extractor (TF-Joint), respectively. We applied the same experiment setting as the main experiment. The experiment result in Table 6.3 demonstrates that our RPG inference outperforms the traditional feature extractors.

6.6 Summary

In this thesis, we propose a novel EEG emotion recognition method, RPGCN, based on the relational thinking. Our method can infer the variations of emotion

relations in the EEG signals and achieve state-of-the-art performance on the Deap and Dreamer datasets with specifically designed summary probabilistic subgraph and variant probabilistic subgraphs. We also demonstrate the importance of RPG inference through ablation studies. In addition, the structure of RPGCN is interpretable, and the universal emotion relation revealed by RPG is consistent with brain physiological phenomena, which can be further used in developing personalized and intelligent applications for many proposes, including VR/AR, product design/recommendation, and medical scenarios applications.

Chapter 7

Conclusion and Future Directions

7.1 Conclusions

This thesis has presented comprehensive advancements in EEG systems, leveraging the power of AI to address significant challenges and improve the interpretability and robustness of various applications. By thoroughly examining the literature on interpretable and robust AI techniques for EEG systems in Chapter 2, we identified and categorized methods to enhance the interpretability and robustness of AI EEG models. Our proposed taxonomy of interpretability methods, to be specific, is backpropagation, perturbation, and rule-based, along with categorizing robustness based on undesirable factors. They are noise and artifacts, human variability, data acquisition instability, and adversarial attacks. This literature review also provides a structured framework for future research.

This thesis introduces several innovative frameworks and architectures that significantly advance the state-of-the-art in EEG applications. In Chapter 3, the HASS framework has demonstrated superior performance in cross-subject sleep staging tasks, utilizing a spatio-temporal attention mechanism to adaptively weight EEG segments based on their spatial and temporal relationships. This approach has significantly improved the MASS and ISRUC datasets, offering a promising solution for clinical and research applications in sleep assessment.

Furthermore, the development of EENED and GlepNet architectures for neural epilepsy detection has shown remarkable improvements in the accuracy of diagnosis

are demonstarted in Chapter 4. By interleaving convolution model with multi-head attention mechanism, these architectures enhance the robustness of epilepsy signal representations with the benefit of Grad-CAM interpretability, thereby elevating their clinical utility.

In Chapter 5, the ICS framework has addressed the challenge of identifying key contributing channels for EEG driver drowsiness detection. By implementing a two-stage training strategy and utilizing class activation mapping, ICS has demonstrated significant performance improvements in cross-subject detection tasks, offering a more reliable solution for driver monitoring systems.

Finally, the RPGCN in Chapter 6 has advanced the field of EEG-based emotion recognition by transforming raw EEG signals into probabilistic graphs. This method effectively models the relationships among EEG channels, providing interpretability and aligning with cognitive neuroscience findings. RPGCN's superior performance in emotion classification tasks underscores the potential for integrating brain activity analysis into more intelligent and personalized HCI systems.

In conclusion, this thesis has significantly advanced EEG systems by addressing critical challenges in interpretability and robustness. The contributions of this thesis are unified by a shared focus on enhancing interpretability and robustness in EEG-based AI systems. While each work addresses distinct research areas, which span sleep staging, neural epilepsy detection, driver drowsiness monitoring, and emotion recognition, they collectively advance the overarching objectives of this research. Together, these studies illustrate how innovative methodologies can effectively tackle critical EEG analysis challenges, highlighting the significance of developing robust and interpretable EEG systems. The proposed frameworks and architectures create opportunities for more reliable and transparent applications across clinical and commercial interaction domains, enhancing human life through advanced AI-driven EEG technologies.

7.2 Limitations

In this section, we discuss several limitations encountered during the research on interpretable and robust AI techniques for EEG systems. These limitations highlight areas for improvement and future research directions:

Data Quality and Quantity: One of the primary challenges in EEG-based research is the variability in data quality. EEG signals are often contaminated with noise and artifacts from various sources, including muscle activity, eye movements, and external electrical interference. Additionally, the quantity of high-quality labeled EEG data is limited, which constrains the training and validation of complex AI models. The reliance on small datasets can lead to overfitting and limit the generalizability of the models.

Inter-Subject and Intra-Subject Variability: EEG signals exhibit significant variability both between different subjects (inter-subject) and within the same subject over different sessions (intra-subject). This variability poses a significant challenge for developing models that can generalize well across different individuals and conditions. Transfer learning and domain adaptation techniques have been explored to address these issues, but there is still a long way to go to achieve robust solutions.

Interpretability vs. Performance Trade-off: While interpretability is a key focus of this research, achieving a balance between model interpretability and performance remains challenging. Highly interpretable models, such as linear models and decision trees, often lack the accuracy and robustness of more complex models. Conversely, models with high performance, such as deep learning models, are often criticized for being "black-box" in nature, making it difficult to understand their decision-making processes.

Computational Complexity: The methodologies employed in this research, such as CAM and hybrid attention mechanisms, often involve significant computational complexity. This complexity can limit the applicability of these methods in real-time or resource-constrained environments. Future research needs to focus on optimizing these techniques to make them more computationally efficient without sacrificing accuracy or interpretability.

Scalability and Adaptability: While the research demonstrates promising results in specific applications like driver fatigue detection and emotion recognition, scalability and adaptability to other EEG-based applications remain a concern. Developing scalable AI models that can adapt to a wide range of EEG applications, including medical diagnostics and cognitive state monitoring, requires further exploration.

Integration with Other Modalities: The current research primarily focuses on EEG signals. However, integrating EEG with other modalities, such as fNIRS, ECG, and behavioral data, could provide a more comprehensive understanding of brain activity and improve the robustness and accuracy of AI models. The complexity of multimodal data fusion and the associated interpretability challenges need to be addressed in future research.

These limitations underscore the need for continued research and development in the field of interpretable and robust AI for EEG systems. Addressing these challenges will pave the way for more reliable and practical applications of EEG-based BCI and AI-driven neurotechnologies.

7.3 Future Directions

Interpretability and robustness techniques offer a promising future for building better EEG systems. Despite much success, there are still some unresolved problems worthy of in-depth study. Therefore, we discuss a few promising directions.

7.3.1 Future Directions for Interpretable AI in EEG Systems

7.3.1.1 Prior Human Knowledge and Brain Inspired Design

A significant limitation of existing interpretable EEG systems lies in their inability to integrate prior information effectively. In EEG systems, prior information refers to established physiological principles that can guide model behavior. Current attribution methods, while capable of identifying correlations between features and predictions, often fail to ensure that models focus on relevant features or avoid those that are less meaningful. This limitation can result in model learning patterns that deviate from established domain knowledge. To address this, interpretable EEG systems in real-world scenarios should aim to jointly learn the relationship between prior information and feature importance, ensuring that explanations are grounded in features deemed essential by domain experts.

Weinberger *et al.* [306] introduced the deep attribution prior (DAPr) framework, which integrates prior knowledge into models by imposing constraints through a prior model. For example, in neuroscience, seizures are known to result from sudden abnormal discharges of neurons in the temporal lobe. However, EEG systems for seizure diagnosis are often susceptible to noise generated by patient movements, such as muscle artifacts. By incorporating the seizure region as prior knowledge, models can be guided to prioritize relevant features, enabling predictions to align with established medical knowledge while reducing the influence of noise.

Beyond the integration of prior human knowledge, brain-inspired design provides an additional pathway to enhance the interpretability of EEG systems. By emulating the structural and functional organization of the brain, these designs ensure that models align more closely with neurophysiological principles, offering biologically plausible explanations. For instance, modular architectures inspired by hierarchical brain structures can facilitate more interpretable decision-making. Separate modules can be designed to process low-level signal features, such as oscillatory patterns, and high-level cognitive features, such as task-related dynamics. This approach mirrors how different brain regions specialize in distinct functions, making decision-making more transparent and easier to interpret. By bridging neuroscience and machine learning, brain-inspired designs have the potential to advance both the interpretability and robustness of EEG systems significantly.

7.3.1.2 High-dimensional Feature Interpretation

The existing interpretable methods in EEG systems mainly reveal the contributions of features to predictions, but lack insight into why features are assigned specific contributing values. Providing dynamic feature descriptions, rather than only linear relationships between features and predictions, can be a promising way to reveal the inner logic of model predictions. For example, Zhang *et al.* [307] designed special loss for each convolutional layer, instructing them to focus on certain regions within the input image. Sabour *et al.* [308] proposed capsule networks to parse the entire object into a parsing tree of capsules by a dynamic routing mechanism, in which each capsule may encode a specific meaning of input data.

