Novel Deep Learning Techniques for the Detection and Classification of Neurodegenerative Diseases using Resting State Electroencephalography

Christopher Almeida Neves

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

December 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Christopher Almeida Neves**

Entitled: **Novel Deep Learning Techniques for the Detection and Classification of Neurodegenerative Diseases using Resting State Electroencephalography**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Thomas Fevens

_____ Examiner
Dr. Sabine Bergler

_____ Examiner
Dr. Thomas Fevens

_____ Co-supervisor
Dr. Yiming Xiao

_____ Co-supervisor
Dr. Yong Zeng

Approved by _____
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2024

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Novel Deep Learning Techniques for the Detection and Classification of Neurodegenerative Diseases using Resting State Electroencephalography

Christopher Almeida Neves

Neurodegenerative diseases are debilitating conditions that progressively deteriorate the life quality of those affected. Compared with traditional neuroimaging modalities, such as Magnetic Resonance Imaging, Electroencephalography (EEG) can provide a more cost-effective and accessible alternative to help underprivileged populations obtain an early diagnosis of their condition, which is paramount for effective patient care. Resting-state EEG (rs-EEG), which records signals while a subject is at rest, offers an alternative to the commonly used task-based experiments for easier-to-adopt data acquisition protocols. While deep learning techniques have been shown to be effective for automatically classifying most EEG signals, they struggle with modeling the long-range temporal dependencies, complex spatial relationships, and the lack of time-locked events in rs-EEG. Aiming to address these issues, we first propose an explainable Graph Neural Network technique for rs-EEG-based Parkinson's disease detection. Our method uses structured global convolutions to model long-range dependencies and novel multi-head graph structure learning to capture the complex spatial relationships in EEG data. We also propose a head-wise gradient-weighted graph attention explainer to obtain rich connectivity insights. Our second major contribution leverages recent innovations in state space modeling techniques to classify individuals with dementia, and we explore spectral and spatial approaches for learning relationships between EEG channels for the designated task. Additionally, we probe our model's outputs with explainability techniques and demonstrate that our model learns physiologically relevant features. This thesis puts forth novel deep-learning methods that show promise in addressing challenges in neurodegenerative disease classification using rs-EEG.

Acknowledgments

I would like to thank my supervisors, Dr. Yong Zeng and Dr. Yiming Xiao, as this thesis and research would not have been possible without their guidance. Their flexibility, support, and unending patience throughout this process have been immeasurable, and I greatly appreciate the mentorship and the intellectual liberties that they afforded me. To say that I have learned a lot from them would be short-selling it. I would also like to thank all of my colleagues at both the Design Lab and Healthy-X Lab for the conversations, encouragement, and support that were indispensable throughout this process.

This journey has been the most challenging yet rewarding experience I have had the opportunity to undertake in my life so far. To my family, friends, and partner, words cannot thank you enough for your patience and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What are Neurodegenerative Diseases?

Neurodegenerative diseases are a group of conditions that progressively deteriorate or damage parts of the nervous system. The most widespread types of neurodegenerative diseases include Alzheimer's disease (AD), multiple sclerosis, amyotrophic lateral sclerosis (ALS), and Parkinson's disease (PD), with Parkinson's disease and Alzheimer's disease being the two leading disorders, according to the National Institute of Environmental Health Sciences [94]. Neurodegenerative diseases are debilitating and often life-threatening, and an early diagnosis is imperative for effective treatment. Parkinson's and Alzheimer's disease are usually diagnosed based on assessments of symptom manifestations, including self-evaluation, cognitive tests, and behavioral evaluations. As these diagnostic tools can be subject to inconsistency due to variations in the delivery of such tests by the evaluators, diagnostic accuracy can benefit from more objective evidence, such as blood samples and neuroimaging data. Therefore, this thesis will study the potential of Electroencephalography (EEG) as a diagnostic tool for Parkinson's and Alzheimer's disease, the most prevalent neurodegenerative conditions.

Structural and functional Magnetic Resonance Imaging (sMRI and fMRI) is often used to study PD and AD with the goal of discovering reliable structural and functional imaging-based biomarkers within the brain. Such biomarkers can help better understand the mechanisms and progression of the diseases. These imaging modalities boast strong spatial resolution, but suffer in terms of temporal

resolution with prohibitively high costs to use, thus limiting their availability in less privileged nations. Their lack of accessibility also means that identifying potential disease biomarkers early is challenging for many, limiting the opportunity for rapid interventions that can significantly affect patient outcomes. A strong alternative to these modalities is electroencephalography, which also captures functional changes in the brain, but with a more portable and cost-effective device.

1.2 Neurodegenerative Disease Functional Neuroimaging Using EEG

EEG is a non-invasive functional neuroimaging technique that measures the electric potentials on the scalp surface arising from the synchronized post-synaptic activity of neuron groups within the cortex. EEG imaging setups vary depending on the applications and can include portable brain-computer interface devices (BCI) embedded in everyday consumer electronics, electrode caps used in clinical settings, and even intracranial EEG, where electrodes are surgically inserted into the skull. EEG data is recorded at very high sampling rates, meaning that massive amounts of data are produced for relatively short recording times. This high recording frequency means that EEG has a temporal resolution far exceeding fMRI. Still, the relatively small magnitude of the measured electrical signals means that EEG lacks fine-grained spatial resolution, is susceptible to environmental interference, and struggles to locate signals from sources deep in the brain. Nevertheless, studies have shown that EEG is a powerful imaging modality, especially in resource-limited regions. Notably, one key application of EEG lies in Epilepsy detection, where temporal resolution is crucial in dynamic brain imaging.

For individuals who are severely impaired by their neurodegenerative diseases, performing task-based experiments in functional neuroimage acquisition can be inconvenient, overly demanding, and complex. This makes many of the experimental paradigms common in EEG trials difficult and sometimes impossible to administer, particularly for the senior population. Resting-state EEG (rs-EEG) offers an alternative recording protocol and is collected while a participant is resting and not performing any specific actions. However, this increased accessibility and ease of recording comes at a cost. The lack of clear signal responses to real-time events makes interpreting rs-EEG more challenging. Thus, techniques that can accurately extract and use salient biomarkers from rs-EEG

data for pathology classification are sorely needed. To that end, this thesis will focus primarily on EEG recorded during resting state experiments.

1.3 Deep Learning: A Boon for EEG Classification

The large quantity of data produced by EEG makes manual annotation and interpretation tedious and time-consuming. Furthermore, extracting relevant features from the data requires significant amounts of domain expertise. Performing this analysis for a large quantity of subjects becomes infeasible without automation. Recently, the rise of machine learning (ML) has made these processing pipelines substantially more efficient. However, many machine learning approaches still rely on hand-crafted features extracted by domain experts. EEG is also severely susceptible to artifacts and requires substantial preprocessing before features can be extracted. This preprocessing step is not only time-consuming, but has been shown to influence the statistical significance of downstream analyses depending on which preprocessing implementation is used [24]. In addition, manually engineered features introduce a bias to the task and can supersede more predictive qualities of the underlying signals in favor of more commonly studied features [64]. We experience this phenomenon in Chapter 4, where frequencies often discarded in traditional preprocessing pipelines turn out as the most predictive features when performing dementia classification.

The advent of powerful deep learning (DL) techniques has opened the door for automated feature learning and has been applied to EEG tasks with great levels of success. They have facilitated the analysis of vast amounts of EEG data and have been imperative in enabling the use of BCI devices. By automatically learning which portions of the input signals are relevant to the task, they can learn to ignore artifacts, and perform well on minimally preprocessed data, removing the need for elaborate preprocessing pipelines [61][124]. Interestingly, post-hoc analysis of trained deep learning models shows that many of the learned features are physiologically relevant, and some architectures can learn processing steps that mimic signal processing filters used in more traditional analyses [29]. Although powerful, deep learning techniques still face many challenges when applied to EEG, particularly rs-EEG, which will be described in the following section.

1.4 Challenges of Deep Learning Applied to Resting State EEG

Deep learning techniques struggle to model EEG signals due to long sequence lengths, non-stationarity, and low signal-to-noise ratio. The majority of studies using deep learning on time-series data use benchmarks with input sequences that are substantially shorter than those seen in EEG. Successful deep learning models must not only extract important temporal features from input data measuring in the thousands of timesteps, but must also model complex spatial relationships between EEG electrodes. Architectures developed specifically for sequential modeling, such as LSTMs and Transformers, still struggle with raw EEG signals for reasons we elaborate on in Chapter 2, and the task becomes more difficult when using data collected from resting-state paradigm experiments. Resting-state EEG is more prone to random signal fluctuations and lacks the time-locked events recorded using experimental paradigms, which means that models need to exploit latent temporal and spatial patterns more effectively. Many of the state-of-the-art deep learning models for EEG are developed for event-triggered tasks, including BCI, sleep stage classification, and epilepsy, all exhibiting clearer temporal patterns than those found in rs-EEG.

Another important issue to consider is the relatively small size of typical EEG datasets compared to other imaging modalities, and the strong inter-subject and inter-site variability present in the data. Subject-specific physiological and affective differences can introduce strong variability into recorded signals. In addition, large variations may be introduced in the signals by different recording locations and times. This necessitates rigorous validation schemes to ensure the robustness of deep learning methods, and novel sequential and spatial modeling techniques to handle EEG signals effectively.

1.5 Thesis Contributions

In this thesis, we begin by proposing a novel Graph Neural Network technique that emphasizes learning the spatial relationships between EEG electrodes for Parkinson's disease detection from resting state EEG data. Our work also proposes a novel method of learning the latent graph structure connecting the electrodes with respect to disease detection. In our second work, we shift our

attention to dementia classification. In particular, we emphasize leveraging state-of-the-art techniques in sequence modeling to extract long-range temporal features from a subject's resting state EEG signals. More specifically, we exploit recent advances in state-space models to classify participants' stage of cognitive decline while exploring the modeling of relationships between EEG electrodes in the spectral domain. Our contributions aim to advance deep learning techniques for the two most common neurodegenerative diseases while simultaneously extracting relevant physiological insights from model outputs.

Our first work presents the following major contributions in Chapter 3:

- We combine structured global convolutions [73] and self-supervised contrastive learning to better model long sequences of EEG data with a limited dataset for the first time.
- We propose a novel dynamic multi-head graph structure learning technique to model relationships between EEG electrodes without making any assumptions about underlying connectivity in contrast to conventional Graph Neural Network methods.
- To enhance the interpretability of our model, we introduce a new technique based on head-wise gradient-weighted attention scores to generate more informative explanations in contrast to more common attention score aggregation techniques.

Our subsequent work offers the following major contributions in Chapter 4:

- We develop a novel Mamba-based architecture to address the need for long-range sequential modeling techniques in rs-EEG signal classification.
- Our method uses a channel-independent modeling approach to extract robust features from the underlying data, and we explore spectral and spatial approaches for learning relationships between channels.
- We are the first to benchmark a Mamba-based architecture on the first large-scale dementia rs-EEG dataset, recently released by [61], and show improved classification performance over previous benchmarks while using substantially fewer DL model parameters.
- We probe our model's outputs to extract clinically relevant insights from the data and show that our model can learn physiologically relevant features.

1.6 Thesis Outline

This thesis begins with a review of important deep learning models used for EEG classification, emphasizing deep sequential modeling techniques. We present each method's strengths and where they fall short for modeling rs-EEG signals. We also give a brief review of unsupervised learning and domain adaptation techniques that have been used to address inter-subject variability. In Chapter 3, we present a novel Graph Neural Network model that jointly classifies subjects with Parkinson's disease while learning important relationships between electrodes. Additionally, we employ a contrastive learning strategy adapted for EEG signals to extract robust features from a limited dataset. In Chapter 4, we introduce a Mamba-based modeling technique that efficiently models long rs-EEG signals while using fewer parameters than baseline models, and we study the physiological significance of the features learned by our model. Finally, in Chapter 5, we conclude the thesis, describe the limitations of our work, and elaborate on promising future avenues of study.

Chapter 2

Background

This chapter begins by introducing EEG as a functional neuroimaging modality. We then provide an overview of traditional machine learning methods applied to EEG followed by a summary of the most important deep learning techniques used to classify EEG signals and describe their strengths and weaknesses. The chapter ends with a brief review of unsupervised learning techniques and domain adaptation methods used to address subject differences in EEG.

2.1 What is EEG?

As previously touched upon in Chapter 1, EEG records the electrical potential emanating from the synchronized activity of cortical neurons. More specifically, it is the postsynaptic activities of pyramidal neurons in the cortical portions of the brain that constitute most of the recorded signal. This is due to their perpendicular alignment and proximity to the cortical surface. Neurons located in deeper parts of the brain are too far from the surface of the head and lack the alignment required to project signals to the scalp, and thus cannot be accurately recorded using EEG devices.

EEG signals record activity from multiple cognitive processes occurring simultaneously. It is common in EEG research to sub-divide the frequency spectrum of the signals into five major sub-bands, and many argue that each sub-band has a unique signature across the scalp surface and reflects different affective or cognitive states. There are no universally agreed-upon start and end values for each of the frequency bands, with many papers reporting delineations that vary by

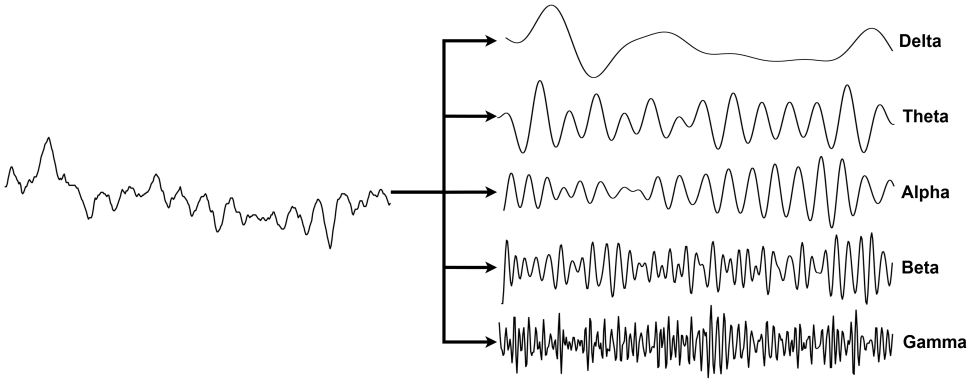

Figure 2.1: Raw EEG signal decomposed into canonical frequency sub-bands using a pass-band filter.

a few Hertz. However, the general spectrum range remains similar between studies. The most commonly agreed upon frequency sub-bands include the Delta (0.5-4 Hz), Theta (4-8 Hz), Alpha (8-13 Hz), Beta (13-30 Hz), and Gamma (30-90 Hz) ranges and are shown in Figure 2.1. Different psychiatric disorders often show abnormalities localized in one or more of the frequency bands, and the power of the band-wise signal is one of the most common biomarkers that is studied in EEG. In Chapter 4, we show that our model learns salient patterns in the Theta, Beta and Gamma bands while classifying individuals with dementia, and many of these learned features are echoed in clinical literature.

EEG devices typically consist of an array of electrodes placed on a subject's head, an amplification unit that amplifies the low-voltage EEG signals, and an output device. Electrode caps, often made out of mesh, are typically used to help arrange the electrodes at specific positions over an individual's scalp and help hold them in place. Setups vary greatly (see Figure 2.2), and the number of electrodes used can vary from 4 to 256 and typically depend on the phenomena being studied. These electrodes are placed over the scalp following standardized systems, with the most common being the 10-20 System, which splits the scalp area into segments of 10 to 20% of the total distance of the skull and defines the relative positions of each electrode. The sampling rates of EEG acquisition devices also vary depending on the task but can extend to thousands of hertz. This high sampling rate, coupled with the number of electrodes used in a study, means that EEG devices produce large amounts of data every second and can be used to image the brain at its native temporal resolution. This sets EEG apart from other functional imaging modalities, such as fMRI and PET,

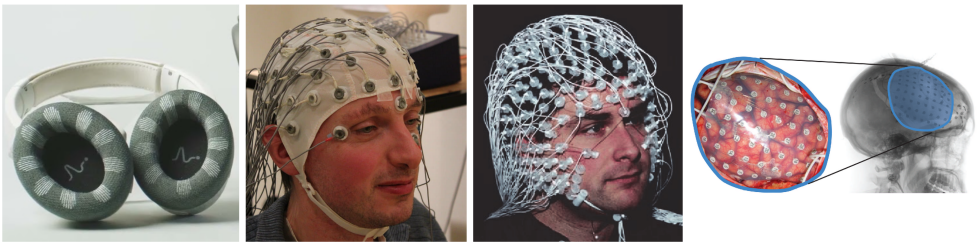

Figure 2.2: EEG recording devices can take many forms. From left to right: Wearable BCI headphones (Neurable), standard EEG setup with mesh cap to hold electrodes in place, high-density EEG electrode array, intra-cranial EEG electrode array placed over the surface of the brain (Electrocorticography).

which boast high spatial resolutions but temporal resolutions that pale in comparison to EEG. However, some studies aim to perform multi-modal imaging of brain activity by pairing EEG with fMRI, which can help optimally couple brain structure and dynamics. Magnetoencephalography (MEG) is one of the only other functional imaging modalities that can rival the temporal resolution of EEG while achieving a better spatial resolution. Whereas EEG measures the electrical activity of neurons, MEG records the magnetic fields that arise from the electrical currents in the brain. However, MEG requires more specialized equipment to shield interference from external magnetic fields and is much more expensive than EEG.

To summarize, the accessibility and cost-effectiveness of EEG means that it is uniquely positioned to offer functional neuroimaging to underserved populations. Although many studies exist linking EEG biomarkers to different pathologies, the low spatial resolution makes it difficult to study more abstract cognitive processes or phenomena that occur deep within the brain. However, as automated detection techniques become more sophisticated, EEG can provide an opportunity for rapid and accurate diagnosis of many disorders. In the next section, we will briefly summarize some of the most common EEG features used with traditional Machine Learning techniques.

2.2 Machine Learning for EEG

Many traditional non-deep learning machine learning (ML) algorithms have been effectively applied to classification tasks for neurodegenerative disorders using EEG. The most popular methods include Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and Random Forests (RF),

and are almost always used in conjunction with hand-crafted feature extraction techniques. There exists a vast number of features that have been proposed and used in conjunction with machine learning algorithms, with the most commonly used features largely falling into three main categories: time-domain features, frequency-domain features, time-frequency domain features [27][120].

Time-domain features express signal characteristics with reference to their variation in time. These features are important when observing time-locked events, such as epileptic seizures and motor imagery tasks. Common time domain features include complexity measures, such as the Higuchi Fractal Dimension, Hurst Exponent and Katz's fractal dimension, which is a measure of self-similarity and quantifies the predictability and regularity of the signals. These non-linear complexity measures have been used along with machine learning algorithms to classify individuals with Alzheimer's disease [2]. Mean or absolute signal value, zero crossings, and slope sign changes are other common time-domain measures that, while simple, have shown success when coupled with machine learning [27].

Frequency-domain features characterize the distribution of a signal across a spectrum of frequencies. Analyzing a signal in the frequency domain makes it easier to observe periodic components of the signal and one of the most common frequency-domain features is the power spectral density (PSD), which describes the distribution of the underlying signal's power over a given frequency range. The relative power of each canonical frequency band is also often used with machine learning methods and has been shown to be discriminative for detecting certain neurodegenerative conditions [62].

Time- and frequency-domain features alone only capture a fraction of the behavior of the underlying signal. Time-domain features are susceptible to artifacts, and frequency-domain features ignore transient events crucial for many classification tasks. Time-frequency domain features jointly extract time and spectral information to generate features that describe how frequency components of the data change with respect to time. These features are some of the most commonly used with machine learning techniques and include the Short-Time Fourier Transform (STFT), the Wavelet transform, and the Filter Bank Common Spatial Pattern (FBCSP).

