h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hochschule Darmstadt

– Fachbereiche Mathematik und
Naturwissenschaften & Informatik –

A Representative Subset of EEG Datasets for Generalizable Deep Learning Models

Abschlussarbeit zur Erlangung des akademischen Grades

Master of Science (M.Sc.)

vorgelegt von

Phuc Bao Nhi Nguyen

Matrikelnummer: 1118550

Referent	:	Prof. Dr. Timo Schürg
Korreferent	:	Prof. Dr. Florian Heinrichs

DECLARATION

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Darmstadt, December 18, 2024

Phuc Bao Nhi Nguyen

ABSTRACT

Electroencephalography (EEG) is a powerful neuroimaging technique widely used across various applications. Despite its advantages of non-invasiveness, portability, and high temporal resolution, EEG analysis remains challenging due to its complexity, high dimensionality, and low signal-to-noise ratio. Deep Learning (DL) models have shown significant potential in addressing these challenges by automatically extracting meaningful features from raw EEG signals and achieving high classification accuracy. However, most existing DL models are task-specific, and their generalizability across diverse EEG datasets and domains remains an open question.

This thesis investigates the feasibility of constructing a small yet representative subset of EEG datasets that encapsulates the diversity of a broader dataset collection. Such a subset would support the development and evaluation of generalizable DL models that perform robustly across various EEG classification tasks without requiring domain-specific adjustments, while also significantly reducing computational costs. Additionally, the thesis explores the generalizability and compatibility of state-of-the-art end-to-end DL models across multiple EEG datasets spanning different domains.

11 end-to-end DL models were trained and tested across 17 EEG datasets. A principled approach to subset selection was proposed, combining Ridge regression with correlation analysis to predict the median F_1-score of models based on the selected subset. The final subset should minimize prediction error while ensuring diversity and avoiding redundancy.

The findings indicate that none of the selected DL models demonstrated outstanding generalizability across all EEG datasets, reflecting limitations in their ability to adapt to diverse domains. The representative subset, comprising SEED, DREAMER-Valence, DREAMER-Arousal, EEGMAT, and CHB-MIT datasets, demonstrated a low prediction error and weak pairwise correlations, highlighting its diversity. This work provides a foundational framework for advancing generalizable DL models for EEG classification tasks.

Keywords: Electroencephalography (EEG), Deep Learning (DL), generalizability, EEG datasets, task-specific models, subset selection, representativeness, Ridge regression, correlation analysis, classification accuracy, diversity, computational efficiency, end-to-end deep learning models, compatibility, median *F*1-score.

ZUSAMMENFASSUNG

Elektroenzephalographie (EEG) ist eine leistungsstarke Neuroimaging-Technik, die in vielen Anwendungen eingesetzt wird. Trotz ihrer Vorteile, wie Nichtinvasivität, Portabilität und hoher zeitlicher Auflösung, bleibt die Analyse von EEG-Daten aufgrund ihrer Komplexität, hohen Dimensionalität und des niedrigen Signal-Rausch-Verhältnisses herausfordernd. Deep-Learning-Modelle (DL) haben großes Potenzial gezeigt, diese Herausforderungen zu bewältigen, indem sie aussagekräftige Merkmale aus EEG-Daten extrahieren und hohe Klassifikationsgenauigkeiten erzielen. Dennoch sind die meisten DL-Modelle auf spezifische Aufgaben beschränkt, und ihre Generalisierbarkeit über verschiedene EEG-Datensätze und Domänen bleibt fraglich.

Diese Masterarbeit untersucht die Konstruktion eines kleinen, repräsentativen Subsets von EEG-Datensätzen, das die Diversität einer größeren Sammlung abbildet. Dieses Subset soll die Entwicklung generalisierbarer DL-Modelle fördern, die robust über verschiedene EEG-Klassifikationsaufgaben hinweg performen, ohne domänenspezifische Anpassungen zu benötigen, und gleichzeitig Rechenkosten reduzieren. Zudem wird die Generalisierbarkeit und Kompatibilität state-of-the-art End-to-End-DL-Modelle über verschiedene EEG-Domänen hinweg analysiert.