We expect interpretable AI in future EEG systems to gain insight from the above two works to provide hidden semantics to explain feature correlations. For instance,

in the MI task, high-dimensional features could be interpreted as concurrency between low-latitude motor cortex signal features and visual cortex signal features. Alternatively, with models' hidden semantics, we can know how models' attention is drawn to noises if high-dimensional features can be interpreted as similarities between noise and essential features. By incorporating dynamic routing algorithms or specialized loss functions into EEG systems, we can guide the model to focus on specific semantic features, and enhance the robustness of the system while maintaining interpretability.

7.3.2 Future Directions for Robust AI in EEG Systems

7.3.2.1 Artificial Synthetic Data and Large Models

Large models have been used in NLP and CV, demonstrating impressive performance coupled with robustness. This is achieved by training with large amounts of data and thus naturally performing well against anomalies. However, there has been little work on applying large models to EEG systems due to the scarcity of available EEG data. Some existing works utilize traditional generative models, such as GANs [309, 310], to artificially synthesize new EEG data for data augmentation.

However, the performance of these works is rather limited because of a lack of proper EEG generation mechanisms. In the future, it is imperative to develop more sophisticated models for EEG data synthesis. Additionally, EEG-oriented data augmentation based on signal processing or adversarial examples remains a potential direction for overcoming data scarcity. Once sufficient EEG data is available, these approaches, alongside brain foundation models, can facilitate the application of large pre-trained models to EEG systems. By adapting foundational architectures to the specificities of EEG data, researchers can develop more robust, generalizable, and interpretable systems for clinical and practical applications.

Moreover, the concept of brain foundation models holds significant promise for advancing EEG-based AI systems. Brain foundation models, aim to create universal representations by pretraining on diverse and large-scale EEG datasets. These models can act as a backbone for a wide range of EEG tasks, such as emotion recognition, sleep staging, and epilepsy detection, enabling efficient transfer learning across tasks with minimal fine-tuning.

A critical component of building brain foundation models is the use of self-supervised learning techniques. By leveraging unlabeled EEG data, self-supervised methods can extract robust and transferable features that capture the underlying structure of EEG signals. Techniques like contrastive learning, masked signal modeling, and predictive coding have shown promise in learning meaningful representations without requiring extensive labeled data. These features, when combined with task-specific fine-tuning, can significantly enhance the robustness and adaptability of EEG systems.

7.3.2.2 Decoupling of EEG Signals for Robust Feature

EEG signals contain diverse information, including subject identity and task-related information (for example, MI and emotion recognition). On the one hand, the identity information in EEG signals is more difficult to forge than other biometric information (for example, face, iris and fingerprint), so it can be used for more reliable identity recognition. On the other hand, the identity information is also a kind of noise that affects the performance of other tasks. Therefore, disengaging the identity information from the EEG signals can make the EEG systems more resilient to subject variations, thus enhancing the robustness of EEG systems and allowing for better cross-subject applications.

However, various types of information in EEG signals are highly coupled and interfere with each other, which hinders their applications. Thus, how to decouple the EEG Signals for designing robust features is a promising direction.

7.3.3 Building Interpretable and Robust EEG Systems

Building human-trusted EEG systems have been a long-term goal pursued by academics for many years. Using interpretability to identify potential problems and vulnerabilities in models can improve the robustness of EEG systems. In addition, incorporating prior human knowledge and interpreting hidden semantics can allow systems to better learn from experts. It also enables consumers to better understand how the systems work. This will be helpful for future academics to improve models and develop better EEG systems to meet consumers' requirements.

Bibliography

[1] Xinliang Zhou, Chenyu Liu, Liming Zhai, Ziyu Jia, Cuntai Guan, and Yang Liu. Interpretable and robust ai in eeg systems: A survey. *arXiv preprint arXiv:2304.10755*, 2023. 5, 56

[2] Xinliang Zhou, Chenyu Liu, Jiaping Xiao, and Yang Liu. Eeg-based sleep staging with hybrid attention. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 112–115. IEEE, 2023. 6

[3] Chenyu Liu, Xinliang Zhou, and Yang Liu. Eened: End-to-end neural epilepsy detection based on convolutional transformer. *arXiv preprint arXiv:2305.10502*, 2023. 6

[4] Xinliang Zhou, Chenyu Liu, Ruizhi Yang, Liangwei Zhang, Liming Zhai, Ziyu Jia, and Yang Liu. Learning robust global-local representation from eeg for neural epilepsy detection. *IEEE Transactions on Artificial Intelligence*, 2024. 6, 20

[5] Xinliang Zhou, Dan Lin, Ziyu Jia, Jiaping Xiao, Chenyu Liu, Liming Zhai, and Yang Liu. An eeg channel selection framework for driver drowsiness detection via interpretability guidance. *arXiv preprint arXiv:2304.14920*, 2023. 7

[6] Xinliang Zhou, Chenyu Liu, Jiaping Xiao, Yang Liu, et al. Learning relational probabilistic graphs for eeg-based emotion recognition. *Preprint on ResearchGate*, March 2024. 7

[7] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021. 11

[8] Zhongke Gao, Xinmin Wang, Yuxuan Yang, Chaoxu Mu, Qing Cai, Weidong Dang, and Siyang Zuo. EEG-based spatio–temporal convolutional neural network for driver fatigue evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2755–2763, 2019. 11, 86

[9] Ji-Hoon Jeong, Kyung-Hwan Shim, Dong-Joo Kim, and Seong-Whan Lee. Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(5):1226–1238, 2020. 11

[10] Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent M Berry, and Yogatheesan Varatharajah. Evaluating latent space robustness and uncertainty of EEG-ML models under realistic distribution shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 11

[11] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing*, 2022. 11, 17, 96

[12] Takashi Nishimoto, Hiroshi Higashi, Hiroshi Morioka, and Shin Ishii. EEG-based personal identification method using unsupervised feature extraction and its robustness against intra-subject variability. *Journal of Neural Engineering*, 17(2):026007, 2020. 11

[13] Zhanyu Ma, Xiaoou Lu, Jiyang Xie, Zhen Yang, Jing-Hao Xue, Zheng-Hua Tan, Bo Xiao, and Jun Guo. On the comparisons of decorrelation approaches for non-gaussian neutral vector variables. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 13

[14] Yimin Hou, Shuyue Jia, Xiangmin Lun, Ziqian Hao, Yan Shi, Yang Li, Rui Zeng, and Jinglei 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. 14

[15] Hubert Cecotti and Axel Graser. Convolutional neural networks for p300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):433–445, 2010. 14, 16

[16] Jun Zhang, Meng Wang, Shengping Zhang, Xuelong Li, and Xindong Wu. Spatiochromatic context modeling for color saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1177–1189, 2015. 14

[17] Richard B Berry, Rohit Budhiraja, Daniel J Gottlieb, David Gozal, Conrad Iber, Vishesh K Kapur, Carole L Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F Quan, et al. Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine. *Journal of Clinical Sleep Medicine*, 8 (5):597–619, 2012. 15

[18] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. SalientSleepNet: Multimodal salient wave detection network for sleep staging. *ArXiv Preprint arXiv:2105.13864*, 2021. 16

[19] Ziyu Jia, Junyu Ji, Xinliang Zhou, and Yuhan Zhou. Hybrid spiking neural network for sleep electroencephalogram signals. *Science China Information Sciences*, 65(4):140403, 2022. 16, 42, 43

[20] Mohamed Ragab, Emadeldeen Eldele, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Self-supervised autoregressive domain adaptation for time series data. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[21] Yuchen Liu and Ziyu Jia. BSTT: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2023. 16, 42

[22] Afraim Salek-Haddadi, Beate Diehl, Khalid Hamandi, Martin Merschhemke, Adam Liston, Karl Friston, John S Duncan, David R Fish, and Louis Lemieux. Hemodynamic correlates of epileptiform discharges: an EEG-fMRI study of 63 patients with focal epilepsy. *Brain Research*, 1088(1):148–166, 2006. 16

[23] Fabrice Bartolomei, Patrick Chauvel, and Fabrice Wendling. Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG. *Brain*, 131(7):1818–1830, 2008. 16

[24] Jian Cui, Zirui Lan, Olga Sourina, and Wolfgang Müller-Wittig. EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 16, 20, 25, 85, 86, 88, 90, 91

[25] Yu-Ting Liu, Yang-Yin Lin, Shang-Lin Wu, Chun-Hsiang Chuang, and Chin-Teng Lin. Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network. *IEEE transactions on neural networks and learning systems*, 27(2):347–360, 2015. 16