When the optimal choice of features for a task is known, manual feature engineering and simple machine learning algorithms are powerful choices for effective classifiers. Unfortunately, this is

often not the case and many times, features that are discriminative for one task may be spurious for another. This necessary domain expertise, along with the heavy preprocessing and time investment required to extract many of these manual features, makes the automated representation learning aspects of deep learning a much more attractive choice. In the following section, we will provide an overview of important deep-learning architectures for extracting salient features from EEG data.

2.3 EEG Representation Learning for Neurodegenerative Disease Classification

Extracting manual features for EEG classification tasks is a time-consuming process that requires task-specific expertise. In-depth knowledge is required to understand which features are relevant to the task at hand and how to properly preprocess the data before feature extraction. In addition, using manual features in a machine learning pipeline can introduce biases into the process. Deep learning can automatically learn representative features and achieve performance that is on par with, or exceeds manual feature engineering methods. In the following sections, we describe the most commonly used architectures in deep learning for EEG, such as Convolutional Neural Networks, Recurrent Neural Networks, and Transformers, and discuss issues that these architectures may have when dealing with rs-EEG in particular. We then provide an overview of more recent techniques that have shown great promise in modeling spatial or temporal relationships in EEG signals, such as Graph Neural Networks and State Space models. Finally, we provide a brief introduction to unsupervised learning and domain adaptation techniques that can help minimize the negative effects of inter-subject variability in EEG tasks.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) revolutionized the field of deep learning and transformed how practitioners approach computer vision tasks. The vast majority of deep learning techniques in medical imaging involve CNNs in some capacity, and although they are not specifically designed to handle sequences, they have achieved state-of-the-art performance in many 1D sequence tasks.

Similar to how 2D CNN filters learn primitive patterns that become more abstract as the number of layers increases, the filters in 1D CNNs learn local patterns that are composed to perform abstract tasks such as attenuating artifacts and emphasizing valuable frequencies [29]. A CNN is typically composed of interleaved layers of learned 1D filters, Batch-Normalization layers, non-linear activation functions, and pooling operations to learn salient features from the underlying data. These features are then used by a number of fully connected layers to perform the downstream task, either regression or classification.

CNNs make up the majority of deep learning architectures applied to EEG data. For motor imagery tasks, the EEGNet architecture is the most widely used [71]. It employs a paradigm often seen in temporal convolutional networks (TCN), which is splitting temporal and channel-wise convolutions into two distinct steps [29][112]. EEGNet also uses a temporal convolution filter size that is equal to half of the sampling frequency. Trying to increase CNN kernel sizes without having to design deeper models while achieving stable training is a trend that is becoming increasingly more common for time series tasks [80][73]. Others have modified popular 2D CNNs to accept 1D signals with great success, instead of designing new models from the ground up. For example, the well-known Inception architecture famous for its proficiency in computer vision tasks [123] was adapted for 1D signals by creating an ensemble model made of Inception modules with kernels of various lengths [53]. This adapted model has shown success on EEG data [104] and other time-series tasks. ResNet [45] and VGG [117], two foundational vision models, also perform well on EEG data [61]. This suggests that many of the design patterns used to create effective CNNs for image tasks translate to 1D signals. Others have introduced innovative designs like Omni-Scale CNNs, specifically for time series tasks. They use a rule-based approach for selecting the quantity and size of the kernels used [127], removing the need for manual tuning.

Although effective, CNNs excel when there are clear local patterns in the underlying data, such as in epilepsy detection, sleep stage classification, or motor imagery tasks. However, in tasks using rs-EEG, obvious responses to event-related stimuli are not always present. In these cases, being able to model global temporal patterns and intricate relationships between electrodes becomes more important. This reduces the efficacy of CNNs and calls for more sophisticated architectures to capture task-related structures in the signals.

2.3.2 Sequential Modeling

Although effective on 1D signals, CNNs have a locality bias that leads them to place higher importance on adjacent features [6]. This makes modeling long-range relationships challenging. Models that are built from scratch for sequential processing are crucial for learning important features from time series, and the following subsections will outline some of the most prominent architectures designed for this purpose.

Recurrent Architectures

A Recurrent Neural Network (RNN) generalizes feed-forward neural networks to sequences [122] and allows for input and output vectors to vary in length. This is useful for medical time series, which can be heterogeneous. RNNs allow information to flow across time steps by maintaining a hidden state, which functions like a memory of past inputs. They integrate information from each time step sequentially into their hidden state, and their recurrent formulation considers the previous hidden state along with the current input to generate an output. The RNN updates its current hidden state h_t with a new input x_t and information from the previous hidden state h_{t-1} to produce an output y_t at time t according to Equation 1 [122].

$$\begin{aligned} h_t &= \sigma(W^{hx}x_t + W^{hh}h_{t-1}) \\ y_t &= W^{yh}h_t \end{aligned} \tag{1}$$

Although RNNs were a pioneering step towards designing neural architectures specifically for sequences, they suffer from vanishing or exploding gradients. This occurs due to gradients accumulating across multiple time steps during back-propagation, causing them to either disappear (vanish) or grow disproportionately large (exploding) due to instabilities in the initial conditions of the network's initialization. This made handling very long sequences challenging. Long Short-Term Memory networks (LSTM) were soon developed to address these issues [48]. The LSTM replaces the RNN units with a memory cell, depicted alongside the simple structure of an RNN in Figure 2.3.

The introduction of the cell state C is the core improvement of the LSTM. It acts as a persistent

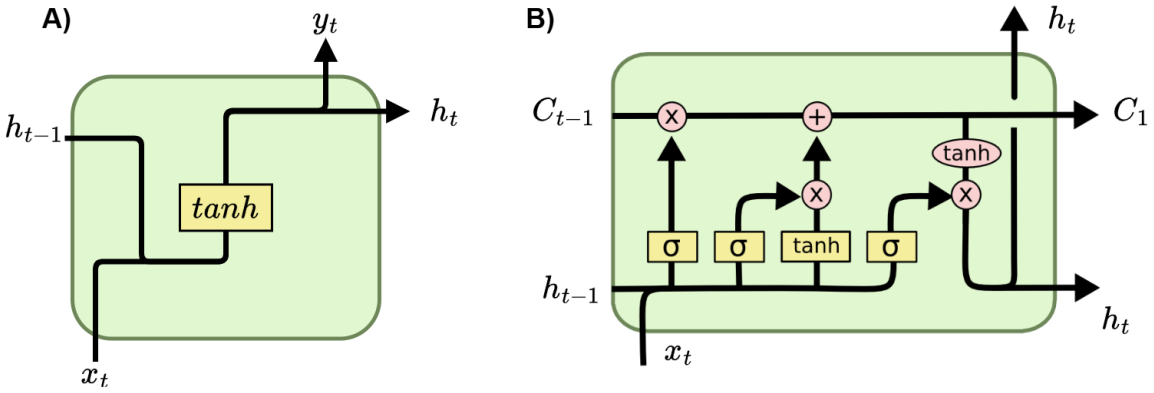

Figure 2.3: **A)** The cell of an RNN. **B)** The LSTM cell.

long-term memory that uses three main gates to modulate the information it retains, depending on current inputs and past hidden states. The input gate controls the portions of input data that enter the cell memory, the forget gate decides which data should be discarded from the cell state, and the output gate determines what information from the cell state should be used to produce the output. By introducing the cell state, the LSTM avoids the issue of vanishing gradients by providing at least one path, where the gradients will not accumulate sequentially and disappear.

LSTMs are overwhelmingly used over their more dated RNN counterparts for EEG time-series classification. They can also be used in conjunction with CNN layers to improve local feature extraction [21]. Although effective for capturing both long- and short-term dependencies, LSTMs still face problems when input sequences get too long. Their sequential processing characteristics make them hard to parallelize, and longer input sequences can result in lengthy training times. They also have a limited memory capacity since information must be compressed into a cell state and they struggle to revise information storage decisions made by their gates [7]. In general sequential modeling, they have largely fallen out of favor to Transformer architectures that are significantly easier to parallelize. However, some of these shortcomings have been addressed by the recent release of the xLSTM model [7], a revised LSTM variant that may be well poised for very long sequence modeling in the future.

Transformers

The Transformer architecture, first introduced by Vaswani et al. for the purpose of natural language processing (NLP) [132], quickly became the state of the art for sequence modeling tasks. By leveraging the self-attention mechanism, Transformers are able to attend to different tokens in an input sequence with various levels of importance, which allows them to ignore noise and focus on portions of the input that are important to the downstream task. Self-attention, given by Equation 2, first linearly maps an input sequence to query Q, key K and value V vectors. A softmax operation then generates attention weights for each element in the value vector, and the resulting multiplication with V generates a weighted output. In Transformers, the self-attention mechanism is further extended to multi-head self attention, which performs the same operation H times, once for each "head" and linearly combines the result from each head. The multi-head extension can be compared to the multiple kernels in CNNs, and is able to learn multiple different representations in parallel, increasing expressivity.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

For medical time series, Transformers are rarely applied naively to the input sequence due to the quadratic computational complexity of the self-attention mechanism relative to the input length. Instead, the time steps in the input time series are often grouped into windows or patches, and projected to a higher-dimensional embedding space. These patches are then fed to a Transformer.

Although effective on sequences, Transformers typically make very little assumptions about the underlying structure of the data compared to CNNs (locality assumption [138]) and LSTMs (recurrent bias [129]). This means that Transformers require significantly greater amounts of training data compared to CNNs and LSTMs in order to learn latent relationships. However, unsupervised or semi-supervised pretraining can alleviate this issue to certain degree. In EEG, an example of this is LaBraM, a large Transformer model that is pretrained by learning how to reconstruct the phase and amplitude of input signals. The pretrained Transformer is used for a BCI downstream task [56]. BENDR is another Transformer model that is instead pretrained using a contrastive learning objective in order to perform well on BCI applications and even outperforms other fully-supervised

benchmarks [67]. Although pretrained Transformers show promising results in EEG tasks, there are still no true foundation Transformer models akin to those in computer vision and NLP. Currently, Transformer architectures are created for narrow experimental paradigms, and how they work on other tasks remains to be studied. The question about the optimal Transformer architecture for time-series analysis is still an open question, and recent models developed for other time-series tasks suggest that inverted Transformer models, where attention is applied per input channel (i.e., recording from an electrode position) instead of per token, may prove to be more effective [76]. Most recently, a novel class of models has emerged, termed deep State Space models, and have presented a promising alternative to Transformers for sequence modeling while addressing many of their downsides. These models will be introduced in the following section.

State Space Models

State space models (SSM) are commonly used in the field of control systems to model dynamic systems. They have recently been adapted for deep learning tasks, where they show great success in efficiently modeling very long-range sequences. These models use two core equations, the **state equation** and the **output equation**. These equations govern how new inputs are integrated into a hidden state and how the hidden state will generate the next output. This hidden state is conceptually similar to hidden states in feed-forward networks [36]. The state and output equations, shown in 3, relate the input, hidden state, and output by using four distinct parameter matrices A, B, C, and D.

$$\begin{aligned} h^{'}(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{3}$$

A determines how the previous hidden state $h(t)$ will affect the current state $h^{'}(t)$ and the matrix B dictates the effect that the input $x(t)$ has on the hidden state. Both A and B are part of the **state equation**. In the output equation, C modulates the influence that the hidden state has on the output, and D is usually compared to a residual connection [39] and allows the input to have a direct effect on the output.

The class of state space models used in deep learning are usually referred to as Deep State Space Models. This distinction refers to models that allow the A, B, C, and D matrices to be

learned using gradient descent. The key to success for deep state space models on time-series tasks lies in how the state space matrices are initialized. One of the first works to propose a deep SSM for long-range sequence modeling was the Legendre Memory Unit (LMU) [136]. Voelker et al. use an SSM based on the dynamics of spiking neural networks, where the A and B state matrices are initialized in a way such that the system is capable of remembering a sliding window of history as a linear projection of Legendre Polynomials [137][136]. By projecting onto a basis of polynomials, the LMU is capable of remembering a greater number of timesteps than other recurrent architectures while using fewer parameters. Voelker et al. are able to efficiently model upwards of 100,000 time steps using their method, whereas a traditional LSTM would struggle past 2,000. The interaction of this memory unit with an RNN's hidden state is trainable through gradient descent, but the actual state matrices A and B are not. In a follow-up paper released in 2021, Chilkuri and Eliasmith show how to view their SSM equation as a convolution as opposed to a recurrence [18]. The parameterization of the A and B state equations, as well as the convolutional view of the SSM equation, paved the way for modern SSM variants

Gu et al. expand on the LMU and introduce Linear State Space Layers (LSSL) [40]. Their method allows the state matrices themselves to be trainable using gradient descent, and they use the recurrent and convolutional views of an SSM proposed by Chilkuri et al. to allow for fast inference and training, respectively. Their tests showed the state-of-the-art performance on very long sequence processing at the time. However, the LSSL suffered from a prohibitive computational complexity, which prevented their widespread use. Their follow-up work introduces Structured State Spaces (S4) [39] and S4-v2 [41]. These innovations use special parametrizations for the A matrix, reducing the computation time of their deep SSM by orders of magnitude compared to previous iterations. S4 and S4-v2 achieve great performance on long-range sequential modeling tasks and close the gap to Transformers for modeling sequences of discrete data, like images and text. S4 has also shown promising results for modeling EEG and has been used for seizure classification [125] and for generating EEG data [59] due to its impressive ability to capture relationships present in very long sequences.

These innovations in deep state space models resulted in an explosion of novel sequence modeling techniques, like Structured Global Convolutions (SGConv) [73], which we use in Chapter 3

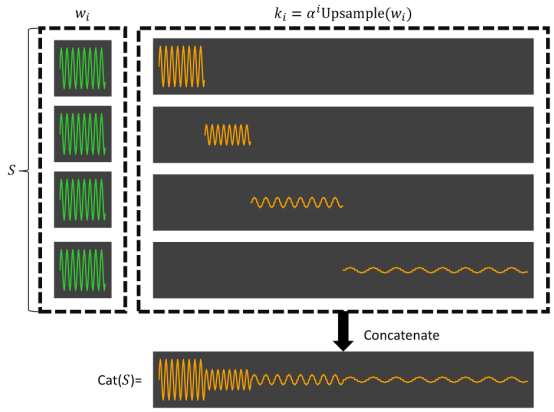

Figure 2.4: SGConv decaying kernel parameterization [73].

to extract temporal features from EEG data. SGConv was introduced to simplify the S4 structure by using insights from its success to develop a global convolution kernel. It does this by creating a decaying multi-scale kernel through the concatenation of a set of weighted sub-kernels, shown in Figure 2.4. This global convolution method has shown great results in Parkinson's disease detection [92] and in the generation of EEG data [135]. Other powerful SSMs include H3 [33] and Hyena [102]. However, the recently released Mamba SSM [37] is the first true sequential modeling technique that presents a viable alternative to Transformers.

Mamba, proposed by Albert Gu and Tri Dao, has taken the world of sequential modeling by storm. Previous state space models, like S4, are not capable of selectively filtering out parts of their hidden state depending on current inputs. This means that it cannot use the current input to decide which parts of the hidden state to forget, which leads to inefficient use of hidden state memory. Up until the introduction of Mamba, self-attention was more expressive as it enabled Transformers to dynamically change attention scores in an input-dependent fashion. Previously, state space models lagged behind in tasks, such as selective copying and induction heads due to this weakness [36]. However, Mamba improves on its predecessors by removing the linear time invariance constraint imposed on the state space matrices. That is, it allows the state parameters A, B, C, and D to vary with the inputs to the model, leading to the updated state and output equations shown in Equation 4.

$$
\begin{aligned}
h^{'}(t) &= A(t)h(t) + B(t)x(t) \\
y(t) &= C(t)h(t) + D(t)x(t)
\end{aligned}
\tag{4}
$$

We use Mamba in chapter 4 to model temporal dependencies in EEG data for classifying individuals with dementia, and although it performs quite well, Mamba still has certain drawbacks compared to Transformers. Its use of a hidden state means that it still performs the compression of previous time steps. However, Mamba's efficient implementation allows it to attend to a significantly longer input length while maintaining a sub-quadratic computational complexity, allowing for more efficient scaling than Transformers.

2.3.3 Graph Neural Networks

Whereas the previous sequence modeling techniques excel at capturing the long-range temporal dependencies in EEG data, Graph Neural Networks (GNN) have been proposed as a promising way to model spatial relationships in EEG recordings. Having a useful inductive bias for a task means that acceptable results can be achieved with fewer parameters, increased robustness, and quicker convergence. Graph neural networks introduce a relational inductive bias to a task, meaning that they place greater importance on relationships between graph elements [108]. This can take advantage of advances in network neuroscience for EEG, which analyzes EEG signals as complex graphs and studies network properties, such as functional connectivity. However, instead of extracting features and statistics from these EEG graphs manually and experiencing many of the same disadvantages as other manual feature extraction techniques mentioned in Section 2.2, GNNs can take advantage of the graph structure of EEGs and extract important relationships that are relevant to the task at hand [64].

We define an EEG graph as a set $G = (V, E, H)$, where V represents the set of all nodes, E is the set of edges, and H is the matrix of D-dimensional node features. In an EEG graph, D can be a window of raw EEG signals or even features extracted manually from a single electrode. In the EEG graph, individual electrodes are represented by graph nodes, and the edges between the nodes can be represented by a range of connectivity or distance measures. The two most often used metrics for the edges include the physical distance [25] and functional connectivity between

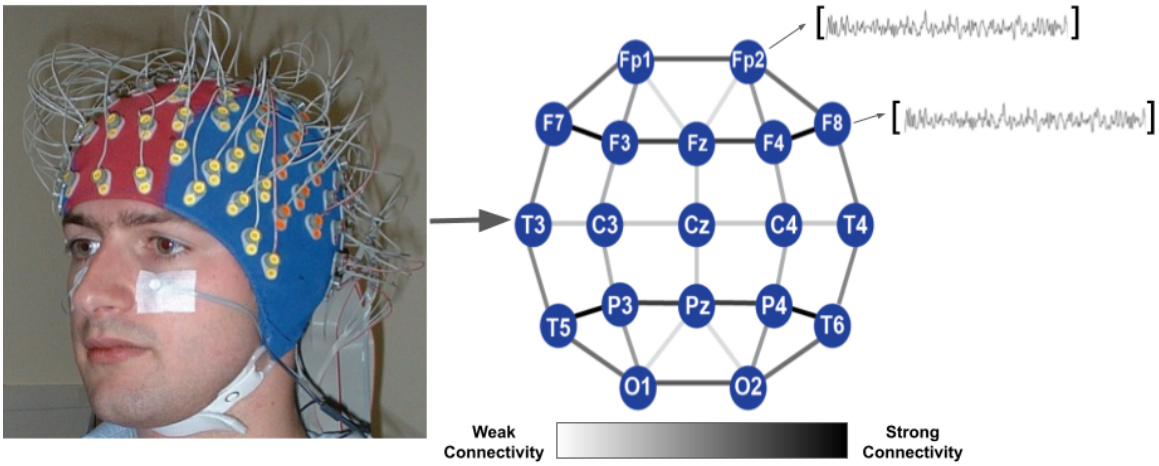

Figure 2.5: EEG modeled as a graph.

two electrodes (e.g., Pearson Correlation Coefficient [49], Absolute Cross-Correlation [124], and Granger Causality [65]). Figure 2.5 shows this graph construction approach and illustrates how relational information between EEG electrodes can be integrated into a graph.