11 End-to-End-DL-Modelle wurden auf 17 EEG-Datensätzen trainiert und getestet. Ein systematischer Ansatz zur Auswahl eines Subsets wurde entwickelt, der Ridge-Regression mit Korrelationsanalysen kombiniert, um den Median-F_1-Score der Modelle basierend auf dem Subset vorherzusagen. Das finale Subset minimiert die Vorhersagefehler, gewährleistet Diversität und vermeidet Redundanz.

Die Ergebnisse zeigen, dass kein DL-Modell herausragende Generalisierbarkeit über alle EEG-Datensätze aufwies, was ihre begrenzte Anpassungsfähigkeit an verschiedene Domänen verdeutlicht. Das ausgewählte Subset, bestehend aus SEED, DREAMER-Valence, DREAMER-Arousal, EEGMAT und CHBMIT, zeigte geringe Vorhersageabweichungen und schwache Korrelationen, was seine Diversität unterstreicht. Diese Arbeit bietet einen Rahmen für die Entwicklung generalisierbarer DL-Modelle für EEG-Klassifikationsaufgaben.

Schlüsselwörter: Elektroenzephalographie (EEG), Deep Learning (DL), Generalisierbarkeit, EEG-Datensätze, aufgabenspezifische Modelle, Subset-Auswahl, Repräsentativität, Ridge-Regression, Korrelationsanalyse, Klassifikationsgenauigkeit, Diversität, Recheneffizienz, Kompatibilität, End-to-End-Deep-Learning-Modelle, Median-F_1-Score.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABKÜRZUNGSVERZEICHNIS

EEG	Electroencephalography
AE	Auto-encoder [Auto-encoder Networks]
VAE	Variational Auto-encoder [Variational Auto-encoder Networks]
CNN	Convolutional Neural Network [Convolutional Neural Networks]
GAN	Generative Adversarial Network [Generative Adversarial Networks]
DBN	Deep Belief Network [Deep Belief Networks]
FC	Fully Connected
LSTM	Long Short-Term Memory [Long Short-Term Memory Networks]
MLPs	Multi-Layer Perceptrons
RBM	Restricted Boltzmann Machine [Restricted Boltzmann Machines]
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
PSD	Power Spectral Density
STFT	Short-Time Fourier Transform
BCI	Brain-Computer Interface
ML	Machine Learning
DL	Deep Learning
SNR	Signal-to-Noise Ratio
DNN	Deep Neural Network [Deep Neural Networks]
ANN	Artificial Neural Network [Artificial Neural Networks]
NLP	Natural Language Processing
MSE	Mean Squared Error
RNN	Recurrent Neural Network [Recurrent Neural Networks]

MI	Motor Imagery
ICA	Independent Component Analysis
WT	Wavelet Transform
BLSTM	Bidirectional Long Short-Term Memory
MWL	Mental Workload
ErrPs	Error-related Potentials
ERPs	Event-related Potentials
RL	Reinforcement Learning
GRU	Gated Recurrent Unit

Part I

THESIS

1 INTRODUCTION

1.1 BACKGROUND AND CONTEXT

Electroencephalography (EEG) is a widely used neuroimaging technique that captures the electrical activity of the brain with high temporal resolution. Its non-invasive nature, portability, and cost-effectiveness make it a versatile tool for a wide range of applications in fields such as healthcare, brain science, and artificial intelligence. EEG is particularly prominent in Brain-Computer Interface (BCI) technology, which has been experiencing significant developments due to its potential to offer a direct communication channel that interconnects the human brain with the outside environment [129]. For instance, by using BCI, users can control the movement of a cursor on a computer screen simply by imagining left or right hand movements, corresponding to the desired direction [130].

Despite its wide applicability, certain inherent characteristics of EEG signals pose some challenges for effective analysis and classification [93]. EEG signals are complex, high-dimensional, and non-stationary, and they have a low Signal-to-Noise Ratio (SNR) in the temporal domain [96]. Therefore, various advanced Machine Learning (ML) and Deep Learning (DL) algorithms have been proposed to effectively process and decode such complex brain data. However, most existing DL models for EEG classification are designed for specific domains and tasks. While excelling in task-specific scenarios, they often struggle to generalize across diverse EEG datasets and tasks.