[26] Jiaxin Ma, Yu Zhang, Andrzej Cichocki, and Fumitoshi Matsuno. A novel EOG/EEG hybrid human–machine interface adopting eye movements and erps: Application to robot control. *IEEE Transactions on Biomedical Engineering*, 62(3):876–889, 2014. 16

[27] Manoj Thulasidas, Cuntai Guan, and Jiankang Wu. Robust classification of EEG signal for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):24–29, 2006. 16

[28] Hubert Cecotti, Miguel P Eckstein, and Barry Giesbrecht. Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering. *IEEE Transactions on Neural Networks and Learning Systems*, 25(11):2030–2042, 2014. 16

[29] Jing Jin, Zhiqiang Wang, Ren Xu, Chang Liu, Xingyu Wang, and Andrzej Cichocki. Robust similarity measurement based on a novel time filter for

SSVEPs detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 16

[30] Phairot Autthasan, Rattanaphon Chaisaen, Thapanun Sudhawiyangkul, Phurin Rangpong, Suktipol Kiatthaveephong, Nat Dilokthanakul, Gun Bhakdisongkhram, Huy Phan, Cuntai Guan, and Theerawit Wilaiprasitporn. MIN2Net: End-to-end multi-task learning for subject-independent motor imagery EEG classification. *IEEE Transactions on Biomedical Engineering*, 69 (6):2105–2118, 2021. 16

[31] Yiliang Liu, Wenbin Su, Zhijun Li, Guangming Shi, Xiaoli Chu, Yu Kang, and Weiwei Shang. Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks. *IEEE Transactions on Cognitive and Developmental Systems*, 11(3):414–424, 2018. 17

[32] Yu Zhang, Tao Zhou, Wei Wu, Hua Xie, Hongru Zhu, Guoxu Zhou, and Andrzej Cichocki. Improving eeg decoding via clustering-based multitask feature learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (8):3587–3597, 2021. 17

[33] Kang Wang, Di-Hua Zhai, Yuhan Xiong, Leyun Hu, and Yuanqing Xia. An mvmd-cca recognition algorithm in SSVEP-based BCI and its application in robot control. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):2159–2167, 2021. 17

[34] Yang Yu, Yadong Liu, Jun Jiang, Erwei Yin, Zongtan Zhou, and Dewen Hu. An asynchronous control paradigm based on sequential motor imagery and its application in wheelchair navigation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(12):2367–2375, 2018. 17

[35] Dongdong Li, Li Xie, Zhe Wang, and Hai Yang. Brain emotion perception inspired EEG emotion recognition with deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 17

[36] Smith K Khare and Varun Bajaj. Time–frequency representation and convolutional neural network-based emotion recognition. *IEEE transactions on neural networks and learning systems*, 32(7):2901–2909, 2020.

[37] Cunbo Li, Peiyang Li, Yangsong Zhang, Ning Li, Yajing Si, Fali Li, Zehong Cao, Huafu Chen, Badong Chen, Dezhong Yao, et al. Effective emotion recognition by learning discriminative graph topologies in EEG brain networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 17

[38] Unsoo Ha, Yongsu Lee, Hyunki Kim, Taehwan Roh, Joonsung Bae, Changhyeon Kim, and Hoi-Jun Yoo. A wearable EEG-HEG-HRV multimodal system with simultaneous monitoring of tes for mental health management. *IEEE Transactions on Biomedical Circuits and Systems*, 9(6):758–766, 2015. 17

[39] Wolfgang Klimesch. Memory processes, brain oscillations and EEG synchronization. *International Journal of Psychophysiology*, 24(1-2):61–100, 1996. 17

[40] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 17, 96, 99

[41] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface. *IEEE transactions on neural networks and learning systems*, 24(4):610–619, 2013. 17

[42] Shuailei Zhang, Kai Keng Ang, Dezhi Zheng, Qianxin Hui, Xinlei Chen, Yang Li, Ning Tang, Effie Chew, Rosary Yuting Lim, and Cuntai Guan. Learning EEG representations with weighted convolutional siamese network: A large multi-session post-stroke rehabilitation study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2824–2833, 2022. 17

[43] VK Benzy, AP Vinod, R Subasree, Suvarna Alladi, and K Raghavendra. Motor imagery hand movement direction decoding using brain computer interface to aid stroke recovery and rehabilitation. *IEEE transactions on neural systems and rehabilitation engineering*, 28(12):3051–3062, 2020. 17

[44] Kai Keng Ang and Cuntai Guan. Brain-computer interface in stroke rehabilitation. *Journal of Computing Science and Engineering*, 7(2):139–146, 2013. 17

[45] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011. 18, 19, 111

[46] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1): 98–107, 2017. 18, 19, 111

[47] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016. 18, 19

[48] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. 19

[49] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, 2018. 19

[50] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021. 19

[51] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011. 19

[52] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12 (2):479–493, 2018. 19

[53] Arman Savran, Koray Ciftci, Guillaume Chanel, Javier Mota, Luong Hong Viet, Blent Sankur, Lale Akarun, Alice Caplier, and Michele Rombaut. Emotion detection in the loop from brain signals and facial images. In *Proceedings of the eNTERFACE 2006 Workshop*, 2006. 19

[54] Julie A Onton and Scott Makeig. High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in Human Neuroscience*, page 61, 2009. 19

[55] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6(1):19, 2019. 19, 90

[56] Sadegh Arefnezhad, James Hamet, Arno Eichberger, Matthias Frühwirth, Anja Ischebeck, Ioana Victoria Koglbauer, Maximilian Moser, and Ali Yousefi. Driver drowsiness estimation using EEG signals with a dynamical encoder–decoder modeling framework. *Scientific Reports*, 12(1):1–18, 2022. 19

[57] Manasa Kalanadhabhatta, Chulhong Min, Alessandro Montanari, and Fahim Kawsar. FatigueSet: A multi-modal dataset for modeling mental fatigue and fatigability. In *Pervasive Computing Technologies for Healthcare*, pages 204–217. Springer, 2022. 19

[58] Ali H Shoeb and John V Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 975–982, 2010. 19

[59] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity:

Dependence on recording region and brain state. *Physical Review E*, 64(6): 061907, 2001. 19, 60, 72

[60] Matthias Ihle, Hinnerk Feldwisch-Drentrup, César A Teixeira, Adrien Witon, Björn Schelter, Jens Timmer, and Andreas Schulze-Bonhage. EPILEPSIAE–a european epilepsy database. *Computer Methods and Programs in Biomedicine*, 106(3):127–138, 2012. 19

[61] Nathan J Stevenson, Karoliina Tapani, Leena Lauronen, and Sampsa Vanhatalo. A dataset of neonatal EEG recordings with seizure annotations. *Scientific Data*, 6(1):1–8, 2019. 19

[62] Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. 19

[63] Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The NMT scalp EEG dataset: an open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling. *Frontiers in Neuroscience*, 15:1764, 2022. 19

[64] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. 19

[65] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018. 19

[66] Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, 23(6):628–635, 2014. 19, 43

[67] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. 19, 43

[68] D Alvarez-Estevez and RM Rijsman. Haaglanden medisch centrum sleep staging database (version 1.0. 1). *PhysioNet*, 2021. 19

[69] Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, Robert Thomas, and Arthur Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep. *Sleep Medicine*, 2(6):537–553, 2001. ISSN 1389-9457. doi: 10.1016/S1389-9457(01)00149-6. 19

[70] Miles E Drake Jr, Ann Pakalnis, Jodie M Andrews, and Janet E Bogner. Nocturnal sleep recording with cassette EEG in chronic headaches. *Headache: The Journal of Head and Face Pain*, 30(9):600–603, 1990. 19

[71] Thomas Penzel, Martin Glos, Carmen Garcia, Christoph Schoebel, and Ingo Fietze. The siesta database and the siesta sleep analyzer. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8323–8326. IEEE, 2011. 19

[72] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot Mueller-Putz, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, page 55, 2012. 19

[73] Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. 19

[74] Kai Keng Ang, Cuntai Guan, Kok Soon Phua, Chuanchu Wang, Ling Zhao, Wei Peng Teo, Changwu Chen, Yee Sien Ng, and Effie Chew. Facilitating effects of transcranial direct current stimulation on motor imagery brain-computer interface with robotic feedback for stroke rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 96(3):S79–S87, 2015. 19

[75] Kai Keng Ang, Cuntai Guan, Kok Soon Phua, Chuanchu Wang, Longjiang Zhou, Ka Yin Tang, Gopal J Ephraim Joseph, Christopher Wee Keong Kuah, and Karen Sui Geok 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*, 7: 30, 2014. 19