Once signals are converted to graphs, they can be used alongside graph neural networks for a range of tasks. GNNs can perform edge imputation, node classification, or whole graph classification, the latter of which is the most common task in EEG studies. The standard GNN architecture for graph classification is shown in Figure 2.6 and is composed of three main components: a node representation learning stage, a readout function, and a classification layer [140]. The node representation learning stage updates node representations with information from itself and the connected neighboring nodes. In this stage, graph pooling layers can be added similarly to the mean or max pooling layers used in CNNs to reduce the number of nodes in a graph. The readout layer is a function that maps the set of node representations to a graph representation using Equation 5. Without a node aggregation step, performing classification using a set of nodes can be computationally expensive, as an MLP classifier would have to classify a feature vector of size $V \times D$ resulting from the concatenation of all node feature vectors in a graph.

$$h_G = Readout(h_1^K, h_2^K, ..., h_n^K) \tag{5}$$

where K is the index of the last graph convolution layer, h_G is the final graph embedding, and h_n

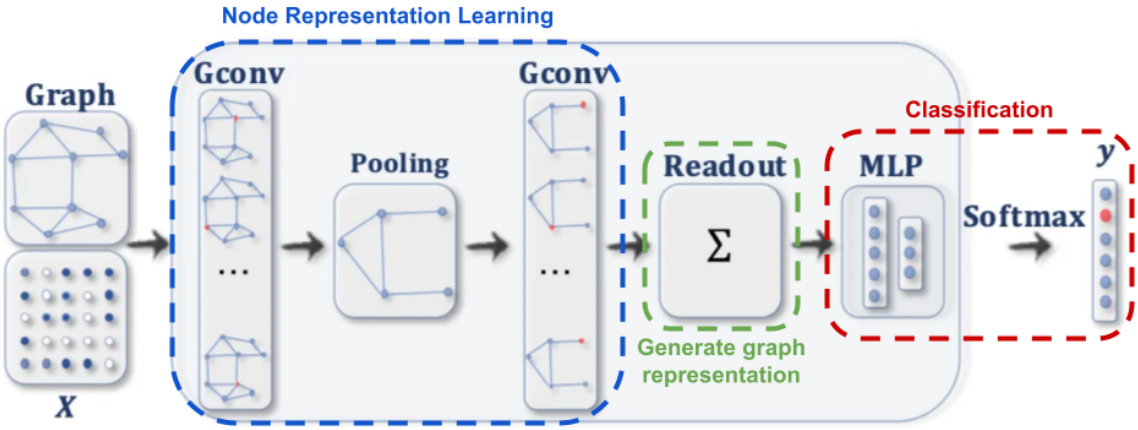

Figure 2.6: General architecture of a Graph Neural Network. GNNs are often composed of graph convolution operators, readout functions and classification heads.

is the embedding of the n-th node in the graph. The mean, max, and sum readout functions are the most commonly used [140] because they are permutation invariant (which is a highly desirable property when working with graphs), are quick to compute, and often produce great results. However, there has been more recent work that tries to develop adaptive readout functions using techniques like self-attention or gating to generate more expressive graph embeddings from node representations [12]. Once a graph embedding is obtained, it is classified using a multi-layer perceptron (MLP).

The graph convolution operators responsible for learning node representations fall into two general categories: spatial graph convolutions and spectral graph convolutions. Both are described in the following sections.

Spatial Graph Neural Networks

Spatial graph neural networks work by iteratively passing information between nodes. A node receives information from its neighbors, which is used to update its own representation [10]. This process is known as *message passing*, and a node updates its information according to Equation 6.

$$h_i^{(l+1)} = \sigma\left(W_1^{(l)} h_i^{(l)} + \sum_{j \in \mathcal{N}_{(v_i)}} W_2^{(l)} h_j^{(l)} e_{ji}\right) \tag{6}$$

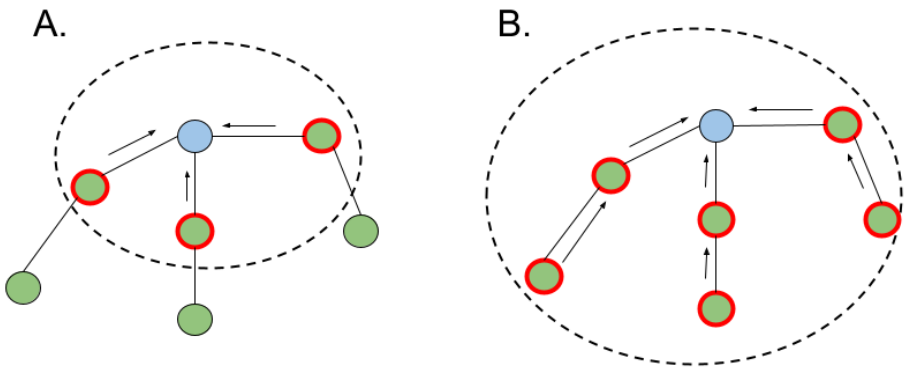

Figure 2.7: Neighboring nodes used to update a node's embedding. **A)** 1-Hop Neighborhood of the current node. **B)** 2-Hop Neighborhood of current node.

A node v_i with an embedding h_i in layer l updates its embedding to h_i^{l+1} through a linear combination of nodes in its neighborhood $\mathcal{N}$. W_1 and W_2 are trainable parameters, $\sum$ is the aggregation function that determines how the neighbor node embeddings are combined, $e_j i$ is the edge weight of the edge connecting neighbor v_j to v_i, and σ is non-linearity. Information can be transferred from longer-distance nodes through the addition of more graph convolution layers in the GNN. A single message-passing operation is performed by a spatial graph convolution and will exchange information between a node and its 1-hop neighbors. That is, nodes that are a single connection away, as shown in Figure 2.7 A). A second GNN layer will allow communication between 2-hop neighbors, and so on. This is similar to the concept of receptive fields in CNNs, where deeper networks allow the model to extract information from elements that are much further apart in the input.

Spatial GNNs are powerful as they are capable of handling graphs with inconsistent numbers of nodes. This makes them flexible and an ideal choice for datasets where graph sizes vary. However, deep spatial GNNs suffer from a phenomenon known as over-smoothing. This describes a collapse in message variance that occurs when messages travel through too many nodes and converge to similar information [64]. This ultimately leads to nodes having similar embeddings and yielding graph representations that lack expressiveness.

Spectral Graph Neural Networks

Instead of the message-passing operation described in the previous section, spectral GNNs rely on the Graph Fourier Transform (GFT). The GFT decomposes a graph into its spectrum, which is defined as the eigen-decomposition of the Laplacian matrix into its eigenvalues and eigenvectors [64, 140]. The graph Laplacian is defined as $L = D - A$, where D is the degree matrix, describing the degree of each node, and A is the adjacency matrix. The GFT is then given by $\hat{H} = U^T H$, where H is the set of node features $H \in \mathbb{R}^{V \times D}$ and U is the eigenvectors of the graph Laplacian. A spectral GNN can then be written as the convolution of a graph and a spatial kernel g in the spectral domain, which gives rise to the following element-wise multiplication[64]:

$$\mathbf{H} * g = \mathbf{U}((\mathbf{U}^T\mathbf{H}) \odot (\mathbf{U}^T g)) \tag{7}$$

where g is a kernel in the spectral domain, and $\mathbf{U}^T g$ is a learnable matrix. This defines a learned convolution on the spectral representation of a graph. Since a spectral convolution requires the adjacency matrix to be known a priori, it can only be applied to graphs with a fixed number of nodes. They are also more computationally expensive than spatial GNNs. However, spectral convolutions inherently capture global graph information, which is something that spatial GNNs cannot do with a single layer. Spectral GNNs are also more interpretable than spatial GNNs, as the learned filters applied to the spectral graph representation can be directly analyzed [10].

One of the most popular spectral convolution variants is the ChebConv graph convolution operator [22]. It reduces the computational complexity of performing a full spectral graph convolution by using a series of localized filters based on Chebyshev polynomials. This approach avoids having to perform the full eigenvalue decomposition of the Laplacian, which is what makes spectral convolutions so taxing to compute for large graphs. ChebConv is also one of the most commonly used spectral convolution filters for EEG tasks [64] and is the method we use in our GNN in Chapter 3.

Latent Graph Structure Learning

For certain tasks, the optimal functions used to calculate edge weights are not already known. Edge weights calculated using hand-crafted features, such as functional connectivity values are prone to

the same biases as manually extracted features used in machine learning tasks. We show this in Chapter 3 by demonstrating that the Pearson Correlation Coefficient, which is often used to construct the edges in an EEG graph, can severely overemphasize the connection strengths of adjacent nodes due to volume conduction effects in rs-EEG. One way to partially remedy this is by using Graph Attention networks (GAT) [133], which assign attention weights to edges between electrodes, thereby creating a proxy for edge values. However, this suffers from two main issues. First, although GATs can modulate the importance of a neighboring node's features, they still require an adjacency matrix that describes whether nodes are connected or not, meaning that connectivity needs to be determined beforehand. Second, the message-passing operation used in a GAT may not be optimal for the given task, so decoupling the graph structure learning operation from the message-passing operation is valuable.

Latent graph structure learning (GSL) aims to learn the underlying graph topology most relevant to the downstream task. That is, they apply a data-driven approach to learning the underlying adjacency matrix. Graph structure learning techniques can be categorized into two broad categories: unsupervised and supervised. Unsupervised methods do not require labeled data to learn node relationships. This makes them highly valuable in low-data regimes. However, because the downstream task has no influence on the generation of these graphs, they may not represent graph structures that are important for the downstream task. Supervised methods obtain relevant graph representations at the cost of requiring larger quantities of labeled data [16].

Within these two broad categories, GSL techniques can be further classified as either metric-based, neural, or direct approaches [148]. *Metric-based* approaches use pre-defined functions to generate similarity values between two node embeddings. For example, Zhang et al. [147] use cosine similarity to determine edge weights between nodes. On the other hand, *neural* approaches use neural networks to generate edge weights given node embeddings as inputs. For example, Pilco et al. [101] use local and global features extracted from nodes and a simple neural network to iteratively learn the edges between vertices. Finally, *direct* approaches treat the adjacency matrix itself as a set of learnable parameters and do not depend on input features or node embeddings to determine edge weights. [118] use this technique to learn an adjacency matrix for an EEG graph used in emotion recognition. In Chapter 3, we use a self-attention-based neural approach to uncover

key task-relevant edges for the task of Parkinson's disease detection and show how it can surpass results obtained using the more common Pearson correlation coefficient approach when used with rs-EEG.

2.4 Unsupervised Learning and Domain Adaptation for Electroencephalography

2.4.1 Unsupervised Learning

Medical imaging datasets are expensive to collect, need to be properly anonymized, and require substantial domain expertise to label and curate. This demanding time and resource requirement leads to dataset sizes that pale in comparison to those found in fields like natural vision and NLP. EEG datasets are comparatively smaller than many other imaging modalities, such as CT and MRI, and have only recently begun to grow in size [95][61]. In order to design effective classifiers that can make use of limited amounts of data, unsupervised learning techniques are often employed. They use pretext tasks to extract expressive feature representations that can then be used by a secondary network for the downstream task. Although there is no general consensus on the official taxonomic division of unsupervised learning techniques for EEG, many studies commonly refer to contrastive and generative pretext tasks as the main categorical divisions [23][139].

For unsupervised learning, it is useful to think of the deep learning network as a combination of a feature encoder and a classification head. The feature extractor maps the input data to a latent space, and the classification head uses this lower-dimensional feature representation for classification. In unsupervised learning, the weights of the feature encoder are learned using a pretext task with unlabeled data. The classification head can then be trained along with the feature encoder, or the feature encoder weights can be frozen (not having their weights adjusted) during training with the labeled dataset.

Contrastive Pretext Tasks

Contrastive pretraining has achieved excellent results in computer vision tasks, with techniques like SimCLR [15] and MoCo [46] sometimes outperforming fully supervised methods. Contrastive learning techniques aim to learn robust features by minimizing the distance between an anchor and its positive samples and maximizing the distance between negative sample pairs. Specifically, a positive sample pair is often an augmented view of a data sample (i.e., anchor), and a negative pair refers to a separate instance or a data point from a different class. The most critical components of a contrastive learning framework are the data augmentations used to form positive and negative pairs, and the loss function. For example, SimCLR uses a form of the InfoNCE loss function (more specifically, NT-Xent), shown in Equation 8.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{8}$$

where $\text{sim}(z_i, z_j)$ is the cosine similarity between a sample's feature vector and its augmented counterpart, τ is a temperature parameter used to adjust the weight of negative samples. The values z_i, z_j, and z_k are the outputs from a projection head attached to the feature extractor of the model being pretrained. The SimCLR framework does not explicitly sample negatives, but instead uses all other samples in a batch as negative pairs. For EEG data, Mohsenvand et al. [88] use this framework to learn rich features from signals that can then be used in a downstream classification task. They apply augmentations to EEG signals similar to those applied to images in computer vision. These augmentations, along with an overview of the SimCLR framework, are shown in 2.8.

Although powerful, the downside of many contrastive learning techniques is the process of selecting negative samples. When using a dataset with few distinct classes, there is a high chance that the negative and positive samples will belong to the same class. This reduces the effectiveness of contrastive pretraining methods. Contrastive pretraining methods also require large batch sizes to increase the odds of selecting positive and negative pairs from distinct classes, which means that computational costs are higher compared to other pretraining pretext tasks. However, contrastive methods are simple to implement and can still perform well for small datasets, even with low label counts. In Chapter 3, we make use of the SimCLR framework to increase the performance of a

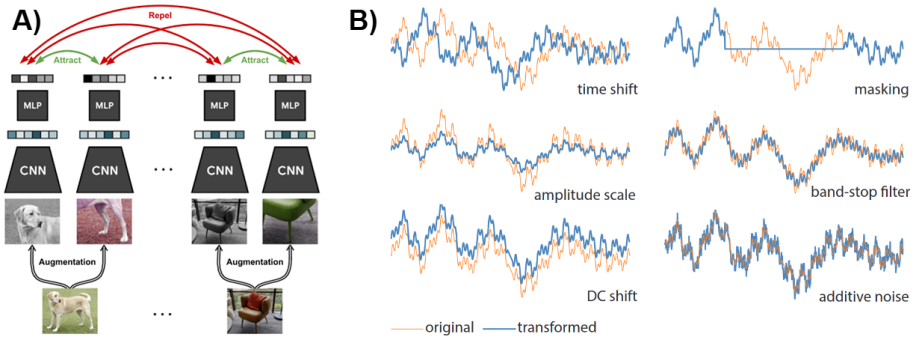

Figure 2.8: **A)** SimCLR contrastive learning framework applied to 2D images. The loss function encourages bringing augmented views of a data point closer in latent space while repelling unrelated samples. **B)** Data augmentations applied to EEG [88].

classifier for Parkinson's disease detection with a small dataset.

Generative Pretext Tasks

Generative pretext tasks either generate unseen signals that can extend outside of the range of the input data or reconstruct input signals in order to learn robust and generalizable features that capture contextual information and correlations [139]. In EEG pretraining, the most common generative pretext task is masked signal reconstruction. In this task, portions of input signals are occluded, and the model predicts the values of the original signals [97][56]. Another less-used but effective technique when pretraining models for event-based tasks is future signal forecasting. This task involves using a historical portion of an EEG sample to predict future time steps and has previously been used in seizure classification [125] and motor imagery tasks [47].

The commonly used loss functions in generative pretext tasks are the Mean Average Error (MAE) and Mean Squared Error (MSE) [23]. There is no general consensus as to which is best for generative tasks involving signals, but MAE tends to be more resilient against outliers, while MSE penalizes outliers more severely. In general, generative pretext tasks require the training of an encoder and decoder, which makes the training process more complex, and contrastive pretext tasks are largely favored for EEG.

2.4.2 Domain adaptation

Medical imaging is very vulnerable to domain shift, and EEG is no exception. Domain shift describes a change in the distribution of the training and testing datasets. In EEG data, strong distribution shifts are encountered between subjects. Individual differences in a subject's physiology and cognitive processes can lead to signals that vary substantially, and a model that performs well for one individual may completely fail when tested on another. Even intra-subject differences can heavily influence results [107]. The mental and emotional state of a subject can have an effect on the recorded signals during the same trial, especially in rs-EEG [85]. Variations in sensor placement, sensor type, and even environmental changes can affect signal recordings and can skew the distributions of the data. This problem is exacerbated for very large datasets [95][61] as it is even more difficult to ensure experimental consistency across recording sessions spread over the span of many years. This is the main reason why studies that perform validations using training and testing sets containing data from the same subjects on the same recording days usually tend to be inflated, as they do not account for the domain shift issues in EEG. A model trained on a certain distribution will deteriorate when tested on a different one, and domain adaptation and generalization techniques refer to the methods used to address this problem.

Domain adaptation is a technique used to adapt the trained model to the testing dataset. Meanwhile, domain generalization refers to a model trained to be generalizable to any unseen domain without access to the target dataset [111]. Domain adaptation is practiced more often than domain generalization in EEG tasks, and techniques vary greatly depending on the available data and downstream tasks. For example, Dose et al. [29] use the first few seconds of an unseen subject's data to calibrate their model. By including some labeled examples from the target domain, their model is capable of adapting to the distribution of a test subject's data, and they experience a significant performance boost in doing so. Although effective for BCI tasks, this method cannot be used for neurodegenerative disease classification using rs-EEG as that would require knowing a subject's diagnosis for the domain adaptation task.

Asgarian et al. [5] use the mix-up data augmentation technique to align representations between source and target domains, with the source domain being the trials in a subject's training set and the

target domain being unseen trials. They employ this technique in the context of a motor imagery task for BCIs and show significant improvements in performance. Chai et al. [13] use an unsupervised domain adaptation technique, which uses a labeled source and unlabeled target dataset to align distributions, to improve generalizability for an emotion recognition task. They employ an autoencoder to project source and target data to a shared feature space and use a distance-based subspace alignment technique to bring the representations closer. Peng et al. [100] propose a domain adaptation method for seizure classification that first learns a generalized feature vector through the adversarial training of an auto-encoder, then aligns the feature vector to the target domain using a transform along the Riemannian manifold.

The methods mentioned above fall into the "alignment" class of domain adaptation algorithms, which aims to align either a data instance, classifier, or domain to the target distribution. Although an official taxonomy has not been settled upon, domain adaptation methods can also include techniques, such as pseudo-labeling and data manipulation [111]. Pseudo-labeling refers to when a model is trained on the source domain using pseudo-labels (labels different from the labels of the target domain and the downstream task) and asked to generate pseudo-labels for the target domain [79]. Data manipulation techniques can also help alleviate distribution shifts and work on the preprocessing level to correct input data before being fed to the classifier. It has been shown that simple normalization techniques, such as channel-wise standardization, can reduce the loss of generalizability caused by domain shift [63].

Chapter 3

Multi-Head Graph Structure Learning using Gradient Weighted Graph Attention Explanations for Parkinson's Disease Detection from Resting State EEG

A version of this chapter was presented at the MLCN (Machine Learning for Clinical Neuroimaging) 2024 workshop hosted in conjunction with the 27th Medical Imaging Computing and Computer Assisted Interventions (MICCAI) conference. In addition, the paper received the Best Paper Award (1st place) at the workshop.

C. Neves, Y. Zeng, Y. Xiao, "Parkinson's Disease Detection from Resting State EEG using Multi-Head Graph Structure Learning with Gradient Weighted Graph Attention Explanations," *The 7th MICCAI workshop on Machine Learning in Clinical Neuroimaging (MLCN)*, LNCS 15266, in press, 2024. (arxiv:2408.00906) [92]

3.1 Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative disorder worldwide [128]. Primarily characterized by motor symptoms, the complex disease can also include psychiatric and cognitive issues. MRI-based biomarkers have attracted major attention, including biochemical alteration shown in quantitative MRI and structural/functional connectivity changes revealed by diffusion and functional MRI [91]. However, electroencephalography (EEG), which records electric signals from a network of locations on the scalp is a much more cost-effective neuroimaging tool with higher temporal resolution than MRI that has also been investigated to provide neurological insights and potential biomarkers for the disease. This is especially true for remote or less privileged regions, where MRI scanners are difficult to access.