Inspired by the breakthroughs in the natural language processing (NLP) field, where powerful large language models have successfully demonstrated the exceptional ability to handle diverse tasks such as translation, chatbots, text generation, and creative writing, one might wonder whether similar generalization is possible for EEG data. Can DL models be developed that classify EEG signals across all domains without requiring domain-specific expertise or complex feature engineering?

This thesis investigates the generalizability of DL models for EEG classification tasks. A key challenge lies in the vast number of datasets across various domains, making it impractical to validate a model's effectiveness by evaluating it on every available dataset. To address this, an essential first step is to identify a representative subset of EEG datasets that encapsulates the inherent diversity of EEG signals across domains. Such a subset would serve as a foundation for developing and evaluating novel models capable of robust performance across a wide range of EEG tasks.

1.2 MOTIVATION AND PROBLEM STATEMENT

In recent years, significant research efforts have been put into the development of ML and DL models for EEG data classification. DL, in particular, has been extensively explored for a wide range of EEG classification problems, including emotion recognition, motor imagery, mental workload, seizure detection, sleep stage scoring, event related potential detection, and brain disorders detection [26, 93]. These advancements highlight the potential of DL for decoding complex EEG signals and enabling applications in neuroscience, healthcare, and beyond.

Despite these successes, most existing studies have focused on developing domain-specific DL models, which were exclusively applied to datasets within a particular domain. While these models have achieved impressive performance within their respective domains, their ability to generalize to diverse EEG classification tasks beyond their specific domains remains largely unproven. For example, a model excelling at emotion recognition might not effectively transfer to seizure detection due to the unique characteristics of each task.

This domain-specificity hinders the flexibility and generalization capability of current EEG-based technologies [93]. Consequently, designing generalization-oriented DL models for EEG analysis is highly desirable. Such models could capture the underlying patterns in EEG signals across various contexts, regardless of the specific task or domain. This would also enhance the scalability and adaptability of EEG analysis, opening doors to applications in various fields. To achieve this objective, a comprehensive understanding of the generalizability of existing DL models is essential. This would pave the way for the future development of generalization-oriented approaches.

However, when working with extensive EEG datasets encompassing diverse contexts and domains, achieving generalizability poses significant challenges. The landscape of EEG datasets spans across numerous types of EEG signals, ranging from spontaneous EEG (e.g., emotion, motor imagery, mental disorders, sleep stages) to evoked potentials (EP) (e.g., event-related potentials (ERP), steady-state evoked potentials (SSEP)) [136]. Developing DL models and generating results on this vast and diverse data would require substantial computational resources and time investment.

One promising solution lies in the identification of a small yet representative subset of EEG datasets that captures the most information with low redundancy and effectively represents the entire spectrum of EEG activity. Such a subset would enable systematic testing and training of DL models while significantly reducing computational costs. By focusing on a **coreset** of datasets, researchers can evaluate the generalizability of DL models more efficiently, ultimately fostering a more unified understanding of model performance on EEG data.

This thesis undertakes an initial exploration of a principled approach to selecting a representative subset of EEG datasets across multiple domains, inspired by the approach proposed by Aitchison, Sweetser, and Hutter [5].

1.3 OBJECTIVES

The primary objective of this thesis is to investigate the feasibility of constructing a coreset that captures the inherent diversity of EEG signals. A well-designed subset could serve as a standardized benchmark, facilitating the development of robust DL models that adapt to variations across different EEG signal types and tasks. This approach has the potential to significantly enhance the generalizability of DL models, paving the way for future advancements in models capable of accurately classifying EEG data across diverse domains.

To achieve this goal, the following specific objectives are outlined:

- Dataset and Model Collection: Gather a diverse collection of EEG datasets spanning various domains to build a foundation for subset selection. Identify suitable DL architectures relevant to this study.
- Performance Evaluation: Train multiple models on the selected EEG datasets to assess performance metrics, gaining insights into task variability and model generalization.
- Systematic Subset Selection: Develop a systematic method for selecting a small yet representative subset from a large collection of EEG datasets, ensuring minimal redundancy while retaining the most informative features.
- Subset Identification: Demonstrate the approach by identifying an example subset of EEG datasets as a proof of concept.