[76] Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. 19

[77] Hangyu Zhu, Yonglin Wu, Ning Shen, Jiahao Fan, Linkai Tao, Cong Fu, Huan Yu, Feng Wan, Sio Hang Pun, Chen Chen, et al. The masking impact of intra-artifacts in EEG on deep learning-based sleep staging systems: A comparative study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1452–1463, 2022. 20

[78] Ji-Seon Bang, Min-Ho Lee, Siamac Fazli, Cuntai Guan, and Seong-Whan Lee. Spatio-spectral feature representation for motor imagery classification using convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 20, 23

[79] Charles A Ellis, Rongen Zhang, Vince D Calhoun, Darwin A Carbajal, Robyn L Miller, and May D Wang. A gradient-based approach for explaining

multimodal deep learning classifiers. In *21st IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1–6. IEEE, 2021. 22

[80] Aarthy Nagarajan, Neethu Robinson, and Cuntai Guan. Relevance based channel selection in motor imagery brain-computer interface. *Journal of Neural Engineering*, 2022. 23

[81] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods*, 274:141–145, 2016. 23

[82] Minji Lee, Leandro RD Sanz, Alice Barra, Audrey Wolff, Jaakko O Nieminen, Melanie Boly, Mario Rosanova, Silvia Casarotto, Olivier Bodart, Jitka Annen, et al. Quantifying arousal and awareness in altered states of consciousness using interpretable deep learning. *Nature Communications*, 13(1): 1064, 2022. 20

[83] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. 20, 24, 91, 114

[84] Valentin Gabeff, Tomas Teijeiro, Marina Zapater, Leila Cammoun, Sylvain Rheims, Philippe Ryvlin, and David Atienza. Interpreting deep learning models for epileptic seizure detection on EEG signals. *Artificial Intelligence in Medicine*, 117:102084, 2021. 24

[85] Amirhessam Tahmassebi, Jennifer Martin, Anke Meyer-Baese, and Amir H Gandomi. An interpretable deep learning framework for health monitoring systems: A case study of eye state detection using EEG signals. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 211–218. IEEE, 2020. 20, 28

[86] Christoph Jansen, Stephan Hodel, Thomas Penzel, Martin Spott, and Dagmar Krefting. Feature relevance in physiological networks for classification of obstructive sleep apnea. *Physiological Measurement*, 39(12):124003, 2018.

[87] Chengxuan Tong, Yi Ding, Kevin Lim Jun Liang, Zhuo Zhang, Haihong Zhang, and Cuntai Guan. TESANet: Self-attention network for olfactory EEG classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2022. 20

[88] Abdulnasir Yildiz, Hasan Zan, and Sherif Said. Classification and analysis of epileptic EEG recordings using convolutional neural network and class activation mapping. *Biomedical Signal Processing and Control*, 68:102720, 2021. 20, 25

[89] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization.

IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31: 710–719, 2022. 56

[90] Fei Wang, Shichao Wu, Weiwei Zhang, Zongfeng Xu, Yahui Zhang, Chengdong Wu, and Sonya Coleman. Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia*, 146:107506, 2020. 26

[91] Ce Zhang, Young-Keun Kim, and Azim Eskandarian. EEG-inception: an accurate and robust end-to-end neural network for EEG-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014, 2021. 20, 34, 35

[92] Bingxiu Liu, Jifeng Guo, CL Philip Chen, Xia Wu, and Tong Zhang. Fine-grained interpretability for EEG emotion recognition: Concat-aided grad-cam and systematic brain functional network. *IEEE Transactions on Affective Computing*, 2023. 20

[93] Yurong Li, Hao Yang, Jixiang Li, Dongyi Chen, and Min Du. EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam. *Neurocomputing*, 415:225–233, 2020. 26

[94] Chunguang Chu, Zhen Zhang, Jiang Wang, Shang Liu, Fei Wang, Yanan Sun, Xiaoxuan Han, Zhen Li, Xiaodong Zhu, and Chen Liu. Deep learning reveals personalized spatial spectral abnormalities of high delta and low alpha bands in EEG of patients with early parkinson's disease. *Journal of Neural Engineering*, 18(6):066036, 2021.

[95] Yosuke Fujiwara and Junichi Ushiba. Deep residual convolutional neural networks for brain–computer interface to visualize neural processing of hand movements in the human brain. *Frontiers in Computational Neuroscience*, 16:882290, 2022.

[96] He Chen, Yan Song, and Xiaoli Li. Use of deep learning to detect personalized spatial-frequency abnormalities in EEGs of children with adhd. *Journal of Neural Engineering*, 16(6):066046, 2019. 20

[97] Michele Lo Giudice, Nadia Mammone, Cosimo Ieracitano, Maurizio Campolo, Arcangelo Ranieri Bruna, Valeria Tomaselli, and Francesco Carlo Morabito. Visual explanations of deep convolutional neural network for eye blinks detection in EEG-based BCI applications. In *International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022. 20, 27

[98] Haneen Alsuradi and Mohamad Eid. Trial-based classification of haptic tasks based on EEG data. In *IEEE World Haptics Conference (WHC)*, pages 37–42. IEEE, 2021. 27

[99] Tao Xu, Wang Dang, Jiabao Wang, and Yun Zhou. Dagam: A domain adversarial graph attention model for subject independent EEG-based emotion recognition. *Journal of Neural Engineering*, 2022. 27

[100] Smith K Khare and U Rajendra Acharya. An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals. *Computers in Biology and Medicine*, 155:106676, 2023.

[101] Taegyun Jeong, Ukeob Park, and Seung Wan Kang. Novel quantitative electroencephalogram feature image adapted for deep learning: Verification through classification of alzheimer's disease dementia. *Frontiers in Neuroscience*, 16:1033379, 2022. 20

[102] Dominik Raab, Andreas Theissler, and Myra Spiliopoulou. XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Neural Computing and Applications*, pages 1–18, 2022. 20, 28

[103] Md Rashed-Al-Mahfuz, Mohammad Ali Moni, Shahadat Uddin, Salem A Alyami, Matthew A Summers, and Valsamma Eapen. A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–12, 2021.

[104] Pankaj Pandey and Krishna Prasad Miyapuram. Nonlinear EEG analysis of mindfulness training using interpretable machine learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3051–3057. IEEE, 2021.

[105] Simone A Ludwig. Explainability using shap for epileptic seizure recognition. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5305–5311. IEEE, 2022. 20

[106] Guanjin Wang, Zhaohong Deng, and Kup-Sze Choi. Detection of epilepsy with electroencephalogram using rule-based classifiers. *Neurocomputing*, 228: 283–290, 2017. 20

[107] Xiashuang Wang, Guanghong Gong, Ni Li, and Shi Qiu. Detection analysis of epileptic EEG using a novel random forest model combined with grid search optimization. *Frontiers in Human Neuroscience*, 13:52, 2019.

[108] Cristian Donos, Matthias Dümpelmann, and Andreas Schulze-Bonhage. Early seizure detection algorithm based on intracranial EEG and random forest classification. *International Journal of Neural Systems*, 25(05):1550023, 2015.

[109] Maouia Bentlemsan, ET-Tahir Zemouri, Djamel Bouchaffra, Bahia Yahya-Zoubir, and Karim 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*, pages 235–238. IEEE, 2014.

[110] Vishal Vijayakumar, Michelle Case, Sina Shirinpour, and Bin He. Quantifying and characterizing tonic thermal pain across subjects from EEG data using random forest models. *IEEE Transactions on Biomedical Engineering*, 64(12):2988–2996, 2017. 20

[111] Hao Feng, Yaxin Peng, Guixu Zhang, and Chaomin Shen. Joint distribution adaptation based tsk fuzzy logic system for epileptic EEG signal identification. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 340–345. IEEE, 2016. 20, 31

[112] Yizhang Jiang, Yuanpeng Zhang, Chuang Lin, Dongrui Wu, and Chin-Teng Lin. EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1752–1764, 2020. 31

[113] Li-Wei Ko, Yi-Chen Lu, Humberto Bustince, Yu-Cheng Chang, Yang Chang, Javier Ferandez, Yu-Kai Wang, Jose Antonio Sanz, Gracaliz Pereira Dimuro, and Chin-Teng Lin. Multimodal fuzzy fusion for enhancing the motor-imagery-based brain computer interface. *IEEE Computational Intelligence Magazine*, 14(1):96–106, 2019.

[114] Zehong Cao and Chin-Teng Lin. Inherent fuzzy entropy for the improvement of EEG complexity evaluation. *IEEE Transactions on Fuzzy Systems*, 26(2): 1032–1035, 2017.