Recently, deep learning (DL)-based techniques have provided excellent outcomes for EEG analysis, but several challenges remain. **First**, most existing DL techniques for EEG rely on Convolutional Neural Networks (CNNs) that aggregate signals across channels [29][71], but such approaches can miss key spatial characteristics of EEG signals, limiting clinically relevant brain connectivity insights and explainability. **Second**, to better incorporate spatial information, graph neural networks (GNNs) that model different EEG sensors and their relationships as nodes and edges of a graph (often represented as an adjacency matrix) have been proposed. However, although stationary connectivity metrics, such as the Pearson Correlation Coefficient (PCC) or Absolute Cross-Correlation (ACC) are straightforward for deriving the graph for GNN, they often fail to capture non-stationary connectivity, overestimate the correlation between adjacent nodes due to mixing of electrical signals over the scalp surface, and may not provide true functional connectivity insights in many situations. **Third**, EEG data sampled at high frequencies often involves very long sequences, which can pose challenges for commonly used sequential DL models to capture task-relevant features. Recently, Li *et al.* [73] tackled this issue with an effective convolutional model called Structured Global Convolution (SGConv) that has surpassed state-of-the-art sequence models, including Transformers [132] and Structured State Spaces [39], by designing a global convolutional kernel that can span the length of the entire sequence. **Finally**, compared with other medical imaging data, the typically small cohort sizes of EEG datasets can pose challenges for developing robust DL

techniques in the domain.

In this work, we aim to address the aforementioned issues with three contributions. **First**, we combined structured global convolutions [73] and self-supervised contrastive learning to better model complex and long sequences of EEG data with a limited cohort for the first time; **Second**, we proposed a novel dynamic multi-head graph structure learning technique to learn the adjacency matrix of the underlying EEG data without imposing potential biases in contrast to conventional static GNN methods; **Third**, to enhance explainability of our DL model for potential clinical insights, we introduced a new technique based on head-wise gradient-weighted attentions to generate an informative adjacency matrix to reveal key task-relevant connectivities in the learnt graph. The proposed method is demonstrated for PD detection with resting state EEG.

3.2 Related Works

To date, several GNN-based methods [64] have been explored for EEG analysis, particularly for seizure detection in epilepsy. Traditionally, manually defined EEG features, such as Short Time Fourier Transform [20], power spectral density [58], and selective frequency bands [118] have been used in machine/deep learning, but can introduce biases while being time-consuming and expertise-demanding. Therefore, automatic feature extraction methods have become more desirable to reduce biases and improve efficiency. Among these, Dissanayake *et al.* [28] and Sun *et al.* [121] used stacked Long Short-Term Memory (LSTM) networks and Transformers to generate feature embeddings. Li *et al.* [73] proposed the Structural Global Convolution, which showed superior ability to model long and complex sequential signals than prior approaches. Using EEG feature embeddings as node features, different GNN designs incorporating temporal features and spatial properties of EEG data have been devised. One notable trend is the rise of attention-based GNNs, which allow for the visualization of salient edges relevant to the designated tasks to enhance DL model transparency. He *et al.* [44] used a graph attention network (GAT) in conjunction with a bi-directional LSTM for seizure detection and Demir *et al.* [26] used a GAT with additional temporal convolutions to decode motor signals. To mitigate issues with static graphs, Tang *et al.* [125] and Song *et al.* [118] employed the concept of attention to learn the graph adjacency matrix instead of the attention weights

between nodes (as in GATs). However, both of their formulations use a single attention head. In EEG-based PD analysis, Chang *et al.* [14] developed a GNN that learns attention coefficients with a graph sparsity constraint to modulate the node feature vectors for PD detection during an auditory oddball task. Further explorations are still required to enhance the efficiency, accuracy, robustness, and transparency of DL-based EEG analysis, especially for GNN-based approaches.

3.3 Methods and Materials

Figure 3.1 outlines an overview our proposed DL architecture, which is composed of a feature encoder (LongConv feature encoder), a multi-head graph structure learner (MH-GSL), a Chebyshev GNN, and a classifier made of fully connected layers for PD vs. Healthy classification.

3.3.1 Feature Encoder with Contrastive Learning

Following the success of Structured Global Convolutions (SGConv) [73] for modeling long sequential data in deep learning tasks, we incorporate it into our EEG feature encoder design, which encodes the input EEG signal to $\tilde{X}_e \in \mathbb{R}^{C \times d_m}$ (C is the number of channels and d_m is model dimension). Specifically, we follow the feature extraction network setup in the work of Vetter *et al.* [135], who modify the Structured Global Convolution layer from its original formulation to have more fine-grained control over its kernel size (referred to as SLConv in Fig. 3.1). The feature extraction network (called LongConv) consists of interleaved masked 1D convolutions, which project the input channels to a set of hidden ones while SLConv layers extract long-range temporal information from each hidden channel. Each masked 1D convolution is followed by a batch normalization layer and a GELU activation. In our adapted LongConv feature encoder design, we add an additional max pooling operation followed by a 1D convolution (Conv1D) to their network structure before the MH-GSL and Chebyshev GNN layers. To alleviate some of the issues presented by the large inter-subject variability of EEG and the relatively small dataset size, we pretrained the LongConv encoder using the SimCLR [15] framework. First proposed for natural images, SimCLR learns self-supervised data representation by maximizing agreement between differently augmented versions

of the same data sample based on a contrastive loss in the latent space. For EEG contrastive learning (CL), we adopted the data augmentations by Mohsenvand *et al.*[88], including combinations of random additive Gaussian noise, random signal masking, a flip along either the signal or electrode dimension or random DC shifts. During training, we used a simple two-layer feed forward network as the projector after the LongConv encoder to obtain a latent space representation used to compute the InfoNCE loss [96]. We used a learning rate of 0.0001, a temperature of 0.005 [88], and a batch size of 100 over 160 SimCLR training epochs.

3.3.2 Multi-Head Graph Structure Learner

Graph topology of EEG signals obtained from stationary connectivity measures and/or the physical distance between electrodes for GNN learning can be misleading and sub-optimal. To tackle this, we proposed a novel graph structure learner (GSL) using multi-head attention. Based on the graph structure layer by Tang *et al.* [125], which adopts the self-attention mechanism [132] to learn edge weights, we extended this approach to include multiple attention heads. Thus, the resulting graph structure learner can attend to different graph representations (adjacency matrices) in parallel, with each attention head providing the edge weights for its paired graph representation. Then, each head-wise learnt graph representation, together with the encoded EEG features are passed to a Chebyshev GNN, updating the features with the learnt spatial relationships. The output of the Chebyshev GNN for each head is then concatenated and projected back to the model dimension d_m using a linear layer. The adjacency matrix $A_h \in \mathbb{R}^{C\times C}$ for a single attention head h out of H heads is given by:

$$\begin{aligned} Q_h = \tilde{X}_e W_{q_h}, K_h = \tilde{X}_e W_{k_h} \\ A_h = softmax(\frac{Q_h K_h^T}{\sqrt{d_K}}) \end{aligned} \tag{9}$$

where $\tilde{X}_e \in \mathbb{R}^{C\times d_m}$ are the feature embeddings, and W_{q_h} and W_{k_h} are the parameter matrices projecting $\tilde{X}_e$ to query Q_h and key K_h, respectively.

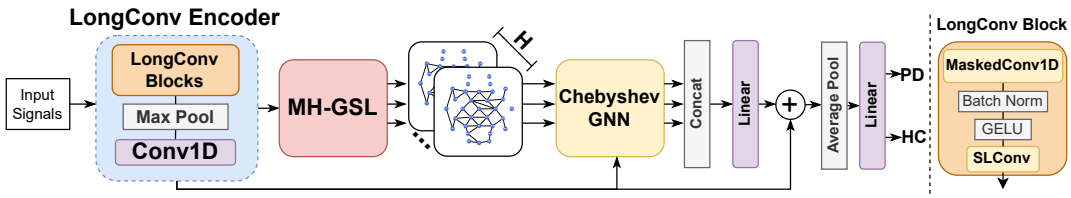

Figure 3.1: Overview of the model architecture for PD detection.

3.3.3 Graph-based EEG Classification

As shown in Fig. 3.1, the final EEG classification is achieved by first adding the head-wise aggregated output from the Chebyshev GNN and EEG feature embeddings from the temporal feature encoder, and average pooling the result along the electrode dimension to yield a final representation of shape $\tilde{X}_g \in \mathbb{R}^{C\times 1}$. A linear layer is then used to perform Healthy vs. PD classification. We use the cross-entropy loss and AdamW optimizer [78] to train our model. Here, we use the Chebyshev GNN in our model, as it has previously been used for EEG analysis [28] [78] and is an effective method of integrating an adjacency matrix with EEG feature embeddings by efficiently approximating graph convolutions using Chebyshev polynomials.

3.3.4 Head-wise Gradient-Weighted Graph Attention Explainer

In multi-head self-attention networks, the average or maximum of the head-wise attention scores [133] are often used to provide graph explainations, but this could be insufficient as some heads may carry greater contributions for decision-making. Inspired by the work of Rasoulian *et al.* [105], where head-wise gradient-weighted self-attention maps were used to improve the specificity of the attention map, we adapted the core idea for GNN-based EEG analysis. Specifically, we obtain a graph explanation by first weighing the head-wise graph representation A_h with the norm of its gradient based on the class activation. Then, the final adjacency matrix $A \in \mathbb{R}^{C\times C}$ is generated as:

$$A = \frac{1}{H}\sum_{h=1}^{H}\left\|\frac{\partial Y}{\partial A_h}\right\| \cdot A_h \tag{10}$$

where H is the number of attention heads and Y is the target class to generate a graph representation for. Finally, A is thresholded to keep the attention scores within two standard deviations from the mean, and then are normalized to [0,1].

3.3.5 Dataset and Preprocessing

We used the UC San Diego Parkinson's disease resting-state EEG (rs-EEG) dataset [106] for our study. The dataset contains the resting-state data of 15 PD patients (63.2 ± 8.2 years, 8 females) and 16 healthy controls (63.5 ± 9.6 years, 9 females). All PD patients had mild to moderate disease severity. Each participant had at least 3 minutes of resting state data recorded using a 32-channel Biosemi ActiveTwo EEG system (sampling rate = 512 Hz). We minimally preprocessed each subject's EEG by first setting the reference to the mean of the EXG7 and EXG8 mastoid electrodes and band-pass filtered the raw signal to 0.5-80 Hz. The data was then segmented into 2 sec of non-overlapping windows, resulting in 90 trials per participant.

3.3.6 Experimental Setup and Ablation Studies

To assess the classification performance of our proposed framework, we compared it against a variety of DL models and configurations. With CNN methods dominating EEG analysis, as a baseline, we re-implemented the method by Dose *et al.* [29] that showed great success on small datasets. To further validate the benefits of each design component of our method, we performed a series of ablation studies. **First**, to confirm the contribution of the Chebyshev GNN, we compared the full version of our method (CL-Encoder+Freeze) against PD detection only based on the temporal feature encoder (LongConv Encoder). **Second**, to verify whether our multi-head GSL had a positive impact on the network performance, we replaced the learnt graph structure input to the Chebyshev GNN with a static graph based on PCC, and evaluate the classification accuracy against the original design ("Full Model w/o MH-GSL vs. Full Model with MH-GSL", both without CL). **Third**, to quantify the performance gain from the SimCLR framework, we compared the proposed frameworks with and without self-supervised pre-training ("CL-Encoder+Freeze vs. Full Model with MH-GSL"). **Finally**, as some studies demonstrated the benefit of finetuning pre-trained feature encoder, we further tested our proposed method by finetuning the feature encoder weights that were pre-trained using the SimCLR framework, and compared the outcome to freezing the feature encoder weights after SimCLR pre-training ("CL-Encoder+Finetune vs. CL-Encoder+Freeze"). We computed classification accuracy, precision and recall, macro F1-score, and AUC metrics for all

Table 3.1: PD vs. Healthy classification performance for all model configurations.

Method	Accuracy %	AUC	F1-Score	Precision	Recall
LongConv Encoder	64.68±1.85	0.638±0.039	0.643±0.017	0.649±0.020	0.644±0.018
Full Model w/o MH-GSL	66.97±1.29	0.670±0.013	0.663±0.009	0.677±0.021	0.666±0.011
Full Model with MH-GSL	67.73±0.85	**0.715±0.024**	0.672±0.009	0.682±0.009	0.674±0.009
CL-Encoder + Freeze	**69.40±1.59**	0.656±0.036	**0.682±0.016**	**0.716±0.021**	**0.688±0.015**
CL-Encoder + Finetune	66.34±2.68	0.707±0.010	0.658±0.030	0.668±0.026	0.660±0.027
CNN classifier [29]	62.99±4.07	0.640±0.061	0.629±0.040	0.629±0.041	0.629±0.040

experimental setups over 3 random seeds (i.e., model weight initialization).

We trained and evaluated all configurations using a leave-one-out cross-validation, where a single subject was used for testing and the rest for training to avoid data leakage. For each fold, two subjects (one healthy and one PD) were randomly selected from the training data as a validation set. Unlike the more common sample-wise cross-validation in EEG-related DL algorithms, our subject-wise strategy can better assess the generalizability of the proposed framework to unseen subjects. Each model was trained with a batch size of 8 with a MultiStep learning rate (LR) scheduler at an initial LR of 1E-4 and a gamma of 0.1. The MH-GSL model was trained using 2 attention heads and the Chebyshev GNN used a single layer with K=5 and a dropout rate of 0.2.

3.4 Results

We present the PD vs. Healthy classification performance of all experiments in Table 3.1, and with an accuracy of 69.40±1.59%, our proposed method (CL-Encoder+Freeze) outperformed the CNN baseline [29] (accuracy=62.99±4.07%) and the other model configurations. For the ablation studies, we confirmed the positive impact of Chebyshev GNN, multi-head graph structure learner, simCLR-based encoder pretraining. Furthermore, between CL-Encoder+Finetune and CL-Encoder+Freeze, further finetuning the feature encoder during full model training decreased all evaluation metrics by 3∼5%. In addition, Fig. 3.2 presents the resulting adjacency matrices averaged for the PD and HC groups for all correctly classified samples based on static PCC-based graphs, mean of head-wise attentions from our MH-GSL, and gradient-weighted mean head-wise attention also from our MH-GSL. The gradient-weighted adjacency matrices show a greater amount of connections towards the inion (back) of the skull compared to their non-weighted counterparts. The PCC graphs show

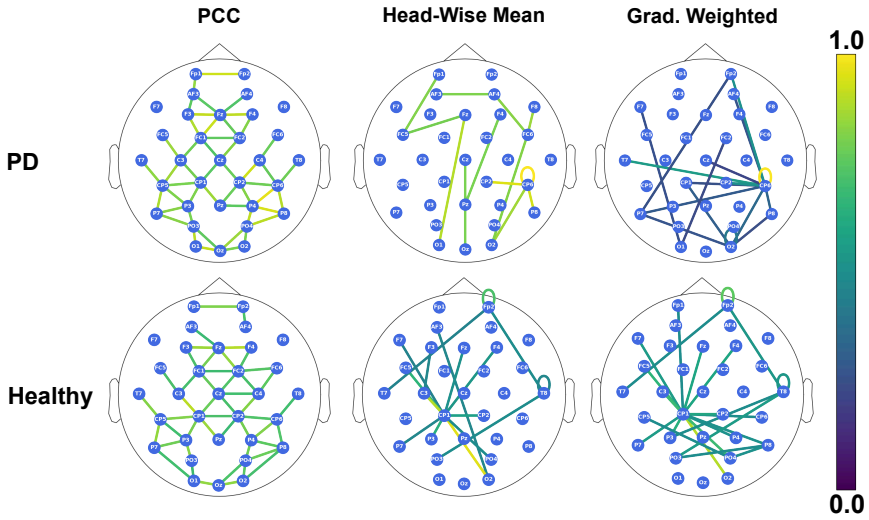

Figure 3.2: Group-wise mean adjacency matrices for PD and healthy subjects for static PCC, mean head-wise attention, and gradient-weighted mean head-wise attention.

almost exclusively connections between neighboring nodes.

3.5 Discussion

Our novel multi-head graph structure learner presents a more dynamic approach that establishes task-driven graphs with improved performance in comparison to static connectivity graphs. This observation agrees with previous studies [125]. So far, despite many attempts to learn graph edge weights using attention mechanisms[125][14], very few extended their formulation to include multiple attention heads despite their great success in vision and language tasks. Different from approaches where attention scores are multiplied with the features in an initial graph [14][72], we directly learn different node features for each adjacency matrix from MH-GSL in parallel, and finally concatenate them for classification. After testing different numbers of attention heads (2, 4 and 8), we found that two heads yielded superior performance for this task. To the best of our knowledge, we are the first to propose a head-wise gradient weighted graph attention explanation to obtain visual interpretation for task-relevant brain connectivity properties. This approach helps further highlight task-relevant graph information. Figure 3.2 reveals that graphs learnt with our method focus more on global connections across the scalp, and overcome the overemphasis on adjacent connections seen in commonly used stationary graphs. It is also interesting to note that weighing the

head-wise adjacency matrices by the norm of their gradients results in a more connected graph structure compared to its unweighted counterpart. Qualitatively, the number of connections seems to greatly increase with gradient-weighing for PD subjects, thus showing a higher connection count to be important for classification. Although an increase in functional connectivity has been shown in PD patients in resting state EEG studies [11], additional analysis of the generated edge explanations is required before drawing neuroscientific conclusions. Nevertheless, the presented technique offers great potential for deriving important connectivity information for the disorder under study. We will further validate the physiological significance of the resulting graph explanation with joint EEG-fMRI studies as the relevant insights could be of more value than PD vs HC classification.

In our experiments, we adopted a subject-wise leave-one-out cross-validation instead of a sample-wise one seen in many reports. The latter approach is often used to accommodate limited subjects in EEG datasets, but can easily cause data leakage issues, resulting in exaggerated accuracy. When adopting this commonly used strategy, our model yields near perfect classification results ($\sim$98% accuracy) potentially due to memorizing subject-specific details instead of task relevant ones. To help address limited data size, we employed contrastive learning to enhance the robustness of our feature encoder, and its benefit is evident in our experiment (1.67% accuracy increase). In comparison to fMRI and task-based EEG, rs-EEG is easier to acquire, but requires more sophisticated feature extraction techniques. Through PD detection, we demonstrated great performance of the proposed DL method and a novel graph explanation technique. We will showcase its adaptability in extended applications in the future.

3.6 Conclusion

We have developed a novel GNN technique for PD detection from resting state EEG based on dynamic graph structure learning, with a head-wise gradient-weighted graph explainer. In addition, we demonstrated the benefit of contrastive learning in efficient and robust feature extraction from a small cohort. With thorough evaluations and ablation studies, the performance of our proposed method has a great potential to offer clinical insights for PD and extended neurological applications with more accessible EEG sensors.

Chapter 4

EEG-SSFormer: Towards a Robust Mamba-Based Architecture for Dementia Detection from Resting State Electroencephalography

A version of this chapter will be submitted to the Imaging Neuroscience journal, published by *The MIT Press*.