As secondary objectives, this study aims to explore two interrelated aspects of deep learning for EEG analysis:

1. **Generalizability**: Evaluate the extent to which state-of-the-art DL architectures generalize across various domains by assessing their performance on tasks outside their respective domains.
2. **Compatibility**: Analyze the compatibility of different DL architectures with various EEG data types, offering insights into optimal pairing of architectures with specific EEG signal domains.

1.4 RESEARCH QUESTION

The main research question of this study is:

> **"Which collection of EEG datasets can best represent the entire spectrum of EEG activity for evaluating and developing generalizable deep learning models?"**

This study hypothesizes that it is possible to identify a coreset of EEG datasets that is significantly smaller in size yet captures the most critical information and effectively represents the entire spectrum of EEG activity. This

coreset would provide a robust foundation for training and testing generalizable DL models in the future.

1.5 STRUCTURE

This thesis is organized into five chapters as follows:

- **Chapter 1: Introduction** – This chapter introduces the research topic, motivation, problem statement, objectives, research questions, and the overall structure of the thesis.
- **Chapter 2: Theoretical Background** – Provides an overview of existing research on EEG analysis, DL models, and the challenges of generalization. It identifies gaps in the literature that this study aims to address.
- **Chapter 3: Methodology** – Details the research methodology, including:
  - **Conceptual Framework**: Describes the theoretical foundation and guiding principles of the study.
  - **Implementation**: Outlines the practical steps involved in dataset collection, preprocessing, model evaluation, and subset selection.
- **Chapter 4: Results** – Presents the findings of the study, including model performance metrics, an analysis of results, insights into generalization, and an evaluation of the subset selection procedure.
- **Chapter 5: Discussion** – Interprets the findings, discusses limitations, and provides recommendations for future research.

2 THEORETICAL BACKGROUND

2.1 ELECTROENCEPHALOGRAM

Figure 2.1 illustrates the raw EEG signals, displaying amplitude variations over time, recorded from the DREAMER dataset across different subjects and channels. EEG data typically appear noisy, with high-frequency oscillations mixed with transient peaks. Additionally, significant differences in amplitude ranges are observed across subjects and channels, reflecting the inherently dynamic nature of EEG signals. Such variability arises from individual differences in brain activity patterns, the brain regions being recorded (channel-specific), and also the presence of physiological or external artifacts.

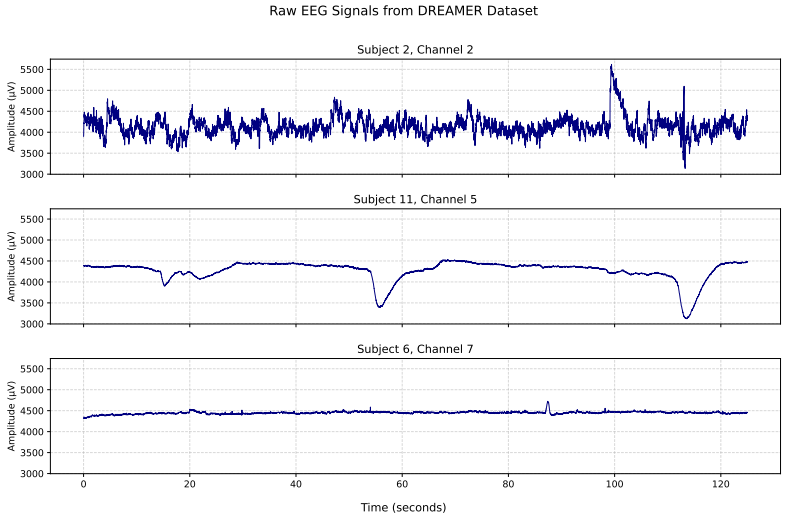

Figure 2.1: Raw EEG Data from the DREAMER Dataset

2.1.1 *Physiological Basis of EEG*

EEG is a non-invasive medical imaging technique that measures the electrical potentials generated by brain activity. The electroencephalogram is a visual representation of these alternating electrical signals, captured from the scalp using metal electrodes and a conductive medium [116]. Small Ag/AgCl disc-shaped electrodes are placed on different locations of the scalp, typically following the standardized 10-20 international electrode placement system (Figure 2.2), which ensures consistency in electrode positioning across different studies and subjects. Electrode positions are labeled with a letter indicating the brain region - F for frontal, C for central, T for temporal, P for parietal,

and O for occipital - followed by an odd or even number to specify the left or right hemisphere [109].