[115] Aysa Jafarifarmand, Mohammad Ali Badamchizadeh, Sohrab Khanmohammadi, Mohammad Ali Nazari, and Behzad Mozaffari Tazehkand. A new self-regulated neuro-fuzzy framework for classification of EEG signals in motor imagery BCI. *IEEE Transactions on Fuzzy Systems*, 26(3):1485–1497, 2017. 20

[116] Dong Qian, Bei Wang, Xiangyun Qing, Tao Zhang, Yu Zhang, Xingyu Wang, and Masatoshi Nakamura. Drowsiness detection by bayesian-copula discriminant classifier based on EEG signals during daytime short nap. *IEEE Transactions on Biomedical Engineering*, 64(4):743–754, 2016. 20, 32

[117] Xingyu Wu, Bingbing Jiang, Kui Yu, and Huanhuan Chen. Separation and recovery markov boundary discovery and its application in EEG-based emotion recognition. *Information Sciences*, 571:262–278, 2021. 32

[118] Weidong Zhou, Yinxia Liu, Qi Yuan, and Xueli Li. Epileptic seizure detection using lacunarity and bayesian linear discriminant analysis in intracranial EEG. *IEEE Transactions on Biomedical Engineering*, 60(12):3375–3381, 2013.

[119] Yu Zhang, Guoxu Zhou, Jing Jin, Qibin Zhao, Xingyu Wang, and Andrzej Cichocki. Sparse bayesian classification of EEG for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2256–2267, 2015.

[120] Haihong Zhang, Huijuan Yang, and Cuntai Guan. Bayesian learning for spatial filtering in an EEG-based brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 24(7):1049–1060, 2013. 20

[121] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *25th International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016. 21, 22

[122] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 21

[123] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE international conference on computer vision*, pages 2921–2929, 2016. 21, 87

[124] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 21, 54, 82

[125] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 21

[126] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 21

[127] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001. 21

[128] Ardalan Aarabi, Reza Fazel-Rezai, and Yahya Aghakhani. A fuzzy rule-based system for epileptic seizure detection in intracranial EEG. *Clinical Neurophysiology*, 120(9):1648–1657, 2009. 21

[129] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012. 21

[130] Yi Wang, Brendan McCane, Neil McNaughton, Zhiyi Huang, Phoebe Neo, et al. Anxietydecoder: an EEG-based anxiety predictor using a 3-d convolutional neural network. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 23

[131] Ce Ju and Cuntai Guan. Tensor-CSPNet: A novel geometric deep learning framework for motor imagery classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 24

[132] Stefan Jonas, Andrea O Rossetti, Mauro Oddo, Simon Jenni, Paolo Favaro, and Frederic Zubler. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40(16):4606–4617, 2019. 26

[133] Zülfikar Aslan and Mehmet Akin. A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals. *Physical and Engineering Sciences in Medicine*, 45(1):83–96, 2022. 26

[134] Enas Abdulhay, Maha Alafeef, Arwa Abdelhay, and Areen Al-Bashir. Classification of normal, ictal and inter-ictal EEG via direct quadrature and random forest tree. *Journal of Medical and Biological Engineering*, 37(6): 843–857, 2017. 29

[135] Adam Li, Chester Huynh, Zachary Fitzgerald, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel O Claudio, Andres M Kanner, Jennifer Hopp, Stephanie Chen, et al. Neural fragility as an EEG marker of the seizure onset zone. *Nature Neuroscience*, 24(10):1465–1474, 2021. 29

[136] Chamandeep Kaur, Amandeep Bisht, Preeti Singh, and Garima Joshi. EEG signal denoising using hybrid approach of variational mode decomposition and wavelets for depression. *Biomedical Signal Processing and Control*, 65: 102337, 2021. 34

[137] Zitong Wan, Rui Yang, Mengjie Huang, Weibo Liu, and Nianyin Zeng. EEG fading data classification based on improved manifold learning with adaptive neighborhood selection. *Neurocomputing*, 482:186–196, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.11.039. 34

[138] Ramy Hussein, Hamid Palangi, Rabab K Ward, and Z Jane Wang. Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. *Clinical Neurophysiology*, 130(1):25–37, 2019. 34, 35

[139] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Transactions on Cybernetics*, 50(7):3281–3293, 2019. 34, 36

[140] He Wang, Peiyin Chen, Meng Zhang, Jianbo Zhang, Xinlin Sun, Mengyu Li, Xiong Yang, and Zhongke Gao. EEG-based motor imagery recognition framework via multisubject dynamic transfer and iterative self-training. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 34

[141] Zhunan Li, Enwei Zhu, Ming Jin, Cunhang Fan, Huiguang He, Ting Cai, and Jinpeng Li. Dynamic domain adaptation for class-aware cross-subject and cross-session EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(12):5964–5973, 2022. 34, 37

[142] Yuan-Pin Lin. Constructing a personalized cross-day EEG-based emotion-classification model using transfer learning. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1255–1264, 2019. 34, 37

[143] Yuan-Pin Lin. Constructing a personalized cross-day EEG-based emotion-classification model using transfer learning. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1255–1264, 2020. doi: 10.1109/JBHI.2019.2934172. 34

[144] Zhong Yin and Jianhua Zhang. Cross-session classification of mental workload levels using EEG and an adaptive deep learning model. *Biomedical Signal Processing and Control*, 33:30–47, 2017. 34, 37

[145] Nooshin Bahador, Jarno Jokelainen, Seppo Mustola, and Jukka Kortelainen. Reconstruction of missing channel in electroencephalogram using spatiotemporal correlation-based averaging. *Journal of Neural Engineering*, 18(5): 056045, 2021. 34, 38

[146] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 14(1):382–393, 2023. doi: 10.1109/TAFFC.2020.3025777. 34

[147] Hubert Banville, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *NeuroImage*, 251:118994, 2022. 34, 38

[148] Jianye Zhang and Peng Yin. Multivariate time series missing data imputation using recurrent denoising autoencoder. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 760–764, 2019. doi: 10.1109/BIBM47256.2019.8982996.

[149] Heba El-Fiqi, Kathryn Kasmarik, Anastasios Bezerianos, Kay Chen Tan, and Hussein A. Abbass. Gate-layer autoencoders with application to incomplete EEG signal recovery. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi: 10.1109/IJCNN.2019.8852101.

[150] Yao Jia, Chongyu Zhou, and Mehul Motani. Spatio-temporal autoencoder for feature learning in patient data with missing observations. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 886–890, 2017. doi: 10.1109/BIBM.2017.8217773. 34

[151] Ziyi Ni, Jiaming Xu, Yuwei Wu, Mengfan Li, Guizhi Xu, and Bo Xu. Improving cross-state and cross-subject visual erp-based BCI with temporal modeling and adversarial training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:369–379, 2022. 34, 39

[152] Xiao Zhang and Dongrui Wu. On the vulnerability of cnn classifiers in eeg-based bcis. *IEEE transactions on neural systems and rehabilitation engineering*, 27(5):814–825, 2019.

[153] Zihan Liu, Lubin Meng, Xiao Zhang, Weili Fang, and Dongrui Wu. Universal adversarial perturbations for cnn classifiers in eeg-based bcis. *Journal of Neural Engineering*, 18(4):0460a4, 2021.

[154] Boyuan Feng, Yuke Wang, and Yufei Ding. SAGA: sparse adversarial attack on EEG-based brain computer interface. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 975–979. IEEE, 2021. 34, 39

[155] Xiao Zhang, Dongrui Wu, Lieyun Ding, Hanbin Luo, Chin-Teng Lin, Tzyy-Ping Jung, and Ricardo Chavarriaga. Tiny noise, big mistakes: adversarial perturbations induce errors in brain–computer interface spellers. *National Science Review*, 8(4):nwaa233, 2021. 39

[156] Siyuan Zhao, Chenyu Liu, Yi Ding, and Xinliang Zhou. Selectivefinetuning: Enhancing transfer learning in sleep staging through selective domain alignment. *arXiv preprint arXiv:2501.03764*, 2025. 42

[157] Xinliang Zhou, Yuzhe Han, Chenyu Liu, Yi Ding, Ziyu Jia, and Yang Liu. Bit-mamsleep: Bidirectional temporal mamba for eeg sleep staging. *arXiv preprint arXiv:2411.01589*, 2024. 42

[158] Emina Alickovic and Abdulhamit Subasi. Ensemble svm method for automatic sleep stage classification. *IEEE Transactions on Instrumentation and Measurement*, 67(6):1258–1265, 2018. 43

[159] Luay Fraiwan, Khaldon Lweesy, Natheer Khasawneh, Heinrich Wenz, and Hartmut Dickhaus. Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier. *Computer methods and programs in biomedicine*, 108(1):10–19, 2012. 43

[160] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018. 43, 47, 48

[161] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 453–456. IEEE, 2018. 43

[162] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. 43

[163] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25 (11):1998–2008, 2017. 43, 47, 48

[164] Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *International Conference on Learning Representations*, 2023. 43

[165] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 46

[166] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012. 47

[167] Akara Supratak and Yike Guo. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 641–644. IEEE, 2020. 47, 48

[168] Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems*, 32, 2019. 47, 48

[169] Jiuwen Cao, Yuanmeng Feng, Runze Zheng, Xiaonan Cui, Weijie Zhao, Tiejia Jiang, and Feng Gao. Two-stream attention 3-D deep network-based childhood epilepsy syndrome classification. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2022. 53

[170] Yuan Zhang, Yao Guo, Po Yang, Wei Chen, and Benny Lo. Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*, 24(2):465–474, 2019.