4.1 Introduction

Dementia affects more than 58 million individuals globally [9] and encapsulates several neurodegenerative diseases, of which Alzheimer's disease (AD) is the most common. The chronic condition can manifest through a progressive deterioration of cognitive abilities, along with drastic psychological changes, and more than 60% of those affected live in low to middle-income countries [9]. Individuals commonly show signs of mild cognitive impairment (MCI) before being diagnosed with dementia, with symptom progression that varies depending on the underlying cause. Many treatments and interventions that may slow the course of the disease need an early diagnosis, ideally

while a patient is still in the mild cognitive impairment phase or earlier [55]. Biomarkers pointing to the dementia-related structural and functional changes of the brain can be used to chart its progression, and structural changes are often identified using Magnetic Resonance Imaging (MRI). Functional MRI (fMRI) and Positron Emission Tomography (PET) are used to study functional and metabolic changes brought on by the disease as potential biomarkers, and recent work has shown that cerebrospinal fluid (CSF) and blood samples can also help monitor the physiological processes underlying Alzheimer's disease [113]. However, most of these potential tests are prohibitively expensive for many, considering that 60% of affected individuals reside in lower-income nations. They are also potentially invasive (e.g., CSF tests and PET scans), and all lack portability, making them challenging to administer to remote and/or underprivileged communities. With a rapidly aging population, there exists an urgent need for practical, inexpensive, and accessible diagnostic methods that can offer objective diagnosis and prognosis of dementia.

Electroencephalography (EEG) is a functional imaging alternative that positions itself favorably thanks to its low cost, portability, non-invasiveness, and high temporal resolution. Earlier, the imaging modality has shown promise in reflecting functional anomalies arising from the structural changes caused by dementia [90]. In addition, irregular EEG patterns have also been observed to be common to early-onset dementia of many causes, and these irregularities become markedly more severe in early-onset Alzheimer's patients [83, 81], making it a promising tool for an early diagnosis. Resting-state EEG (rs-EEG) is recorded while a participant is at rest and does not require any elaborate task-based experimental protocols. This is more convenient than task-based EEG acquisition for both the patients and the EEG technicians, and a technique that can automatically detect dementia from rs-EEG can be of great value for an early and accessible diagnosis.

To date, a wide range of EEG-based biomarkers have been explored to help characterize dementia among the population. Modir et al. [86] show that the onset of dementia leads to a slowing of EEG dynamics. Measuring the latency of specific Event-Related Potentials (ERP) shows differences between the MCI and AD populations [110, 145]. In the spectral domain, studies have linked increased rs-EEG frequency band power to AD [31]. Furthermore, a loss of EEG complexity, which characterizes the regularity and predictability of signals [69], has also been used to differentiate

MCI [50] and AD [34] individuals from healthy controls (HCs). More recently, deep learning algorithms have garnered attention as they can automatically learn discriminative features from raw EEG signals without the need for complex preprocessing, which could have adverse impacts on the downstream analyses [24, 17]. Thanks to the rapid developments in deep learning approaches, developing automated diagnostic tools for the identification of dementia using raw EEG signals, particularly at its early stages, has become a real possibility.

The majority of the existing deep learning solutions applied to the task of AD and MCI detection involve Convolutional Neural Networks (CNN) either applied to raw EEG signals [52], two-dimensional spectrograms [51] or selective frequency spectrum features [66]. Recurrent models, such as Long-Short Term Memory (LSTM) and Recurrent Neural Networks (RNN) that were designed to process sequential data, have also been used to classify both raw EEG signals [4] and hand-crafted features [3] to some degree of success, and Transformer models, which model contextual importance of tokens in sequential data with self-attention, have also achieved promising results when being applied to 2D spectral representations [84] and raw signals [57] for EEG. However, learning salient features from raw rs-EEG data is much more challenging as the lack of apparent signal responses to external event-based stimuli means that deep learning techniques must instead capture sophisticated symptom-related hidden characteristics in the recordings. In this regard, recurrent DL architectures like LSTMs become difficult to train on the lengths of data seen in rs-EEG studies due to their limited memory capacity and non-parallelizable nature, while Transformers often struggle to model complex features unless given large amounts of training samples, which is usually unrealistic for EEG experiments. They also suffer from a computational complexity that scales quadratically with input length due to their self-attention mechanism, making it difficult for them to train on long signal sequences. Therefore, convolutional neural networks are often strong choices for these tasks, as they are more robust to these issues and are typically competitive with other methods [61, 60]. Additionally, the validity of some existing methods on how temporal convolutional networks and Transformers have been applied to time-series data has recently been called into question [80, 143]. Traditionally, deep learning techniques jointly embed the input channels of a multivariate time series. More specifically, as traditional deep learning models process the multivariate time series, they mix all input channels simultaneously during a projection to a higher

dimensional feature space. This means that each feature learned by the model will contain information from all input channels [93]. However, recent work has shown that better outcomes can be achieved by treating each separate input channel as a univariate time series and projecting each one to a separate feature space. This implies that each feature will contain learned patterns from only one input channel. It is believed that the increase in performance that results in using a channel-independent modeling technique may also be due to the observation that greater differences exist between channels in a multivariate time series like EEG than in computer vision tasks, where only the Red-Green-Blue channels are present [80]. Others have argued that methods combining all input channels simultaneously fail in multivariate time-series tasks because they assume that each input channel contains data emanating from the same underlying process [43]. This assumption does not hold true in EEG, where although electrodes may share some signal components due to effects like volume conduction, they ultimately capture the activity of many distinct underlying neural processes [19]. This univariate modeling strategy has found some success in EEG but is still not widely adopted. Notably, some recent EEG DL methods have benefited from the univariate modeling strategy, ranging from EEG data synthesis with generative diffusion DL models [134] to more accurate seizure detection and classification [126]. Integrating these insights and finding more effective ways of dealing with very long sequences is crucial in being able to model lengthy rs-EEG sequences.

Very recently, state space models (SSM) have positioned themselves as a strong option for very long sequence modeling as the framework describes the behavior of a dynamical system by modeling it as a collection of states and how the system transitions between these states. Deep state space models, such as Mamba [38], have achieved state-of-the-art performance in challenging long-range sequence tasks with results that match and often exceed Transformer models while boasting a computational complexity that scales linearly with sequence length. This enhanced scalability, in contrast to Transformers, is beneficial when handling the high sampling rates of EEG data, but what an effective Mamba-based DL model for robust EEG feature extraction looks like remains an open question. For the task of sleep stage and sleep disorder classification, Siddhad et al. [116] propose a dual-branch Mamba architecture, and Zhang et al. [144] combine a bi-directional Mamba with attention. In terms of multi-task EEG classification, Gui et al. [42] propose a bi-directional Mamba architecture with a task-aware mixture of experts to perform epilepsy, sleep stage, emotion, and

motor-imagery classification. Behrouz et al. [8] design a hybrid Mamba and graph neural network model capable of processing EEG and fMRI data, and Panchavati et al. [97] modified a U-Net architecture with Mamba layers for seizure detection. For motor-imagery classification, Yang et al. [141] apply Mamba across both the temporal and channel dimensions to extract relevant EEG features. To the best of our knowledge, the approach by Tran et al. [130] is the only other Mamba-based method for EEG-based differential diagnosis for dementia. Specifically, they attempt to detect Alzheimer's disease and frontotemporal dementia from the resting state EEG of 88 participants. However, they perform a trial-wise validation in their experiments (subjects may have data present in both training and testing splits), which is known to severely overestimate model performance as it trivializes individual differences between subjects, degrading generalization.

In this work, we intend to address the aforementioned issues with deep learning methods applied to rs-EEG with the following contributions. **First**, we design a novel Mamba-based DL model to address the need for long-range sequential modeling techniques in rs-EEG signal classification, which we use to allow differential diagnosis of dementia (i.e., HC vs. MCI vs. dementia classification). Specifically, our method uses a channel-independent modeling approach with effective temporal and channel mixing strategies to extract robust EEG features. **Second**, we are the first to benchmark a Mamba-based architecture using the first large-scale dementia rs-EEG dataset [61], and show improved classification performance over existing methods while using substantially fewer parameters. **Third**, by using occlusion-based explainability methods, we examine the validity of features learned by the proposed DL model and reveal key physiologically relevant insights regarding dementia and mild cognitive decline.

4.2 Materials and Methods

Figure 4.1 outlines our proposed DL architecture for HC, MCI and dementia classification based on rs-EEG. First, input EEG samples are reshaped so that each electrode channel (referred to as **channel** throughout the text) can be treated as a univariate time series throughout the model. Second, we apply patching to parse the input time series into discrete segments of length P while

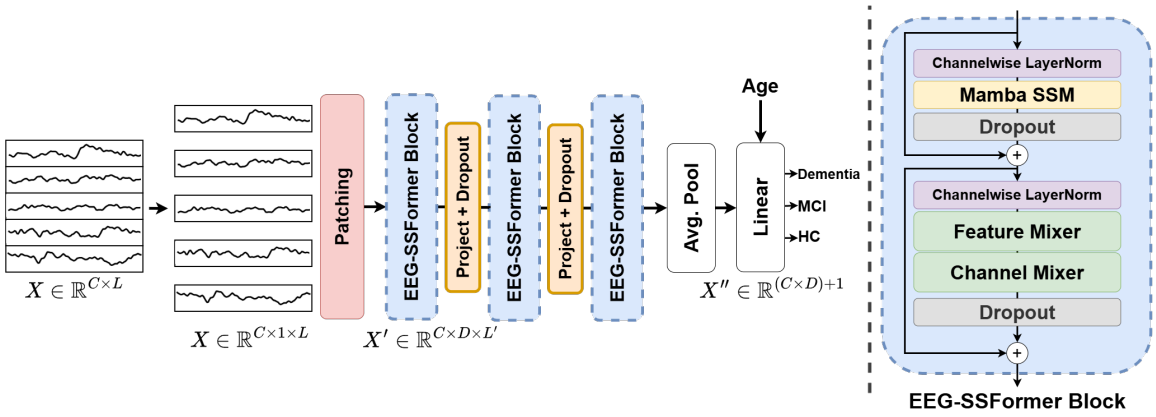

Figure 4.1: Overview of the model architecture for HC, MCI, and Dementia classification. C, D, L represent the sizes of the channel, feature, and temporal dimensions.

independently projecting each channel to a D-dimensional **feature** vector. Next, signals are processed by 3 EEG-SSFormer blocks. Each block after the first is preceded by a fully connected layer that doubles the number of features per electrode (Project + Dropout). Within each EEG-SSFormer block, a channel-wise LayerNorm and a Mamba SSM layer learn relationships between sequence segments. Then, a feature mixer learns interactions between each of the D number of features for each channel individually, after which the channel mixer captures interactions between the C number of channels for each feature. Finally, outputs from the last EEG-SSFormer block are average-pooled and concatenated with the age of the participant to obtain a final vector of shape $X'' \in \mathbb{R}^{(C \times D)+1}$, which is then classified using a linear layer as HC, MCI, or dementia. To mitigate overfitting to the training dataset, we incorporate Dropout layers [119] for regularization. We apply Dropout after the fully connected layer in the Project + Dropout block and in all three of the EEG-SSFormer blocks on the outputs of the Mamba SSM as well as after the feature and channel mixing operation.

In the following sections, we will provide an overview of the different components of our proposed DL model, including the time-series patching, channel-independent feature extraction, and decoupled channel and feature mixing. We then describe the dataset used, experimental setup, and techniques employed to interpret model outputs for potential physiological insights.

4.2.1 Channel independent feature learning

We adopt the channel-independent modeling strategy that was employed previously [80, 77, 93] to better capture distinct electrode-specific features in the input EEG data and reduce the effects of distribution shift [43]. This is done by applying the patching, projection, channel-wise LayerNorm, and Mamba operations separately for each channel.

Time series patching

Before passing through the EEG-SSFormer blocks, the input signal undergoes channel-wise z-score normalization by each channel's own mean and standard deviation instead of the more common global normalization with the mean and standard deviation of the full training dataset. We then perform patching on the input signals. The input time series is parsed into smaller segments of P number of time steps, where P is referred to as the patch length. These segments of time series data are then projected to a higher dimensional feature space of dimension D. Normally, this patching operation (parsing + projection) is performed in two distinct steps [93]. However, we perform patching using a single 1D convolution layer, similar to the work of Luo et al. [80]. Specifically, the fully convolutional patching first reshapes input signals to $X \in \mathbb{R}^{C \times 1 \times L}$. Next, we apply a 1D convolution with a stride and kernel length equal to the patch size P. The convolution operation will result in a patched output $X' \in \mathbb{R}^{C \times D \times L'}$, where L' is equal to the number of segments. Unlike individual words in Natural Language Processing tasks, a single time step does not have semantic meaning or context. Patching extracts local semantic information between groups of time steps and reduces the overall computational complexity of the model [93], motivating its use in our DL architecture. After the patching operation, the input data is normalized in a channel-wise manner using an inverted LayerNorm technique, which will be described in the following section.

Inverted LayerNorm

Unlike standard z-score normalization that is applied to input data at the preprocessing stage, LayerNorm is a mini-batch processing technique that is applied to normalize processed intermediate features between layers. The commonly used LayerNorm module can have two main issues.

First, when normalizing all features for a single time step, the resulting time series will contain few variations between features, reducing the representation power. Second, due to the large differences between time series channels, large fluctuations related to an event in one channel may introduce spurious noise in another when using the standard LayerNorm technique, thus removing the benefits of a channel-independent modeling strategy. To solve these, the inverted LayerNorm technique was introduced by Liu et al. [77] and normalizes data along the time step dimension rather than the feature dimension. This channel-wise normalization has been shown to be more robust to distribution shifts and more effective when dealing with non-stationary signals [77, 63, 75]. The detailed formulation of the inverted LayerNorm is described in Equation 11.

$$LayerNorm(X') = \frac{X'_n - \mu_n}{\sigma_n}, \text{ for all } n \in \{1, ..., N\} \text{ where } N = C \times D \tag{11}$$

Learning temporal dependencies using Selective State Spaces

After the inverted LayerNorm, we use a Mamba state space model (SSM) to extract global temporal relationships between time steps from the data. State space models describe dynamic systems and project an input signal $x(t)$ to a hidden state $h(t)$, which is then used to obtain an output state $y(t)$. This is performed through Equation 12, where the M_A matrix governs how the hidden state $h(t)$ changes over time, M_B decides how the current input affects the hidden state, M_C influences how the hidden state impacts the output and M_D allows the input to directly modulate the output.

$$\begin{aligned} h^{'}(t) &= M_A h(t) + M_B x(t) \\ y(t) &= M_C h(t) + M_D x(t) \end{aligned} \tag{12}$$

Mamba improves upon previous state space models like S4 [39] by allowing the state matrices M_A, M_B, M_C and M_D to vary based on the input. This allows Mamba to filter irrelevant portions of an input sequence while highlighting important information in a data-dependent way. This filtering process requires input channels to be mixed and projected to a higher dimensional space, so we reshape inputs to $X' \in \mathbb{R}^{(B \times C) \times L' \times D}$ before the Mamba layer, with B representing the batch dimension. This is done so that Mamba can learn temporal patterns independently for each input

channel C. In this work, we take advantage of Mamba's powerful long-range sequential modeling to learn global temporal patterns in rs-EEG data.

4.2.2 Decoupled channel-and-feature mixing

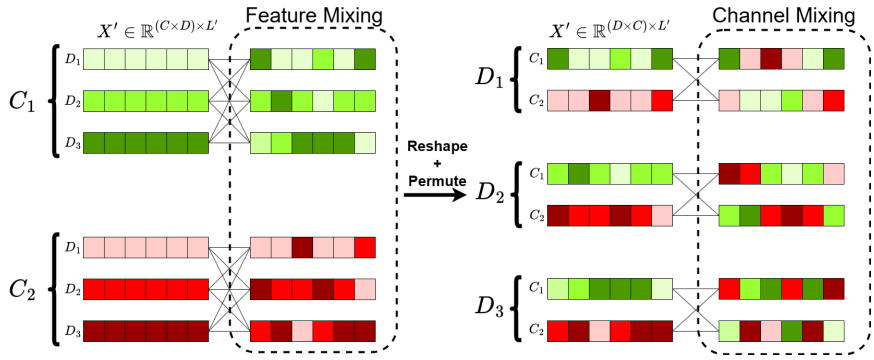

Figure 4.2: Decoupled feature and channel mixing. Feature mixing shares information between features of the same channel. Channel mixing captures cross-channel relationships for a group of features.

The patching layer and Mamba SSM jointly learn both local and global temporal patterns while treating each channel as a separate univariate time series. An important next step is to capture important cross-feature and cross-channel relationships. In this work, we adapt the strategy proposed by Luo et al. [80], who decouples the channel and feature mixing steps into two distinct operations.

Many works learn feature relationships in a coupled approach. That is, they learn inter-channel and inter-feature interactions jointly in a single mixing step, often using a simple convolution [61]. However, when using a channel-independent modeling approach, this coupled channel and feature modeling can substantially increase parameter counts. Decoupling the two operations (feature mixing and channel mixing) not only drastically reduces the parameter count of the model which decreases computational complexity, but also forces the model to use the parameters more efficiently [80].

This decoupled approach is illustrated in Figure 4.2 for a toy example of an input with two input channels and 3 features per channel. It starts with performing mixing across the feature dimension by reshaping the inputs to $X' \in \mathbb{R}^{(C \times D) \times L'}$ and performing a grouped convolution, with the number

of groups equal to the number of EEG electrode channels C. This means that C number of individual filters are used to capture relationships between each of their D number of features. Then, to model relationships between channels, we simply reshape and permute the data to $X' \in \mathbb{R}^{(D \times C) \times L'}$ and perform another grouped convolution with a number of groups equal to the number of feature dimensions per electrode channel, D.

In our final computational framework, we employed the decoupled channel and feature mixing in the spatial domain. However, alternative spectral domain mixers have been shown to be effective, and we further validate our decision to perform spatial domain-based mixing against the spectral domain counterpart for the designated application. We present this comparison as an ablation experiment in Section 4.2.2.

Feature and channel mixing in the spatial domain

We implement the decoupled feature and channel mixing step using a series of point-wise convolutions in the spatial domain. A point-wise convolution is a convolution that uses a kernel size of 1 and is functionally similar to a linear layer. For both the feature and channel mixers, we use two point-wise convolutions following an inverted bottleneck structure. That is, the first point-wise convolution projects either the number of features D (for the feature mixing step) or the number of channels C (for the channel mixing step) to be twice as wide as the input dimension. Therefore, for the feature mixer, the first point-wise convolution projects inputs from $X' \in \mathbb{R}^{(C \times D) \times L'}$ to $X' \in \mathbb{R}^{(C \times 2D) \times L'}$ and for the channel mixer, the first pointwise convolution projects inputs from $X' \in \mathbb{R}^{(D \times C) \times L'}$ to $X' \in \mathbb{R}^{(D \times 2C) \times L'}$. Afterwards, we apply a ReLU non-linearity, followed by a second point-wise convolution, to return the features or channels back to their original dimension.

Feature and channel mixing in the spectral domain

Recent work has shown that frequency domain mixers may sometimes lead to better modeling results [142, 99] for time-series data compared to their spatial counterparts. To that end, we test if this is also the case for rs-EEG classification.

To model the relationships between channels and between features in the spectral domain, we adapt the frequency domain mixer initially proposed by Patro et al. [99] called EinFFT and use

Table 4.1: Mean and standard deviation of the ages (years) for all subjects in the training, validation and testing dataset splits.

Condition	Training	Validation	Testing
Healthy Controls	65.37±9.48	64.91±10.43	63.21±8.43
Mild Cognitive Impairment	73.71±7.83	75.02±7.13	72.75±8.57
Dementia	76.59±8.01	77.25±9.57	77.25±7.17

it for both the feature and channel mixing. First, data that has been patched and processed by the Mamba SSM is reshaped and permuted to $X' \in \mathbb{R}^{B \times L' \times (C \times D)}$. The EinFFT mixer then performs the Fourier transform of the processed input data to obtain $\mathcal{X}' \in \mathbb{C}^{B \times L' \times (C \times D)}$. The data is then linearly transformed using complex-valued weights $\mathcal{W}$ and biases $\mathcal{B}$, which essentially acts as a linear layer in the frequency domain. A ReLU nonlinearity is then applied to the transformed output, followed by a second linear transformation in the frequency domain, after which the signals are finally converted back to the spatial domain using the inverse Fourier transform.