[171] Supriya Supriya, Siuly Siuly, Hua Wang, and Yanchun Zhang. Epilepsy detection from EEG using complex network techniques: A review. *IEEE Reviews in Biomedical Engineering*, 16:292–306, 2021. 53

[172] Ashish Gupta, Hari Prabhat Gupta, Bhaskar Biswas, and Tanima Dutta. Approaches and applications of early classification of time series: A review. *IEEE Transactions on Artificial Intelligence*, 1(1):47–61, 2020. 53

[173] Ziyu Jia, Xiyang Cai, Gaoxing Zheng, Jing Wang, and Youfang Lin. SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Transactions on Artificial Intelligence*, 1(3):248–257, 2020. 53

[174] Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang 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*, 30:2126–2136, 2022. 53

[175] Yunzhe Tao, Tao Sun, Aashiq Muhamed, Sahika Genc, Dylan Jackson, Ali Arsanjani, Suri Yaddanapudi, Liang Li, and Prachi Kumar. Gated transformer for decoding human brain EEG signals. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 125–130. IEEE, 2021. 53

[176] Omar El Ariss and Kaoning Hu. Resnet-based parkinson's disease classification. *IEEE Transactions on Artificial Intelligence*, 2022. 53

[177] Tiehua Zhang, Yuze Liu, Zhishu Shen, Rui Xu, Xin Chen, Xiaowei Huang, and Xi Zheng. An adaptive federated relevance framework for spatial temporal graph learning. *IEEE Transactions on Artificial Intelligence*, 2023. 53

[178] Gustavo Assunção, Bruno Patrão, Miguel Castelo-Branco, and Paulo Menezes. An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 3(6):867–886, 2022. 53

[179] Muhammad Tariq Sadiq, Xiaojun Yu, Zhaohui Yuan, Muhammad Zulkifal Aziz, Siuly Siuly, and Weiping Ding. Toward the development of versatile brain–computer interfaces. *IEEE Transactions on Artificial Intelligence*, 2 (4):314–328, 2021. 53

[180] Muhammad Zubair, Maria Vladimirovna Belykh, M. Umesh Kumar Naik, Mohammad Fareeda Madeen Gouher, Shani Vishwakarma, Shaik Rafi Ahamed, and Ramanjaneyulu Kongara. Detection of epileptic seizures from EEG signals by combining dimensionality reduction algorithms with machine learning models. *IEEE Sensors Journal*, 21(15):16861–16869, 2021. doi: 10.1109/JSEN.2021.3077578. 53

[181] Shufang Li, Weidong Zhou, Qi Yuan, and Yinxia Liu. Seizure prediction using spike rate of intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(6):880–886, 2013. 53, 56

[182] Ali Shahidi Zandi, Reza Tafreshi, Manouchehr Javidan, and Guy A Dumont. Predicting epileptic seizures in scalp EEG based on a variational bayesian gaussian mixture model of zero-crossing intervals. *IEEE Transactions on Biomedical Engineering*, 60(5):1401–1413, 2013. 53

[183] Jerome Engel Jr, Anatol Bragin, Richard Staba, and Istvan Mody. High-frequency oscillations: what is normal and what is not? *Epilepsia*, 50(4): 598–604, 2009. 54, 56, 83

[184] Richard J Staba, Charles L Wilson, Anatol Bragin, Itzhak Fried, and Jerome Engel Jr. Quantitative analysis of high-frequency oscillations (80–500 Hz) recorded in human epileptic hippocampus and entorhinal cortex. *Journal of neurophysiology*, 88(4):1743–1752, 2002. 54

[185] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 54, 56

[186] Poulomi Pal and Manjunatha Mahadevappa. Adaptive multi-dimensional dual attentive DCNN for detecting cardiac morbidities using fused ECG-PPG signals. *IEEE Transactions on Artificial Intelligence*, 2022. 54

[187] Xinwu Yang, Jiaqi Zhao, Qi Sun, Jianbo Lu, and Xu Ma. An effective dual self-attention residual network for seizure prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1604–1613, 2021. 54

[188] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with self-attention. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE, 2019. 54, 56

[189] Shuiling Shi and Wenqi Liu. B2-ViT Net: Broad vision transformer network with broad attention for seizure prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 54

[190] Xinqiao Zhao, Hongmiao Zhang, Guilin Zhu, Fengxiang You, Shaolong Kuang, and Lining Sun. A multi-branch 3D convolutional neural network for EEG-based motor imagery classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(10):2164–2177, 2019. 54

[191] John Milton and Peter Jung. *Epilepsy as a dynamic disease*. Springer Science & Business Media, 2003. 54

[192] Donghyun Park, Hoonseok Park, Sangyeon Kim, Sanghyun Choo, Sangwon Lee, Chang S Nam, and Jae-Yoon Jung. Spatio-temporal explanation of 3D-EEGNet for motor imagery EEG classification using permutation and saliency. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 55

[193] MZ Ahmad, Maryam Saeed, Sajid Saleem, and Awais M Kamboh. Seizure detection using EEG: A survey of different techniques. In *2016 International Conference on Emerging Technologies (ICET)*, pages 1–6. IEEE, 2016. 55

[194] Lasitha S Vidyaratne and Khan M Iftekharuddin. Real-time epileptic seizure detection using EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):2146–2156, 2017. 55

[195] A Liu, JS Hahn, GP Heldt, and RW Coen. Detection of neonatal seizures through computerized EEG analysis. *Electroencephalography and clinical neurophysiology*, 82(1):30–37, 1992. 56

[196] Robin Tibor Schirrmeister, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11): 5391–5420, 2017. 56, 73, 75, 77, 91, 112

[197] Alexander J Casson, David C Yates, Shelagh JM Smith, John S Duncan, and Esther Rodriguez-Villegas. Wearable electroencephalography. *IEEE Engineering in Medicine and Biology Magazine*, 29(3):44–56, 2010. 56

[198] Kais Gadhoumi, Jean-Marc Lina, and Jean Gotman. Discriminating preictal and interictal states in patients with temporal lobe epilepsy using wavelet analysis of intracerebral EEG. *Clinical neurophysiology*, 123(10):1906–1916, 2012. 56

[199] Samanwoy Ghosh-Dastidar, Hojjat Adeli, and Nahid Dadmehr. Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. *IEEE Transactions on Biomedical Engineering*, 54(9):1545–1551, 2007. 56

[200] Jonathan S Smith. The local mean decomposition and its application to EEG perception data. *Journal of the Royal Society Interface*, 2(5):443–454, 2005. 56

[201] Tao Zhang and Wanzhong Chen. Lmd based features for the automatic seizure detection of EEG signals using svm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(8):1100–1108, 2016. 56

[202] Roshan Joy Martis, U Rajendra Acharya, Jen Hong Tan, Andrea Petznick, Ratna Yanti, Chua Kuang Chua, EY Kwee Ng, and Louis Tong. Application of empirical mode decomposition (EMD) for automated detection of epilepsy using EEG signals. *International journal of neural systems*, 22(06):1250027, 2012. 56

[203] Han-Tai Shiao, Vladimir Cherkassky, Jieun Lee, Brandon Veber, Edward E Patterson, Benjamin H Brinkmann, and Gregory A Worrell. SVM-based system for prediction of epileptic seizures from iEEG signal. *IEEE Transactions on Biomedical Engineering*, 64(5):1011–1022, 2016. 56

[204] Jing Jin, Xingyu Wang, and Bei Wang. Classification of direction perception EEG based on pca-svm. In *Third International Conference on Natural Computation (ICNC)*, volume 2, pages 116–120. IEEE, 2007. 56

[205] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2909–2917, 2020. 56

[206] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019. 57, 65

[207] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 57, 82, 83

[208] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 58, 67

[209] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 58, 67

[210] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019. 58, 67

[211] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*, 2020. 59, 69

[212] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017. 59, 69

[213] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021. 61, 72

[214] Khaled Saab, Jared Dunnmon, Christopher Ré, Daniel Rubin, and Christopher Lee-Messer. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *NPJ digital medicine*, 3(1):59, 2020. 61, 63, 73, 77

[215] David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Lars Petersson, Matthew J Aburn, and Clinton Fookes. Neural memory networks for seizure type classification. In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 569–575. IEEE, 2020. 62, 63, 73, 75, 77

[216] Xuyang Zhao, Noboru Yoshida, Tetsuya Ueda, Hidenori Sugano, and Toshihisa Tanaka. Epileptic seizure detection by using interpretable machine learning models. *Journal of Neural Engineering*, 20(1):015002, 2023. 71

[217] Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, and Garth Slaney. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098, 2021. 71

[218] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The uci machine learning repository. `https://archive.ics.uci.edu`. 72

[219] Anubha Gupta, Pushpendra Singh, and Mandar Karlekar. A novel signal modeling approach for classification of seizure and seizure-free EEG signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26 (5):925–935, 2018. 72

[220] Siddharth Panwar, Shiv Dutt Joshi, Anubha Gupta, and Puneet Agarwal. Automated epilepsy diagnosis using EEG with test set evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1106–1116, 2019. 72

[221] Palak Handa, Monika Mathur, and Nidhi Goel. Open and free EEG datasets for epilepsy diagnosis. *arXiv preprint arXiv:2108.01030*, 2021. 73

[222] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006. 73, 77

[223] Yang Li, Yu Liu, Wei-Gang Cui, Yu-Zhu Guo, Hui Huang, and Zhong-Yi Hu. Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(4):782–794, 2020. 73, 77

[224] Chetan Ralekar, Shubham Choudhary, Tapan Kumar Gandhi, and Santanu Chaudhury. Development of character recognition model inspired by visual explanations. *IEEE Transactions on Artificial Intelligence*, 2023. 82

[225] Alexandros T Tzallas, Markos G Tsipouras, and Dimitrios I Fotiadis. Epileptic seizure detection in EEGs using time–frequency analysis. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):703–710, 2009. 82

[226] Luciana Patrízia Alves Andrade-Valença, François Dubeau, Francesco Mari, Rina Zelmann, and Jean Gotman. Interictal scalp fast oscillations as a marker of the seizure onset zone. *Neurology*, 77:524 – 531, 2011. URL `https://api.semanticscholar.org/CorpusID:17978000`. 83

[227] European Commission. Road safety thematic report – fatigue. Technical report, European Road Safety Observatory, 2021. URL `https://road-safety.transport.ec.europa.eu/system/files/2021-07/road_safety_thematic_report_fatigue_tc_final.pdf`. 85

[228] Chin-Shun Hsieh and Cheng-Chi Tai. An improved and portable eye-blink duration detection system to warn of driver fatigue. *Instrumentation Science & Technology*, 41(5):429–444, 2013. 85

[229] Mitesh Patel, Sara KL Lal, Diarmuid Kavanagh, and Peter Rossiter. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert systems with Applications*, 38(6):7235–7242, 2011. 85

[230] Shigeyuki Tateno, Xia Guan, Rui Cao, and Zhaoxian Qu. Development of drowsiness detection system based on respiration changes using heart rate monitoring. In *2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 1664–1669. IEEE, 2018. 85

[231] Zhongmin Liu, Yuxi Peng, and Wenjin Hu. Driver fatigue detection based on deeply-learned facial expression representation. *Journal of Visual Communication and Image Representation*, 71:102723, 2020. 85

[232] Yong Peng, Qian Xu, Shuxiang Lin, Xinghua Wang, Guoliang Xiang, Shufang Huang, Honghao Zhang, and Chaojie Fan. The application of electroencephalogram in driving safety: current status and future prospects. *Frontiers in psychology*, 13, 2022. 85

[233] Jian Cui, Zirui Lan, Yisi Liu, Ruilin Li, Fan Li, Olga Sourina, and Wolfgang Müller-Wittig. A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods*, 202:173–184, 2022. 85

[234] Rifai Chai, Ganesh R Naik, Tuan Nghia Nguyen, Sai Ho Ling, Yvonne Tran, Ashley Craig, and Hung T Nguyen. Driver fatigue classification with independent component by entropy rate bound minimization analysis in an eeg-based system. *IEEE journal of biomedical and health informatics*, 21(3): 715–724, 2016. 86

[235] Carolyn Schwendeman, Ryan Kaveh, and Rikky Muller. Drowsiness detection with wireless, user-generic, dry electrode ear eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 9–12. IEEE, 2022. 86

[236] Prabhavathi C Nissimagoudar, Anilkumar V Nandi, and HM Gireesha. Deep convolution neural network-based feature learning model for eeg based driver alert/drowsy state detection. In *International Conference on Soft Computing and Pattern Recognition*, pages 287–296. Springer, 2019. 86

[237] Sirui Ding, Zhiyong Yuan, Panfeng An, Guotong Xue, Wenxiang Sun, and Jianhui Zhao. Cascaded convolutional neural network with attention mechanism for mobile eeg-based driver drowsiness detection system. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1457–1464. IEEE, 2019. 86

[238] Zaid Abdi Alkareem Alyasseri, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Osama Ahmad Alomari. Person identification using EEG channel selection with hybrid flower pollination algorithm. *Pattern Recognition*, 105:107393, 2020. 87

[239] Hesam Varsehi and S Mohammad P Firoozabadi. An EEG channel selection method for motor imagery based brain–computer interface and neurofeedback using granger causality. *Neural Networks*, 133:193–206, 2021. 87

[240] Liying Yang, Si Chao, Qingyang Zhang, Pei Ni, and Dunhui Liu. A grouped dynamic eeg channel selection method for emotion recognition. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3689–3696. IEEE, 2021. 87

[241] Han Zhang, Xing Zhao, Zexu Wu, Biao Sun, and Ting Li. Motor imagery recognition with automatic eeg channel selection and deep learning. *Journal of neural engineering*, 18(1):016004, 2021. 87

[242] Pramod Gaur, Karl McCreadie, R. B. Pachori, Hui Wang, and Girijesh Prasad. An automatic subject specific channel selection method for enhancing motor imagery classification in EEG-BCI using correlation. *Biomedical Signal Processing and Control*, 68:102574, 2021. 87

[243] Chun-Shu Wei, Yu-Te Wang, Chin-Teng Lin, and Tzyy-Ping Jung. Toward drowsiness detection using non-hair-bearing eeg-based brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 26(2):400–406, 2018. 90

[244] Wenbo Li, Guanzhong Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, Dongpu Cao, and Fei-Yue Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Transactions on Computational Social Systems*, 9(3):667–678, 2021. 95

[245] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios. *IEEE Transactions on Affective Computing*, 2021.

[246] Yekta Said CAN and Cem ERSOYSenior Member. Smart affect monitoring with wearables in the wild: An unobtrusive mood-aware emotion recognition system. *IEEE Transactions on Affective Computing*, 2022. 95

[247] Haiyun Huang, Qiuyou Xie, Jiahui Pan, Yanbin He, Zhenfu Wen, Ronghao Yu, and Yuanqing Li. An EEG-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness. *IEEE Transactions on Affective Computing*, 12(4):832–842, 2019. 95

[248] Mario Ezra Aragon, Adrian Pastor Lopez-Monroy, Luis-Carlos Gonzalez Gonzalez-Gurrola, and Manuel Montes. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Transactions on Affective Computing*, 2021.

[249] Sarah Sarabadani, Larissa C Schudlo, Ali Akbar Samadani, and Azadeh Kushski. Physiological detection of affective states in children with autism spectrum disorder. *IEEE Transactions on Affective Computing*, 11(4):588–600, 2018. 95

[250] Jean Costa, François Guimbretière, Malte F Jung, and Tanzeem Choudhury. Boostmeup: Improving cognitive performance in the moment by unobtrusively regulating emotions with a smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–23, 2019. 95

[251] David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. Technical design space analysis for unobtrusive driver emotion assessment using multi-domain context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–30, 2023. 95

[252] M Prajwal, Ayush Raj, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. Towards efficient emotion self-report collection using human-ai collaboration: A case study on smartphone keyboard interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):1–23, 2023. 95

[253] Albert C Cruz, Bir Bhanu, and Ninad S Thakoor. Vision and attention theory based sampling for continuous facial emotion recognition. *IEEE Transactions on Affective Computing*, 5(4):418–431, 2014. 95

[254] E Pranav, Suraj Kamal, C Satheesh Chandran, and MH Supriya. Facial emotion recognition using deep convolutional neural network. In *2020 6th International conference on advanced computing and communication Systems (ICACCS)*, pages 317–320. IEEE, 2020. 98

[255] Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1906–1916, 2022.