In their work, Patro et al. [99] parameterize the complex weight matrix $\mathcal{W}$ as a block diagonal matrix to reduce the overall parameter count of the model. However, they set a fixed number of 4 blocks in their parametrization, meaning that not all channels or features can share information. To enable the decoupled channel and feature mixing mentioned above, we remove the fixed constraint on the number of blocks in $\mathcal{W}$ and instead set the number of blocks to C (corresponding to the number of EEG channels) for the feature mixing step and D (corresponding to the number of features per channel) for the channel mixing step. This exactly emulates the grouped convolution mentioned in Section 4.2.2. It is important to note that both the spatial and spectral mixers have an identical number of learnable parameters to allow for a fair comparison.

4.2.3 Dataset and preprocessing

For our study, we used the Chung-Ang University Hospital EEG (CAUEEG) dataset [61], which is the largest rs-EEG dataset of patients with various stages of dementia to date. Our experiments focus on the dementia subset of this dataset, which categorizes subjects as healthy controls (HC), having mild cognitive impairment (MCI), or diagnosed with dementia. This dataset includes rs-EEG data of 1,155 subjects, with some recorded during photic stimulation. The dataset is subdivided into

training, validation, and testing sets. The training dataset contains 950 subjects (367 HC, 334 MCI, and 249 dementia), the validation set contains 119 subjects (46 HC, 42 MCI, and 31 dementia), and the testing set contains 118 subjects (46 HC, 41 MCI, 31 dementia). Most participants have well-annotated clinical diagnoses of dementia subtype, and the sex distribution consists of 6 males for every 10 females. Note that the sex of each specific participant in the dataset is removed to preserve their anonymity. The means and standard deviations of the ages for all HC, MCI, and dementia subjects in the training, validation, and testing datasets are detailed in Table 4.1.

In order to be classified as HC, MCI, or dementia, Kim et al. [61] use a series of inclusion criteria, which we summarize here. The criteria for a healthy control include: **1)** no interruption in daily activities **2)** no abnormality (within a standard deviation of age and education-adjusted baselines) on a series of neuropsychological tests [1, 54]. To be classified as having MCI, the following criteria must be met: **1)** no interruption in daily activities **2)** there must have been complaints regarding issues with memory **3)** cognitive impairment is assessed during a range of neuropsychological tests (impairment must be $\geq$ 1 standard deviation of age and education adjusted norms) [1, 54] **4)** 0.5 rating in clinical dementia **5)** the subject is not categorized as demented according to the DSM-IV criteria [32]. Finally, to be considered as having dementia, the participants must conform to the probable dementia criteria of the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association, as well as the DSM-IV [30]. We will refer the readers to the original publication [61] for further details.

For the curated data, each subject has a minimum of 5 minutes of EEG recordings, sampled at 200 Hz and recorded using 19 EEG electrodes placed according to the 10-20 placement system. An EKG or ECG electrode and a channel for the photic stimulus are also included. Since this study aims to explore the classification capabilities of deep learning architectures on raw rs-EEG data, the photic and EKG/ECG channels are left unused in our case. The EEG data is band-pass filtered between 0.5 and 70 Hz at the time of acquisition and is referenced to the common average. As in the work by Kim et al. [61], we do not perform any further pre-processing. We provide models with 10 seconds of EEG data that are randomly sampled from each participant's total available data, following the same procedure as Kim et al. [61]. This study is the first to apply a Mamba-based architecture to this corpus, and the size of the dataset ensures a robust evaluation of model

performance.

4.2.4 Experimental setup and ablation studies

To assess the effectiveness of our proposed methods, we conduct experiments comparing them to various baselines and ablated configurations in terms of HC vs. MCI vs. dementia classification performance. To measure classification performance, we compute the macro-averaged classification accuracy, macro-averaged AUROC, and the class-wise F1-scores for all baselines and model configurations. We calculate each metric over 3 random seeds and report each metric's mean and standard deviation.

As CNNs are the most commonly used DL architectures in EEG classification and the best-performing models tested by Kim et al. [61] on the CAUEEG dataset were CNNs, we implement the strongest CNN models from the previous investigation [61] as baselines. These include a 1D-ResNet-18, the best performing 1D-VGG model (1D-VGG-19), as well as a popular long-short term memory (LSTM) [48] architecture. We use the optimal hyperparameters determined by Kim et al. [61] for the CNN models since they were already subjected to a rigorous hyperparameter tuning scheme. We set the number of layers and hidden units of the LSTM to match those of our proposed method.

In addition, we perform experiments to validate various design choices for our proposed DL architecture. **First**, we assess the optimal domain for feature and channel mixing by comparing our model using the spatial channel mixer described in Section 4.2.2 (EEG-SSFormer-PW) against one model variant using its spectral counterpart (EEG-SSFormer-EinFFT). **Second**, we quantify the effectiveness of the inverted LayerNorm technique described in Section 4.2.1 on EEG signals by testing a variant of our model with the conventional LayerNorm (EEG-SSFormer-PW w/o I-LN). **Third**, we confirm the performance benefits of decoupling the feature and channel mixing into two separate sequential steps described in Section 4.2.1 by using a single coupled feature and channel mixer in the spatial domain (EEG-SSFormer-PW Coupled). **Finally**, since age has been shown to be a critical risk factor for dementia [82], we integrate it into our prediction pipeline similar to other works [61]. We name this model EEG-SSFormer-PW + Age. This model is compared with a model variant without utilizing such information (EEG-SSFormer-PW).

All models are trained using the same random-cropping scheme as in Kim et al. [61], where a training sample consists of 10 seconds of signal randomly cropped from a subject's data. This acts as a form of regularization and helps models see a larger variety of signals. We discard the first 10 seconds of each participant's data to avoid recording artifacts emanating from the trial start. During training, each model sees 100,000 random crops of signals per epoch over a total of 50 training epochs, resulting in 5,000,000 random crops seen per model over their training regime. We use identical training, validation, and testing splits as the previous authors [61], who split data in a subject-wise manner to avoid data leakage for fair evaluation. For the validation and testing sets, we split each participant's data into 10-second non-overlapping windows, resulting in 9123 samples in the validation set (3153 HC, 3435 MCI, 2535 dementia) and 8795 samples in the testing set (3027 HC, 3350 MCI, 2418 dementia). All models are trained using the AdamW optimizer. For the LSTM and SSFormer models, we use base learning rates of 0.0003 and 0.0001, respectively, a minibatch size of 32, and a cosine decay learning rate scheduler for both. We apply Gaussian noise as a data augmentation to the input signals for all methods to improve the robustness of model training.

The EEG-SSFormer architecture in Figure 4.1 employs 3 EEG-SSFormer blocks with hidden feature sizes of 32, 64, and 128 per channel. We use a dropout rate of 0.05 in the Project + Dropout layer before the second and third EEG-SSFormer blocks, a rate of 0.2 applied to the Mamba SSM outputs in all EEG-SSFormer blocks, and a rate of 0.05 for the feature and channel mixers. We use a patch size of 8 for all model variants. For the experiment involving age, we first normalize the age value using the mean and standard deviation of the training set, and add random Gaussian noise to the value. This prevents the model from memorizing the age of the subject, and overfitting to the subjects in the training dataset. We then concatenate it to the average pooled results before classifying them using the Linear layer.

4.2.5 Model interpretability

Beyond the designated classification task, investigation of the discriminative features crucial to the task can also offer relevant clinical knowledge to better understand the diseases. Therefore, we probe our model for physiologically relevant insights using an occlusion sensitivity analysis on both the individual EEG channel electrodes and each of the canonical frequency bands.

Channel occlusion sensitivity topographic maps

We generate channel occlusion sensitivity topographic maps to understand which EEG electrodes/channels (or scalp regions) are the most important for the downstream HC vs. MCI vs. dementia classification task.

We first select the subset of correctly classified HC, MCI, and dementia samples from the testing set consisting of a total of 8795 10-second samples extracted from 118 subjects. We then sequentially occlude each electrode, measure the associated drop in predicted class probability from our DL model for the correct class, and record the average value across all samples in the relevant cohort. Note that the class probabilities are obtained by applying the softmax function to the model's output logits. These channel-wise probability changes are then mapped over an outline of a scalp according to their positions in the 10-20 EEG electrode placement system, and intermediate values between the electrode positions are interpolated in order to generate a smooth heatmap over the scalp surface. The values in the resulting topographic maps further from 0 imply that the signals from that electrode were more relevant to the model's final classification. A positive number represents a drop in predicted class probability, whereas a negative number signifies that the probability for the predicted class increased after occluding the electrode. This may happen if those channels are particularly noisy or do not contain useful information for the predicted class and only serve to confound the model. Since the topographic maps are generated using correctly classified samples, electrodes that drop the class-specific probability better reflect the importance of the class.

Canonical frequency band analysis

To understand which canonical frequency bands are more relevant to the classification task with our model, we iteratively band-stop filter each of the delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-90 Hz) bands of test dataset as specified by Diessen et al. [131]. For each occluded frequency band, we calculate the new classification accuracy of the model. We then compare this new classification accuracy with the accuracy obtained without removing frequency bands and compute the **relative** change. We accumulate this relative accuracy change over 3 random seeds using the EEG-SSFormer-PW + Age model and report the mean and

standard error. If the model made use of features mostly present within a certain frequency band, then we expect a large drop in relative accuracy once that specific band is removed.

4.3 Results

4.3.1 Classification performance of baseline models and ablation studies

Table 4.2 presents the classification accuracies of our proposed method, variants of the proposed technique, the CNN baselines and the LSTM. With an accuracy of $60.14 \pm 0.57\%$, our proposed method outperforms all of the CNN baselines, the LSTM, and model variants. The second-best accuracy is achieved by a variant of our proposed method without the inclusion of the age signal ($58.42 \pm 1.10\%$). The channel and feature mixing in the spatial domain (EEG-SSFormer-PW, $58.42 \pm 1.10\%$) outperforms the model variant with mixing in the spectral domain (EEG-SSFormer-EinFFT, $57.65 \pm 0.66\%$) on average. When switching the LayerNorm module to perform normalization across tokens (EEG-SSFormer-PW) instead of across features (EEG-SSFormer-PW w/o I-LN), we see a rise in all classification metrics with an almost 2% increase in classification accuracy. Performing the channel and feature mixing in a single coupled step (EEG-SSFormer-PW Coupled) sees a drop in classification accuracy of 4.82% when compared to the decoupled counterpart (EEG-SSFormer-PW). The model configuration that includes the age signal (EEG-SSFormer-PW + Age) performs the strongest among all EEG-SSFormer architectures. This configuration includes the token-wise LayerNorm and decoupled spatial channel and feature mixing.

Compared to the CNN baseline models, we confirm the effectiveness of the Mamba-based architecture. EEG-SSFormer-PW and EEG-SSFormer-EinFFT outperform the 1D-VGG-19 model, the 1D-ResNet-19, and the LSTM model on average. All models perform the best on HC classification, and the performance of the MCI group is lowest across all methods despite the MCI group having a greater number of training subjects in comparison to the dementia group (334 vs. 249, respectively).

In terms of parameter counts, the 1D-VGG-19 model is the largest with 20.2 million parameters, followed by the 1D-ResNet-18 model with 11.4 million parameters. The LSTM is the smallest

Table 4.2: Classification results of all model configurations. Best results are in bold, second best results are underlined.

Method	Acc. (%)	Macro AUROC	HC F1	MCI F1	Dementia F1
LSTM	52.14±0.70	0.688±0.010	0.615±0.013	0.426±0.003	0.517±0.007
1D-ResNet-18	52.66±0.73	0.701±0.006	0.654±0.002	0.416±0.010	0.499±0.015
1D-VGG-19	54.99±0.39	0.714±0.003	0.689±0.003	0.460±0.002	0.491±0.016
EEG-SSFormer-PW + Age	**60.14±0.57**	**0.784±0.004**	**0.715±0.007**	**0.483±0.008**	**0.590±0.016**
EEG-SSFormer-PW Coupled	53.60±0.17	0.721±0.004	0.645±0.009	0.474±0.018	0.477±0.016
EEG-SSFormer-PW w/o I-LN	56.53±0.44	0.742±0.002	0.657±0.005	0.469±0.011	0.564±0.002
EEG-SSFormer-EinFFT	57.65±0.66	0.754±0.005	0.678±0.003	0.482±0.007	0.562±0.013
EEG-SSFormer-PW	58.42±1.10	0.759±0.006	0.686±0.005	0.480±0.027	0.584±0.005

model with 399,000 parameters, and the EEG-SSFormer-PW, EEG-SSFormer-EinFFT and EEG-SSFormer-PW + Age models have approximately 5.1 million parameters.

4.3.2 Channel occlusion sensitivity topographic maps

Figure 4.3 shows the channel occlusion scores described in Section 4.2.5. The channel occlusion scores are mapped to a birds-eye view of the scalp surface using their corresponding positions of the 10-20 placement system. Correctly classified samples from subjects in the test set are used to generate the topographic maps, which includes 2229 HC, 1704 MCI and 1336 dementia samples for the no-age group and 2481 HC, 1483 MCI and 1395 dementia samples from the topographic maps with age included. Areas between electrodes use values interpolated from the other electrodes to create smooth contours across the scalp. The top row of the figure shows the occlusion scores for the EEG-SSFormer-PW model trained without the age signal. The bottom row depicts the results for the EEG-SSFormer-PW + Age model, which includes the age signal. Overall, including the age adds robustness to the features learned by the model. There are fewer fluctuations in class probability when using a participant's age, whereas the model without age is more prone to prediction probability changes when zeroing out a channel.

The electrodes that have the greatest effect on predicted probability are over the frontal lobe, left-temporal and central-parietal lobe for HC, the central-parietal, right-temporal, and occipital lobe for MCI, and mostly over the occipital and temporal-right lobe for dementia subjects. The inclusion of age leaves these areas relatively unchanged across conditions but shifts some importance from the

center of the scalp to the inion for the MCI group.

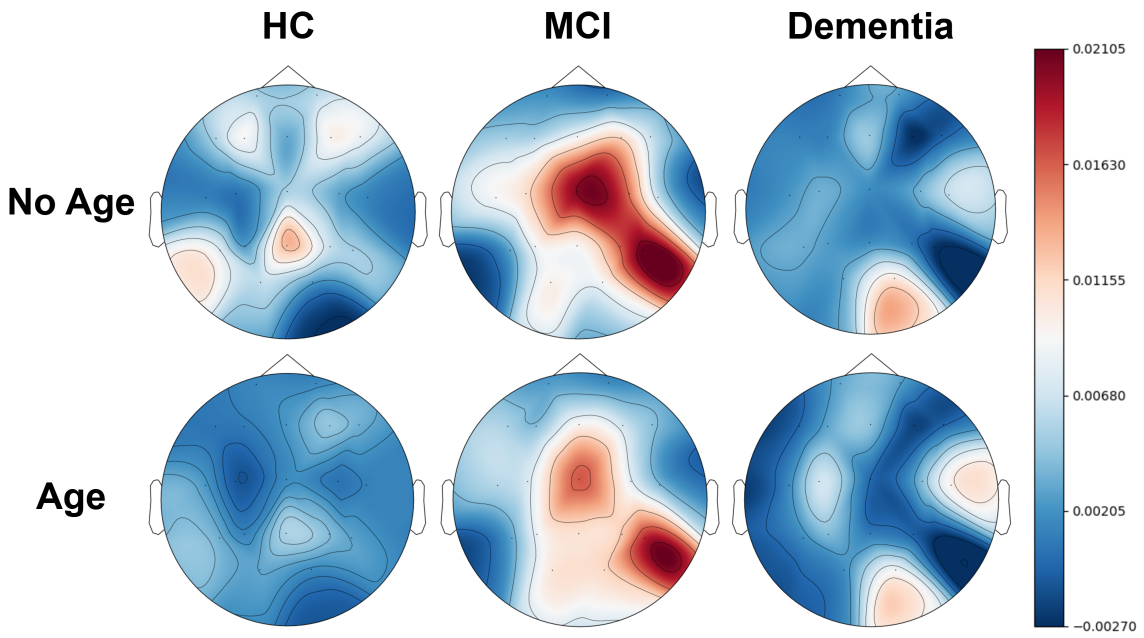

Figure 4.3: Channel occlusion sensitivity topographic maps for EEG-SSFormer model trained without and with the age signal.

4.3.3 Canonical frequency band analysis

The relative changes in class accuracies after iteratively band-stop filtering each canonical frequency band are shown in Figure 4.4. Removing the delta band has the lowest relative accuracy decrease of all the frequency bands but results in an 8.1% relative increase in accuracy for the dementia class. The removal of the theta band shows a substantial decrease in relative classification accuracy of 39.7% in dementia, with little performance change for the HC and MCI cohorts. The absence of the alpha band results in an equal and opposite effect for the MCI and dementia groups, with an 8.6% drop in relative accuracy for the MCI group, albeit with a large standard error, and a corresponding 12.2% increase for the dementia detection accuracy. The filtering of the beta band is responsible for the largest drop in HC performance among all frequency bands, with a 21.6% drop in relative accuracy, but is accompanied by an increase in classification performance for the MCI and dementia groups. Finally, the removal of the gamma band sees a large drop in MCI classification

performance, the largest drop for MCI of all the frequency bands with a 32.9% relative decrease in accuracy performance.

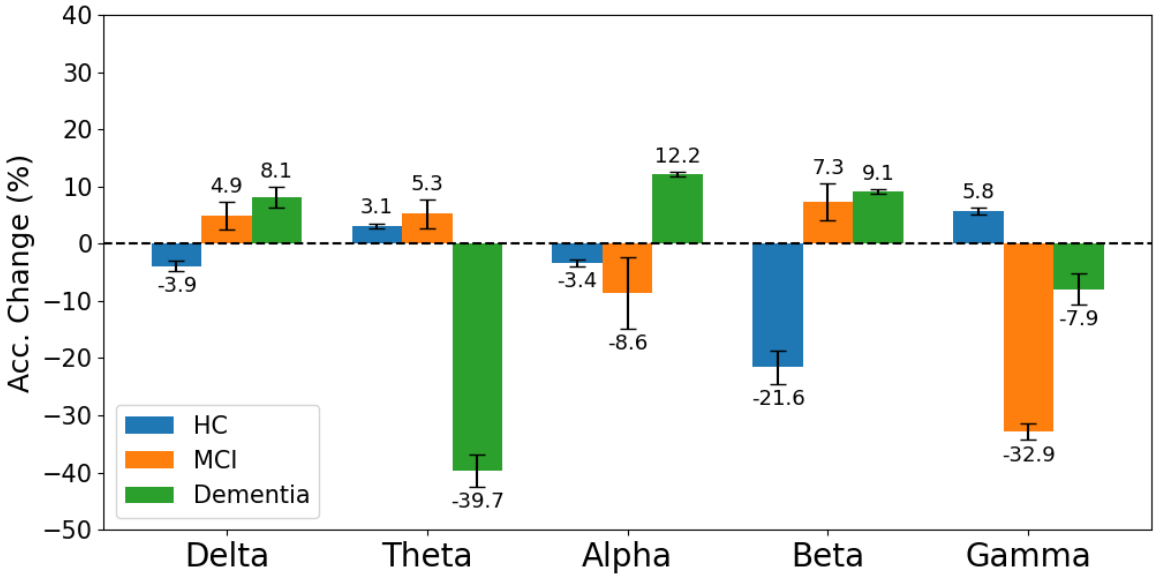

Figure 4.4: Relative accuracy change of best performing EEG-SSFormer model configuration with canonical frequency band-stop filter.

4.4 Discussion

In this work, we present a novel Mamba-based channel-independent architecture that effectively extracts salient features from raw rs-EEG signals for classifying dementia while outperforming models with close to four times as many parameters. This suggests that Mamba-based architectures may be more suitable than pure CNN architectures, which have achieved state-of-the-art results in EEG classification tasks. Additionally, our model is validated on the largest public dementia rs-EEG databases, and exploring the features learned by our model reveals physiologically relevant insights.

While gaining strong popularity in computer vision and natural language processing domains, Mamba has seldom been used for EEG classification and even less for resting-state paradigms and dementia detection tasks. Recently, Tran et al. [130] use an ensemble model featuring Mamba to classify inputs containing both raw EEG and manually extracted spectral features for the task of discriminating between subjects with Alzheimer's disease, frontotemporal dementia, and healthy

individuals. To the best of our knowledge, they are the only existing work attempting to apply Mamba to dementia detection using EEG, but their use of manually extracted features steer the scope of the task away from using minimally processed signals. In addition, they use a trial-based experimental protocol, where they shuffle the extracted EEG segments from all participants and split the segments into training and testing sets. As a result, in this setup, a model will be able to see data from a single subject in both the training and testing splits (i.e., data leakage), allowing the model to memorize subject-specific instead of task-specific details, reducing the algorithm's generalizability to unseen subjects. Other studies have applied more traditional DL techniques for dementia classification. Sen et al. [114] use the intrinsic time-scale decomposition to extract rotation components from EEG signals, then use a 1D CNN to classify patients with Alzheimer's disease in an in-house dataset. However, they also employ the trial-based validation setup and report a classification accuracy of 94%. Radwan et al. [103] use graph neural networks with Granger causality graphs for the binary classification of abnormal EEGs with the CAUEEG dataset [61]. Finally, Farina et al. [31] compare machine learning classifiers trained separately on manually extracted features from fMRI and rs-EEG data for AD vs. HC, MCI vs. HC, and AD vs. MCI classification. Similar to the results reported in Table 4.2, they also show that the MCI group is the hardest to classify correctly out of the three conditions. Due to the difference in algorithm validation setup and differences in the datasets, it is difficult to directly compare the classification accuracy of our method against these aforementioned works.

To validate our approach, we compare our work to the models developed by Kim et al. [61], which include the popular ResNet-18 and VGG-19 CNN architectures adapted for EEG data and tested on the CAUEEG dataset. To compare our model to classic recurrent architectures, we include an LSTM baseline as well. Both our spectral and spatial EEG-SSFormer variants outperform the baseline models, with the spatial variant (EEG-SSFormer-PW) achieving a 3.43% improvement over the next best-performing baseline model (1D-VGG-19) without the inclusion of the age signal. Notably, our proposed architecture contains 5.1 million trainable parameters, substantially less than the 20.2 million parameters of the 1D-VGG-19 and the 11.4 million parameters of the 1D-ResNet-18, suggesting that using a state space model and a channel-independent approach allows

for learning the long-range features unique to each channel that are discriminative for the downstream classification task more efficiently.

In the experiment comparing the mixing of the model's hidden features and input channels in the spatial and spectral domains, the spatial mixing achieves superior classification scores, outperforming the spectral mixing by 0.77%. The frequency mixer applies the Fourier transform across all elements of the sequence, which means that it observes the global periodic components of the data but fails to localize important frequency components in time. On the other hand, time-frequency techniques such as the short-time Fourier transform are so valuable in EEG data as they combine the best of both worlds and are capable of determining *when* an important change in signal frequency occurs [89]. By ignoring the temporal component of the data, the spectral channel mixer may be ignoring important characteristics in the data that could lead to a correct dementia classification. Although frequency domain variate mixing has shown success in time-series tasks before [142, 99], the datasets used contain a maximum input length of 336 time steps. This number is dwarfed by the 2000 time steps used for 10 seconds of rs-EEG data in our experiments, and modeling the variations in frequency components over time may be less important.

Our ablation test comparing the decoupled feature and channel mixing with its coupled counterpart shows that decoupling the mixing yields a substantial increase in classification performance, with the decoupled mixer outperforming coupled mixing by 4.82%. In general, in non-channel-independent approaches, all of a model's input channels are projected to a higher dimensional feature space using a single convolution operation [61, 70]. This produces features that are a combination of data from all input channels. However, using this coupled approach poses problems when treating each input channel as a univariate time series. For a single channel-independent feature and channel mixer, to capture interactions between C number of input channels and D number of features per input channel, a coupled approach would require approximately $(C \times D)^2$ parameters. For comparison, the model variant that was tested with coupled channel and feature mixing (EEG-SSFormer-PW Coupled) contained a total of 34 million trainable parameters, whereas the decoupled version (EEG-SSFormer-PW) has only 5.1 million. This reduction in parameters not only reduces the computational burden of training the model but increases its performance as well.

To better understand the decision-making process of our deep learning algorithm, we investigate the contributions that individual EEG channels and frequency bands have on the classification outcomes. In classic EEG analysis, these factors are often of great interest as potential biomarkers for disease diagnosis or insights to better understand the target neural processes. In the channel occlusion plots in Figure 4.3, we observed similar patterns of channel importance between the models with and without using the age factor in the classification. Overall, the model that makes use of age shows lower changes in prediction probability, suggesting that the introduction of the age signal can lead to greater robustness in the classifier. This is likely due to the fact that advanced aging is often correlated with cognitive function decline [109]. Notably, for MCI, our model attributes greater importance to central, frontocentral, and right parietal areas of the scalp, which is consistent with findings by Chetty et al. [17] showing that these areas exhibit significantly higher gamma/alpha ratios in the prodromal AD participants. Farina et al. [31] also discover that the best differentiator of amnestic MCI, which is also the most common form of MCI in the CAUEEG dataset, is elevated beta power in the right temporoparietal areas. In terms of the dementia subjects, there is a lower emphasis on central regions of the scalp as shown in Figure 4.3, and a higher emphasis is placed on occipital regions, with some important electrodes located on the right temporal regions. In their study, Giustiniani et al. [35] find that the central and occipital areas of Alzheimer's disease individuals exhibit higher theta power than those with frontotemporal dementia and vascular dementia. Although the dataset used in this study includes a collection of dementia subtypes, patients with Alzheimer's disease substantially outnumber those with vascular dementia (230 vs. 79) and thus may be overrepresented in the occlusion plots.

We also investigate the importance of features learned throughout individual canonical frequency bands on the diagnostic task. As shown in Figure 4.4 the theta frequency band is crucial for dementia detection. This is consistent with the conclusion of Farina et al. [31] that theta power is the strongest predictor of Alzheimer's disease status in resting-state EEG, and they report that the significance of the theta band is consistent across multiple regions of the scalp. Chetty et al. [17] come to the same conclusion when discriminating between Alzheimer's disease, prodromal Alzheimer's disease, and healthy controls from resting-state EEG features. They find that the theta power of Alzheimer's patients was elevated compared to their prodromal and healthy counterparts. Besides

theta band, previous studies [68, 35] also suggested increased delta band in rs-EEG to be associated with Alzheimer's disease. In our case, the removal of the delta band contributed to a slight increase in the classification accuracy of dementia. This may be due to the fact that we focus on the symptom of dementia across a number of underlying causes other than specifically AD. For health controls, we find the most discriminative frequency band lies in the beta band. This finding is mirrored in the work of Farina et al. [31], who state that beta band power is the most significant predictor of healthy control status. Finally, increased gamma vs. alpha power ratio and gamma-band functional connectivity in the prodromal Alzheimer's disease population has been shown in previous studies [17], which are consistent with the significant drop in relative classification accuracy that our model experiences for the MCI class when removing the alpha and gamma bands.

Although we show improved performance using our novel DL technique, some limitations to the current work remain that can provide opportunities for future explorations. The CAUEEG dataset is the largest rs-EEG dataset for dementia, but there exists a few sources of heterogeneity among subjects that may impact our classification accuracy. First, the dataset contains two different experimental protocols, with a sub-cohort of patients receiving photic stimulation during data acquisition. Although our model was still capable of learning relevant EEG features for differential diagnosis, the discrepancy in experimental protocols may interfere with the learning process. Future work can explore strategies to harmonize this difference and quantify the impact of photic stimulation in rs-EEG-based dementia detection. Another source of heterogeneity lies in the diversity of disease diagnoses among the population in the dataset. While the majority of the subjects in the dataset were assigned general labels of dementia, MCI, or HC, the causes of dementia vary (e.g., Parkinson's disease dementia vs. vascular dementia), which may contribute to different characteristics in EEG activities. Future investigations that reveal the common and distinct EEG patterns of different causes for MCI and dementia may help better reveal the mechanisms of the symptoms.

4.5 Conclusion

In this work, we propose a novel Mamba-based channel-independent DL model for HC vs. MCI vs. dementia classification. Using a subject-wise validation scheme, we develop and test our

proposed method based on the CAUEEG dataset with the largest corpus of rs-EEG collected from dementia patients. Our results demonstrate superior performance compared to the state-of-the-art CNN models while reducing the model parameter by approximately four times. Furthermore, we show that our model is capable of extracting physiologically relevant features from the resting state signals, with insights in line with the current neuroimaging literature. Our state-of-the-art results offer a promising avenue to leverage rs-EEG for the diagnosis and study of dementia, which is particularly beneficial for remote areas and underprivileged communities.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we address the unique challenges of developing deep learning methods for resting state EEG data and introduce two novel methods. In Chapter 3, we propose a hybrid GNN model using Chebyshev graph convolutions and a novel multi-head graph structure learning framework to extract spatial relationships between EEG electrodes and structured global convolutions to learn temporal patterns in the signals. We introduce a novel graph explainability technique that weighs the learned attention-based edge weights by their head-wise gradients to produce adjacency matrices that are more descriptive than the mean and max aggregation that has been more commonly adopted. Our framework shows promising performance for classifying patients with Parkinson's disease, and we push the capabilities of our model even further by using a contrastive pretraining task designed specifically for 1D signals. Our work shows that manually designed edge weights based on the commonly used Pearson correlation coefficient are sub-optimal for identifying important electrode relations in resting state EEG, and graph structure learning can be a promising alternative.

Our second contribution, presented in Chapter 4, shifts attention to the task of dementia detection. We introduce a novel deep-learning architecture that leverages the power of the Mamba state-space model and a channel-independent modeling technique to classify individuals' cognitive decline stages. Our study also tests the viability of modeling cross-feature and cross-channel interactions in both the spatial and spectral domains, and we show that applying conventional deep

learning methods in a channel-independent strategy helps increase model generalization in EEG. We also show that our model can learn physiologically relevant insights in a data-driven manner instead of the classic hypothesis-driven exploration in typical neuroimaging research.

Diagnostic screening should generally be accessible for individuals regardless of economic status or physical capabilities. EEG, particularly rs-EEG, is a step towards more accessible testing alternatives for many disorders. In this thesis, I presented novel deep learning methods that can address issues encountered when using deep learning with resting state EEG and offer researchers insights into what leads the proposed models to their conclusions. Hopefully, these insights can spur more focused work into relevant physiological biomarkers to help describe neurodegenerative diseases.

5.2 Future Work

In our first contribution, our graph structure learning layer learns adjacency matrices that are more descriptive than commonly used hand-crafted ones. Although there are some similarities between the graphs learned by our method and those extracted from Parkinson's disease patients using network analysis, future studies can help confirm our observation in relation to the existing neuroscience literature by studying the subject-wise graphs learned from a much larger cohort. To expand on this research direction, group-wise differences between Parkinson's disease subtypes can be investigated and corroborated with the existing clinical literature to evaluate the robustness of the physiological insights generated by our technique.

Previous works that employ graph structure learning for EEG have imposed graph regularization constraints on the adjacency matrix during the training process [125]. These constraints can enforce sparsity, low node degree values, and varying degrees of smoothness to encourage the structure learning layers to learn EEG graphs with desirable properties that can be more easily compared with insights from the network neuroscience literature. Regularizing the graph learning process can also help avoid solutions converging to sub-optimal results, such as a single node with a disproportionately high degree value.

In our second main contribution, the Mamba temporal feature extractor models long-range dependencies in the underlying signal. While our model outperforms CNNs that have significantly more model parameters, testing the efficacy of other SSM variants may potentially lead to increased performance in future studies. Mamba uses a purely real-valued hidden state, which allows it to excel at many tasks while remaining very computationally efficient. However, some preliminary results show that real-value SSMs under-perform their complex-valued counterparts (e.g., S4) on some tasks involving audio and video data. It is hypothesized that complex-valued hidden states may be optimal for continuous data modalities. In contrast, real-valued hidden states may be better for discrete data such as texts. So far, a wide range of SSMs have been developed, and it can be interesting to understand how the trade-off of complex vs. real hidden states affects EEG data processing for downstream tasks. In addition, we extend our second main contribution by performing a preliminary study on model pretraining with a forecasting task, which is outlined in Appendix A.

Appendix A

Investigating future data forecasting as a SSL pretext task for dementia classification using resting state EEG

This section shows the results of an investigation on the feasibility of self-supervised pretraining using data forecasting as a pretext task for dementia classification from rs-EEG. Previous research in computer vision using Mamba has demonstrated improved performance through pretraining [74]. Self-supervised (SSL) pretraining tasks may help the model generate valuable representations from limited-size datasets without requiring additional data annotations. These representations may then be used in a downstream classification task. However, in the context of EEG, future data forecasting as a pretext task has been largely under-explored. Most studies have focused on pretraining techniques related to signal reconstruction [97, 57], contrastive learning [115, 87], or synthetic data generation [134, 98]. The work by Tang et al. [124] is notable for being the first to employ signal forecasting as a self-supervised pretraining task for seizure classification using EEG and He et al. [47] use future data forecasting for a motor imagery task. To the best of our knowledge, nobody has applied this pretext task to rs-EEG data.

For this task, we train the EEG-SSFormer-PW model introduced in Chapter 4 to predict the next n time steps of preprocessed EEG data given a 10-second input segment. We replace the final linear

classification head with a forecasting head, which consists of a flattening operation that converts average pooled outputs from $X \in \mathbb{R}^{B\times C\times D\times 1}$ to $X \in \mathbb{R}^{B\times (C\times D)}$, and a linear layer that generates predictions for the following n time steps. We use a Mean-Squared Error (MSE) loss function similar to He et al. [47], shown in Equation 13 where Y_i is the ground truth of the i^{th} time step, and $\tilde{Y_i}$ is the model prediction. We evaluate the model's performance using macro-averaged Accuracy (%).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2 \tag{13}$$

The pretraining pretext task is evaluated in two experimental setups. We finetune the model using the labeled dataset during model training in the first setup. This is "Finetuned Accuracy" in Table A.1. The second setup freezes the model, except for a Linear classification head. This test aims to quantify the quality of the features extracted by the pretrained model. These results are reported under "Frozen Accuracy" in Table A.1. We pretrain the EEG-SSFormer model for up to 350 epochs with a base learning rate of 0.0002 using a cosine-decay learning-rate scheduler and forecast the next 2 seconds (400 time steps) of data. The model is evaluated every 50 epochs. We use the same EEG-SSFormer-PW model hyperparameters as in Chapter 4.

Across the board, pretraining the model with the forecasting task deteriorates performance on the downstream classification task, with the frozen feature encoder obtaining worse results. Our experiments show that this pretext does not improve classification accuracy when applied to rs-EEG data for dementia classification. With epilepsy and motor imagery tasks, there is a clear dependence between time steps. For example, seizures are usually characterized by the prodrome, ictal, and post-ictal stages, indicating that the seizure is incoming, onset, and subsiding. This implies a temporal ordering that may be easier to model for a forecasting pretraining task. According to [146], if a forecasting pretext task is used for signals or sequences with high degrees of uncertainty, randomness, or sudden unexpected events, then the model may struggle to forecast future values. This negatively impacts the effectiveness of the self-supervised learning task. Rs-EEG data lacks the clearer structure of event-based tasks and is prone to seemingly random fluctuations, which may be why the forecasting pretraining is ineffective. The decrease in classification accuracy is noticeably

Table A.1: Performance of frozen and fine-tuned pre-trained models on the classification task outlined in Chapter 4.

Pre-training Epochs	Finetuned Accuracy (%)	Frozen Accuracy (%)
50	54.97	50.10
100	55.60	48.74
150	54.82	48.03
200	56.37	44.24
250	54.96	41.41
350	55.47	43.45

worse for the frozen encoder configuration, suggesting that the features the model learns to extract during the self-supervised learning task are not informative enough for classification.

Bibliography

[1] Hyun-Jung Ahn, Juhee Chin, Aram Park, Byung Hwa Lee, Mee Kyung Suh, Sang Won Seo, and Duk L Na. Seoul neuropsychological screening battery-dementia version (snsb-d): a useful tool for assessing and monitoring cognitive impairments in dementia patients. *Journal of Korean medical science*, 25(7):1071–1076, 2010.

[2] Ali H Al-Nuaimi, Emmanuel Jammeh, Lingfen Sun, and Emmanuel Ifeachor. Higuchi fractal dimension of the electroencephalogram as a biomarker for early detection of alzheimer's disease. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2320–2324. IEEE, 2017.

[3] Michele Alessandrini, Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simona Luzzi, and Claudio Turchetti. Eeg-based neurodegenerative disease classification using lstm neural networks. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 428–432. IEEE, 2023.

[4] Ashik Mostafa Alvi, Siuly Siuly, and Hua Wang. A long short-term memory based framework for early detection of mild cognitive impairment from eeg signals. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(2):375–388, 2022.

[5] Sepehr Asgarian, Ze Wang, Feng Wan, Chi Man Wong, Feng Liu, Yalda Mohsenzadeh, Boyu Wang, and Charles X Ling. Multi-view contrastive learning for unsupervised domain adaptation in brain-computer interfaces. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[6] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius

Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[7] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

[8] Ali Behrouz and Farnoosh Hashemi. Brain-mamba: Encoding brain activity via selective state space models. In *Conference on Health, Inference, and Learning*, pages 233–250. PMLR, 2024.

[9] Deo Benyumiza, Edward Kumakech, Jastine Gutu, Jude Banihani, Joshua Mandap, Zohray M Talib, Edith K Wakida, Samuel Maling, and Celestino Obua. Prevalence of dementia and its association with central nervous system infections among older persons in northern uganda: cross-sectional community-based study. *BMC geriatrics*, 23(1):551, 2023.

[10] Deyu Bo, Xiao Wang, Yang Liu, Yuan Fang, Yawen Li, and Chuan Shi. A survey on spectral graph neural networks. *arXiv preprint arXiv:2302.05631*, 2023.

[11] Taylor J Bosch, Arturo I Espinoza, Martina Mancini, Fay B Horak, and Arun Singh. Functional connectivity in patients with parkinson's disease and freezing of gait using resting-state eeg and graph theory. *Neurorehabilitation and Neural Repair*, 36(10-11):715–725, 2022.

[12] David Buterez, Jon Paul Janet, Steven J Kiddle, Dino Oglic, and Pietro Liò. Graph neural networks with adaptive readouts. *Advances in Neural Information Processing Systems*, 35: 19746–19758, 2022.

[13] Xin Chai, Qisong Wang, Yongping Zhao, Xin Liu, Ou Bai, and Yongqiang Li. Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition. *Computers in Biology and Medicine*, 79:205–214, 2016. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2016.10.019.

[14] Hongli Chang, Bo Liu, Yuan Zong, Cheng Lu, and Xuenan Wang. Eeg-based parkinson's disease recognition via attention-based sparse graph convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Yu Chen and Lingfei Wu. Graph neural networks: Graph structure learning. In Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao, editors, *Graph Neural Networks: Foundations, Frontiers, and Applications*, pages 297–321. Springer Singapore, Singapore, 2022.

[17] Chowtapalle Anuraag Chetty, Harsha Bhardwaj, G Pradeep Kumar, T Devanand, CS Aswin Sekhar, Tuba Aktürk, Ilayda Kiyi, Görsev Yener, Bahar Güntekin, Justin Joseph, et al. Eeg biomarkers in alzheimer's and prodromal alzheimer's: a comprehensive analysis of spectral and connectivity features. *Alzheimer's Research & Therapy*, 16(1):236, 2024.