[256] Tsung-Ren Huang, Shin-Min Hsu, and Li-Chen Fu. Data augmentation via face morphing for recognizing intensities of facial emotions. *IEEE Transactions on Affective Computing*, 2021. 95

[257] Jui-Feng Yeh, Jian-Cheng Tsai, Bo-Wei Wu, and Tai-You Kuang. Deep learning-based emotion spatial regression in speech recognition for human-computer interaction. In *2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII)*, pages 549–552. IEEE, 2019. 95, 98

[258] Xie Ying and Zhang Yizhe. Design of speech emotion recognition algorithm based on deep learning. In *2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, pages 734–737. IEEE, 2021. 98

[259] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015. 95

[260] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2): 505–523, 2018. 95

[261] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 2021. 95

[262] Yuzhe Zhang, Huan Liu, Dalin Zhang, Xuxu Chen, Tao Qin, and Qinghua Zheng. EEG-based emotion recognition with emotion localization via hierarchical self-attention. *IEEE Transactions on Affective Computing*, 2022. 96

[263] Yang Li, Wenming Zheng, Zhen Cui, Tong Zhang, and Yuan Zong. A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1561–1567, 2018. 99

[264] Huayu Chen, Shuting Sun, Jianxiu Li, Ruilan Yu, Nan Li, Xiaowei Li, and Bin Hu. Personal-zscore: Eliminating individual difference for EEG-based cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 2021.

[265] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018. 99, 101, 114, 118

[266] Gopal Chandra Jana, Anshuman Sabath, and Anupam Agrawal. Capsule neural networks on spatio-temporal eeg frames for cross-subject emotion recognition. *Biomedical Signal Processing and Control*, 72:103361, 2022. 99

[267] Rui Li, Chao Ren, Chen Li, Nan Zhao, Dawei Lu, and Xiaowei Zhang. Sstd: a novel spatio-temporal demographic network for eeg-based emotion recognition. *IEEE Transactions on Computational Social Systems*, 10(1):376–387, 2022. 96, 99

[268] Longbin Jin. Emotion recognition based bci using channel-wise features. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–6, 2020. 96

[269] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte. Towards robust neuroadaptive HCI: exploring modern machine learning methods to estimate mental workload from EEG signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018. 96

[270] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. The empathetic car: Exploring emotion inference via driver behaviour and traffic context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–34, 2021. 96

[271] Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020. 96, 101, 114, 118

[272] Tim Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5(7): 583–589, 2004. 96

[273] Paul Smolensky. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987. 97, 101

[274] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019. 97

[275] Zhiyuan Li, Shizhong Han, Ahmed Shehab Khan, Jie Cai, Zibo Meng, James O'Reilly, and Yan Tong. Pooling map adaptation in convolutional neural network for facial expression recognition. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 1108–1113. IEEE, 2019. 98

[276] Stefano Piana, Alessandra Staglianò, Francesca Odone, and Antonio Camurri. Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(1):1–31, 2016. 98

[277] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*, 2019. 98

[278] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 98

[279] Chenyu Liu, Xinliang Zhou, Jiaping Xiao, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vsgt: variational spatial and gaussian temporal graph models for eeg-based emotion recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3078–3086, 2024.

[280] Chenyu Liu, Xinliang Zhou, Yihao Wu, Ruizhi Yang, Zhongruo Wang, Liming Zhai, Ziyu Jia, and Yang Liu. Graph neural networks in eeg-based emotion recognition: a survey. *arXiv preprint arXiv:2402.01138*, 2024. 98

[281] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen. Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 99

[282] Yi Wang, Zhiyi Huang, Brendan McCane, and Phoebe Neo. Emotionet: A 3-d convolutional neural network for eeg-based emotion recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 99

[283] Ehsan Lotfi and M-R Akbarzadeh-T. Practical emotional neural networks. *Neural Networks*, 59:61–72, 2014. 99

[284] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. Generalization and personalization of mobile sensing-based mood inference models: An analysis of college students in eight countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–32, 2023. 100

[285] Kunal Gupta, Sam WT Chan, Yun Suen Pai, Nicholas Strachan, John Su, Alexander Sumich, Suranga Nanayakkara, and Mark Billinghurst. Total vrecall: Using biosignals to recognize emotional autobiographical memory in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–21, 2022.

[286] Mintra Ruensuk, Eunyong Cheon, Hwajung Hong, and Ian Oakley. How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–32, 2020. 100

[287] Matthew G Perich and Kanaka Rajan. Rethinking brain-wide interactions through multi-region 'network of networks' models. *Current opinion in neurobiology*, 65:146–151, 2020. 101

[288] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, pages 1–9, 2023. 101

[289] Patricia A Alexander. Relational thinking and relational reasoning: harnessing the power of patterning. *NPJ science of learning*, 1(1):1–7, 2016. 101

[290] Min Wang, Heba El-Fiqi, Jiankun Hu, and Hussein A Abbass. Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states. *IEEE Transactions on Information Forensics and Security*, 14(12):3259–3272, 2019. 101

[291] Peter T Fox, Marcus E Raichle, Mark A Mintun, and Carmen Dence. Nonoxidative glucose consumption during focal physiologic neural activity. *Science*, 241(4864):462–464, 1988. 104

[292] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in Neural Information Processing Systems*, 28, 2015. 105

[293] Oscar B Sheynin. Laplace's theory of errors. *Archive for history of exact sciences*, 17(1):1–61, 1977. 105

[294] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ArXiv Preprint arXiv:1312.6114*, 2013. 106

[295] Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2022. 106

[296] Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pages 4531–4541. PMLR, 2020. 107

[297] Yi Ding, Neethu Robinson, Qiuhao Zeng, Duo Chen, Aung Aung Phyo Wai, Tih-Shih Lee, and Cuntai Guan. Tsception: a deep learning framework for emotion detection using eeg. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 114, 115

[298] Clarence Tan, Marko Šarlija, and Nikola Kasabov. Neurosense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal eeg patterns. *Neurocomputing*, 434:137–148, 2021. 114, 115

[299] Shuaiqi Liu, Xu Wang, Ling Zhao, Bing Li, Weiming Hu, Jie Yu, and Yu-Dong Zhang. 3dcann: A spatio-temporal convolution attention neural network for eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5321–5331, 2021. 114, 116

[300] Cheng Cheng, Yong Zhang, Luyao Liu, Wenzhe Liu, and Lin Feng. Multi-domain encoding of spatiotemporal dynamics in eeg for emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 27(3):1342–1353, 2022. 114, 115

[301] Shadi Sartipi, Mastaneh Torkamani-Azar, and Mujdat Cetin. A hybrid end-to-end spatio-temporal attention neural network with graph-smooth signals for eeg emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. 114, 116

[302] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020. 114, 118

[303] Xiaoxu Li, Wenming Zheng, Yuan Zong, Hongli Chang, and Cheng Lu. Attention-based spatio-temporal graphic lstm for EEG emotion recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 114, 118

[304] Wei-Bang Jiang, Xu Yan, Wei-Long Zheng, and Bao-Liang Lu. Elastic graph transformer networks for EEG-based emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 114, 118

[305] Jungwon Min, Kaoru Nashiro, Hyun Joo Yoo, Christine Cho, Padideh Nasseri, Shelby L Bachman, Shai Porat, Julian F Thayer, Catie Chang, Tae-Ho Lee, et al. Emotion downregulation targets interoceptive brain regions while emotion upregulation targets other affective brain regions. *Journal of Neuroscience*, 42(14):2973–2985, 2022. 118, 119

[306] Ethan Weinberger, Joseph Janizek, and Su-In Lee. Learning deep attribution priors based on prior knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14034–14045, 2020. 125

[307] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8827–8836, 2018. 125

[308] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 125

[309] Sherif M Abdelfattah, Ghodai M Abdelrahman, and Min Wang. Augmenting the size of EEG datasets using generative adversarial networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018. 126

[310] Fatemeh Fahimi, Strahinja Dosen, Kai Keng Ang, Natalie Mrachacz-Kersting, and Cuntai Guan. Generative adversarial networks-based data augmentation for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):4039–4051, 2020. 126