[18] Narsimha Reddy Chilkuri and Chris Eliasmith. Parallelizing legendre memory unit training. In *International Conference on Machine Learning*, pages 1898–1907. PMLR, 2021.

[19] Michael X Cohen. Where does eeg come from and what does it mean? *Trends in neurosciences*, 40(4):208–218, 2017.

[20] Ian C Covert, Balu Krishnan, Imad Najm, Jiening Zhan, Matthew Shore, John Hixson, and Ming Jack Po. Temporal graph convolutional networks for automatic seizure detection, 2019.

[21] Iago Henrique de Oliveira and Abner Cardoso Rodrigues. Empirical comparison of deep learning methods for eeg decoding. *Frontiers in Neuroscience*, 16:1003984, 2023.

[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[23] Federico Del Pup and Manfredo Atzori. Applications of self-supervised learning to biomedical signals: A survey. *IEEE Access*, 2023.

[24] Arnaud Delorme. Eeg is better left alone. *Scientific reports*, 13(1):2372, 2023.

[25] Andac Demir, Toshiaki Koike-Akino, Ye Wang, Masaki Haruna, and Deniz Erdogmus. Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1061–1067. IEEE, 2021.

[26] Andac Demir, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdoğmuş. Eeg-gat: graph attention networks for classification of electroencephalogram (eeg) signals. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 30–35. IEEE, 2022.

[27] Rohtash Dhiman et al. Machine learning techniques for electroencephalogram based brain-computer interface: A systematic literature review. *Measurement: Sensors*, 28:100823, 2023.

[28] Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Geometric deep learning for subject independent epileptic seizure prediction using scalp eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 26(2):527–538, 2021.

[29] Hauke Dose, Jakob S Møller, Helle K Iversen, and Sadasivan Puthusserypady. An end-to-end deep learning approach to mi-eeg signal classification for bcis. *Expert Systems with Applications*, 114:532–542, 2018.

[30] Bruno Dubois, Howard H Feldman, Claudia Jacova, Steven T DeKosky, Pascale Barberger-Gateau, Jeffrey Cummings, André Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, et al. Research criteria for the diagnosis of alzheimer's disease: revising the nincds–adrda criteria. *The Lancet Neurology*, 6(8):734–746, 2007.

[31] Francesca R Farina, DD Emek-Savaş, L Rueda-Delgado, Rory Boyle, Hanni Kiiski, Görsev Yener, and Robert Whelan. A comparison of resting state eeg and structural mri for classifying alzheimer's disease and mild cognitive impairment. *Neuroimage*, 215:116795, 2020.

[32] Michael B First and Harold Alan Pincus. The dsm-iv text revision: rationale and potential impact on clinical practice. *Psychiatric services*, 53(3):288–292, 2002.

[33] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[34] Sinead Gaubert, Marion Houot, Federico Raimondo, Manon Ansart, Marie-Constance Corsi, Lionel Naccache, Jacobo Diego Sitt, Marie-Odile Habert, Bruno Dubois, Fabrizio De Vico Fallani, Stanley Durrleman, and Stéphane Epelbaum. A machine learning approach to screen for preclinical Alzheimer's disease. *Neurobiology of Aging*, 105:205–216, September 2021. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2021.04.024.

[35] Andreina Giustiniani, Laura Danesin, Beatrice Bozzetto, AnnaRita Macina, Silvia Benavides-Varela, and Francesca Burgio. Functional changes in brain oscillations in dementia: a review. *Reviews in the Neurosciences*, 34(1):25–47, 2023.

[36] Maarten Grootendorst. A Visual Guide to Mamba and State Space Models - Maarten Grootendorst.

[37] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[38] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[39] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2021.

[40] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[41] Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train

your hippo: State space models with generalized orthogonal basis projections. *arXiv preprint arXiv:2206.12037*, 2022.

[42] Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. EEGMamba: Bidirectional State Space Model with Mixture of Experts for EEG Multi-task Classification, October 2024.

[43] Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[44] Jiatong He, Jia Cui, Gaobo Zhang, Mingrui Xue, Dengyu Chu, and Yanna Zhao. Spatial–temporal seizure detection with graph attention network and bi-directional lstm architecture. *Biomedical Signal Processing and Control*, 78:103908, 2022.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[46] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[47] Yanbin He, Zhiyang Lu, Jun Wang, Shihui Ying, and Jun Shi. A self-supervised learning based channel attention mlp-mixer network for motor imagery decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2406–2417, 2022.

[48] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[49] 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.

[50] Yu-Tsung Hsiao, Chien-Te Wu, Chia-Fen Tsai, Yi-Hung Liu, Thanh-Tung Trinh, and Chun-Ying Lee. Eeg-based classification between individuals with mild cognitive impairment and

healthy controls using conformal kernel-based fuzzy support vector machine. *International Journal of Fuzzy Systems*, 23:2432–2448, 2021.

[51] Cosimo Ieracitano, Nadia Mammone, Alessia Bramanti, Amir Hussain, and Francesco C. Morabito. A Convolutional Neural Network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing*, 323:96–107, January 2019. ISSN 09252312. doi: 10.1016/j.neucom.2018.09.071.

[52] Cosimo Ieracitano, Nadia Mammone, Amir Hussain, and Francesco Carlo Morabito. A convolutional neural network based self-learning approach for classifying neurodegenerative states from eeg signals in dementia. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[53] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[54] Seungmin Jahng, Duk L Na, and Yeonwook Kang. Constructing a composite score for the seoul neuropsychological screening battery-core. *Dementia and Neurocognitive Disorders*, 14(4):137–142, 2015.

[55] Vesna Jelic and Jan Kowalski. Evidence-Based Evaluation of Diagnostic Accuracy of Resting EEG in Dementia and Mild Cognitive Impairment. *Clinical EEG and Neuroscience*, 40(2): 129–142, April 2009. ISSN 1550-0594, 2169-5202. doi: 10.1177/155005940904000211.

[56] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[57] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci, 2024.

[58] Ming Jin, Hao Chen, Zhunan Li, and Jinpeng Li. Eeg-based emotion recognition using graph convolutional network with learnable electrode relations. In *2021 43rd Annual International*

Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 5953–5957. IEEE, 2021.

[59] Jaivardhan Kapoor, Auguste Schulz, Julius Vetter, Felix Pei, Richard Gao, and Jakob H Macke. Latent diffusion for neural spiking data. *arXiv preprint arXiv:2407.08751*, 2024.

[60] Ann-Kathrin Kiessner, Robin T Schirrmeister, Joschka Boedecker, and Tonio Ball. Reaching the ceiling? empirical scaling behaviour for deep eeg pathology classification. *Computers in Biology and Medicine*, page 108681, 2024.

[61] Min-jae Kim, Young Chul Youn, and Joonki Paik. Deep learning-based EEG analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset. *NeuroImage*, 272:120054, May 2023. ISSN 10538119. doi: 10.1016/j.neuroimage.2023.120054.

[62] Nam Heon Kim, Ukeob Park, Dong Won Yang, Seong Hye Choi, Young Chul Youn, and Seung Wan Kang. Pet-validated eeg-machine learning algorithm predicts brain amyloid pathology in pre-dementia alzheimer's disease. *Scientific Reports*, 13(1):10299, 2023.

[63] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

[64] Dominik Klepl, Min Wu, and Fei He. Graph neural network-based eeg classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[65] Wanzeng Kong, Min Qiu, Menghang Li, Xuanyu Jin, and Li Zhu. Causal graph convolutional neural network for emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1686–1693, 2022.

[66] Supavit Kongwudhikunakorn, Suktipol Kiatthaveephong, Kamonwan Thanontip, Pitshaporn Leelaarporn, Pathitta Dujada, Tohru Yagi, Vorapun Senanarong, Wanumaidah Saengmolee, and Theerawit Wilaiprasitporn. Psd-cnn approach for subject independent dementia recognition from eeg signals. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 588–594. IEEE, 2024.

[67] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[68] Kiwamu Kudo, Kamalini G Ranasinghe, Hirofumi Morise, Faatimah Syed, Kensuke Sekihara, Katherine P Rankin, Bruce L Miller, Joel H Kramer, Gil D Rabinovici, Keith Vossel, et al. Neurophysiological trajectories in alzheimer's disease progression. *Elife*, 12:RP91044, 2024.

[69] Zen J Lau, Tam Pham, SH Annabel Chen, and Dominique Makowski. Brain entropy, fractal dimensions and predictability: A review of complexity measures for eeg in healthy and neuropsychiatric populations. *European Journal of Neuroscience*, 56(7):5047–5069, 2022.

[70] 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.

[71] 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.

[72] Yang Li, Yang Yang, Qinghe Zheng, Yunxia Liu, Hongjun Wang, Shangling Song, and Penghui Zhao. Dynamical graph neural network with attention mechanism for epilepsy detection using single channel eeg. *Medical & Biological Engineering & Computing*, 62(1): 307–326, 2024.

[73] Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*, 2022.

[74] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024.

[75] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9881–9893. Curran Associates, Inc., 2022.

[76] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[77] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, March 2024.

[78] I Loshchilov. Decoupled weight decay regularization, 2017.

[79] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. *arXiv preprint arXiv:2209.07027*, 2022.

[80] Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024.

[81] N Malek, MR Baker, C Mann, and JJANS Greene. Electroencephalographic markers in dementia. *Acta Neurologica Scandinavica*, 135(4):388–393, 2017.

[82] Catriona D McCullagh, David Craig, Stephen P McIlroy, and A Peter Passmore. Risk factors for dementia. *Advances in psychiatric treatment*, 7(1):24–31, 2001.

[83] Christina Micanovic and Suvankar Pal. The diagnostic utility of eeg in early-onset dementia: a systematic review of the literature with narrative analysis. *Journal of Neural Transmission*, 121:59–69, 2014.

[84] Andreas Miltiadous, Emmanouil Gionanidis, Katerina D Tzimourta, Nikolaos Giannakeas, and Alexandros T Tzallas. Dice-net: a novel convolution-transformer architecture for alzheimer detection in eeg signals. *IEEE Access*, 2023.

[85] Yurui Ming, Weiping Ding, Danilo Pelusi, Dongrui Wu, Yu-Kai Wang, Mukesh Prasad, and Chin-Teng Lin. Subject adaptation network for eeg data analysis. *Applied Soft Computing*, 84:105689, 2019.

[86] Aslan Modir, Sina Shamekhi, and Peyvand Ghaderyan. A systematic review and methodological analysis of EEG-based biomarkers of Alzheimer's disease. *Measurement*, 220:113274, October 2023. ISSN 02632241. doi: 10.1016/j.measurement.2023.113274.

[87] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pages 238–253. PMLR, 2020.

[88] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pages 238–253. PMLR, 2020.

[89] Santiago Morales and Maureen E Bowers. Time-frequency analysis methods and their application in developmental eeg data. *Developmental cognitive neuroscience*, 54:101067, 2022.

[90] D.V. Moretti, C. Miniussi, G.B. Frisoni, C. Geroldi, O. Zanetti, G. Binetti, and P.M. Rossini. Hippocampal atrophy and EEG markers in subjects with mild cognitive impairment. *Clinical Neurophysiology*, 118(12):2716–2729, December 2007. ISSN 13882457. doi: 10.1016/j.clinph.2007.09.059.

[91] Favour Nerrise, Qingyu Zhao, Kathleen L. Poston, Kilian M. Pohl, and Ehsan Adeli. An Explainable Geometric-Weighted Graph Attention Network for Identifying Functional Networks Associated with Gait Impairment. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14221, pages 723–733. Springer Nature Switzerland, Cham, 2023. ISBN 9783031438943 9783031438950.

[92] Christopher Neves, Yong Zeng, and Yiming Xiao. Parkinson's disease detection from resting

state eeg using multi-head graph structure learning with gradient weighted graph attention explanations. *arXiv preprint arXiv:2408.00906*, 2024.

[93] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers, March 2023.

[94] NIH. National Institute of Environmental Health Sciences: Neurodegenerative Diseases, 2022.

[95] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[96] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[97] Saarang Panchavati, Corey Arnold, and William Speier. MENTALITY: A MAMBA-BASED APPROACH TOWARDS FOUNDATION MODELS FOR EEG. 2024.

[98] Sharaj Panwar, Paul Rad, Tzyy-Ping Jung, and Yufei Huang. Modeling eeg data distribution with a wasserstein generative adversarial network to predict rsvp events. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(8):1720–1730, 2020.

[99] Badri N. Patro and Vijay S. Agneeswaran. SiMBA: Simplified Mamba-Based Architecture for Vision and Multivariate Time series, March 2024.

[100] Peizhen Peng, Liping Xie, Kanjian Zhang, Jinxia Zhang, Lu Yang, and Haikun Wei. Domain adaptation for epileptic eeg classification using adversarial learning and riemannian manifold. *Biomedical Signal Processing and Control*, 75:103555, 2022.

[101] Darwin Saire Pilco and Adín Ramírez Rivera. Graph learning network: A structure learning algorithm. *arXiv preprint arXiv:1905.12665*, 2019.

[102] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger

convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR, 2023.

[103] Mohamed Radwan, Pedro G Lind, and Anis Yazidi. An interpretable graph based model for classification of eeg using directional functional connectivity. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[104] Alireza Rafiei, Rasoul Zahedifar, Chiranjibi Sitaula, and Faezeh Marzbanrad. Automated detection of major depressive disorder with eeg signals: a time series classification using deep learning. *IEEE Access*, 10:73804–73817, 2022.

[105] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning. *arXiv preprint arXiv:2304.04902*, 2023.

[106] Alexander P. Rockhill, Nicko Jackson, Jobi George, Adam Aron, and Nicole C. Swann. "uc san diego resting state eeg data from patients with parkinson's disease", 2021.

[107] Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020.

[108] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. A Gentle Introduction to Graph Neural Networks, 2021.

[109] Angela M Sanford. Mild cognitive impairment. *Clinics in geriatric medicine*, 33(3):325–337, 2017.

[110] Jorge E. Santos Toural, Arquímedes Montoya Pedrón, and Enrique J. Marañón Reyes. A new method for classification of subjects with major cognitive disorder, Alzheimer type, based on electroencephalographic biomarkers. *Informatics in Medicine Unlocked*, 23:100537, 2021. ISSN 23529148. doi: 10.1016/j.imu.2021.100537.

[111] Gita Sarafraz, Armin Behnamnia, Mehran Hosseinzadeh, Ali Balapour, Amin Meghrazi, and

Hamid R Rabiee. Domain adaptation and generalization of functional medical data: A systematic survey of brain data. *ACM Computing Surveys*, 56(10):1–39, 2024.

[112] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[113] Michael Schöll, Inge MW Verberk, Marta Del Campo, Constance Delaby, Joseph Therriault, Joyce R Chong, Sebastian Palmqvist, and Daniel Alcolea. Challenges in the practical implementation of blood biomarkers for alzheimer's disease. *The Lancet Healthy Longevity*, 5 (10), 2024.

[114] Sena Yagmur Sen, Ozlem Karabiber Cura, Gulce Cosku Yilmaz, and Aydin Akan. Classification of alzheimer's dementia eeg signals using deep learning. *Transactions of the Institute of Measurement and Control*, page 01423312241267046, 2024.

[115] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2022.

[116] Gourav Siddhad, Sayantan Dey, and Partha Pratim Roy. DrowzEE-G-Mamba: Leveraging EEG and State Space Models for Driver Drowsiness Detection, August 2024. arXiv:2408.16145 [cs].

[117] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[118] 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.

[119] 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.

[120] Igor Stancin, Mario Cifrek, and Alan Jovic. A review of eeg signal features and their application in driver drowsiness detection systems. *Sensors*, 21(11):3786, 2021.

[121] Mingyi Sun, Weigang Cui, Shuyue Yu, Hongbin Han, Bin Hu, and Yang Li. A dual-branch dynamic graph convolution based adaptive transformer feature fusion network for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 13(4):2218–2228, 2022.

[122] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[123] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[124] Siyi Tang, Jared A Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. *arXiv preprint arXiv:2104.08336*, 2021.

[125] Siyi Tang, Jared A Dunnmon, Qu Liangqiong, Khaled K Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel L Rubin. Modeling multivariate biosignals with graph neural networks and structured state space models. In Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 50–71. PMLR, 22 Jun–24 Jun 2023.

[126] Siyi Tang, Jared A Dunnmon, Qu Liangqiong, Khaled K Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel L Rubin. Modeling multivariate biosignals with graph neural networks and structured state space models. In *Conference on Health, Inference, and Learning*, pages 50–71. PMLR, 2023.

[127] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, and Jing Jiang. Omni-scale cnns: a simple and effective kernel size configuration for time series classification. *arXiv preprint arXiv:2002.10061*, 2020.

[128] Eduardo Tolosa, Alicia Garrido, Sonja W Scholz, and Werner Poewe. Challenges in the diagnosis of Parkinson's disease. *The Lancet Neurology*, 20(5):385–397, May 2021. ISSN 14744422.

[129] Ke Tran, Arianna Bisazza, and Christof Monz. The importance of being recurrent for modeling hierarchical structure. *arXiv preprint arXiv:1803.03585*, 2018.

[130] Xuan-The Tran, Linh Le, Quoc Toan Nguyen, Thomas Do, and Chin-Teng Lin. EEG-SSM: Leveraging State-Space Model for Dementia Detection, July 2024. arXiv:2407.17801 [cs].

[131] E Van Diessen, T Numan, E Van Dellen, AW Van Der Kooi, M Boersma, D Hofman, R Van Lutterveld, BW Van Dijk, ECW Van Straaten, A Hillebrand, et al. Opportunities and methodological challenges in eeg and meg resting state functional brain network research. *Clinical Neurophysiology*, 126(8):1468–1481, 2015.

[132] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[133] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[134] Julius Vetter, Jakob H Macke, and Richard Gao. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *Patterns*, 5(9), 2024.

[135] Julius Vetter, Jakob H Macke, and Richard Gao. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. Technical Report 9, 2024.

[136] Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

[137] Aaron R Voelker and Chris Eliasmith. Improving spiking dynamical networks: Accurate delays, higher-order synapses, and time cells. *Neural computation*, 30(3):569–609, 2018.

[138] Zihao Wang and Lei Wu. Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[139] Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. Self-supervised learning for electroencephalogram: A systematic survey. *arXiv preprint arXiv:2401.05446*, 2024.

[140] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[141] Xiaoxiao Yang and Ziyu Jia. Spatial-Temporal Mamba Network for EEG-based Motor Imagery Classification, September 2024. arXiv:2409.09627 [cs].

[142] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

[143] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[144] Chao Zhang, Weirong Cui, and Jingjing Guo. MSSC-BiMamba: Multimodal Sleep Stage Classification and Early Diagnosis of Sleep Disorders with Bidirectional Mamba, May 2024. arXiv:2405.20142 [cs].

[145] Jiahao Zhang, Haifeng Lu, Lin Zhu, Huixia Ren, Ge Dang, Xiaolin Su, Xiaoyong Lan, Xin Jiang, Xu Zhang, Jiansong Feng, Xue Shi, Taihong Wang, Xiping Hu, and Yi Guo. Classification of Cognitive Impairment and Healthy Controls Based on Transcranial Magnetic Stimulation Evoked Potentials. *Frontiers in Aging Neuroscience*, 13:804384, December 2021. ISSN 1663-4365. doi: 10.3389/fnagi.2021.804384.

[146] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[147] Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9263–9275. Curran Associates, Inc., 2020.

[148] Yanqiao Zhu, Weizhi Xu, Jinghao Zhang, Yuanqi Du, Jieyu Zhang, Qiang Liu, Carl Yang, and Shu Wu. A survey on graph structure learning: Progress and opportunities. *arXiv preprint arXiv:2103.03036*, 2021.