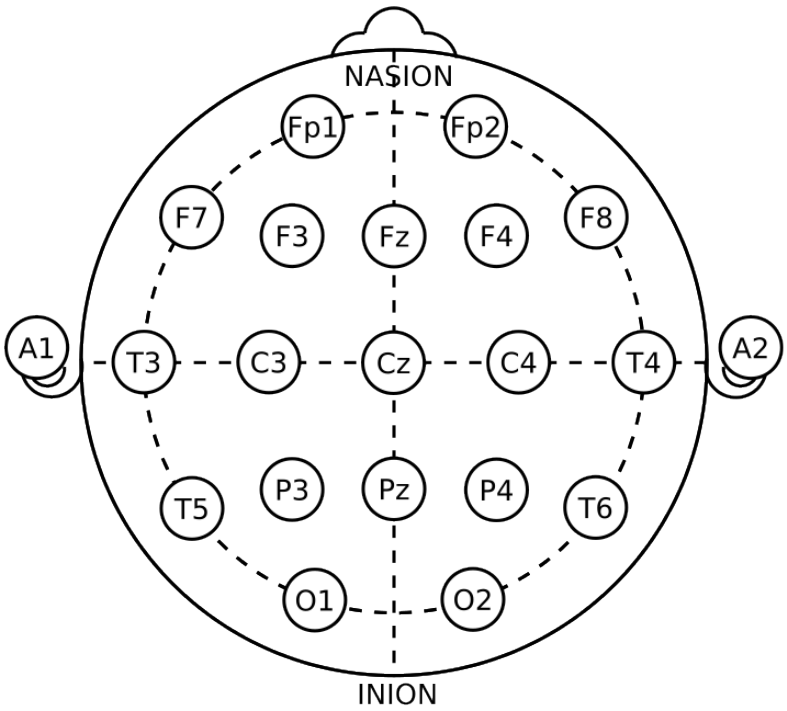

Figure 2.2: 21 Electrodes of International 10-20 System for EEG [119].

Human brain functions are driven by complex neural activations and interactions, which produce electromagnetic signal dynamics as primary effects, alongside secondary hemodynamic and metabolic changes [45]. Electrical activity is generated by the exchange of electrochemical signals between pyramidal neurons. The simultaneous and synchronous transmission of billions of tiny excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs) within large neural populations sum up, creating electrical fields strong enough to be detected from outside the head [25]. These signals thus reflect instantaneous neural currents and can be used to study a wide range of brain processes.

2.1.2 *Characteristics of EEG Signals*

EEG signals are characterized by their complex and dynamic nature through oscillatory patterns, the statistical properties can be in multiple domains: time, frequency, and spatial dimensions. These characteristics are used to extract important features to interpret brain activity.

2.1.2.1 *Time Domain Characteristics*

EEG signals measure voltage fluctuations over time and are characterized in the time domain by changes in signal amplitude and waveform duration relative to time [27]. These signals exhibit highly non-Gaussian and

non-stationary behavior, meaning their statistical properties can vary unpredictably over time. Key features in the time domain include basic statistical metrics (e.g., mean value, skewness and kurtosis), energy- and entropy-related features, Hjorth parameters [48], and counts of zero-crossings and local extrema [77]. Additionally, peaks, troughs, and transient events such as spikes or bursts, which may correspond to specific cognitive or pathological events, are also observed.

2.1.2.2 *Frequency Domain Characteristics*

EEG captures neural oscillations, which are rhythmic or repetitive electrical activities spanning a broad spectrum of frequencies [14]. Hence, frequency domain features are widely used in EEG research to analyze oscillatory and rhythmic patterns [22]. Frequency refers to the number of events occurring within a specified time period and is typically measured in hertz (Hz), where one hertz corresponds to one cycle per second [44]. These signals are typically decomposed into five major frequency bands, each associated with distinct brain functions and states: delta (1–4 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (>30 Hz) [107].

A key feature derived from frequency analysis is the power spectral density (PSD) which reflects the distribution of signal power across different frequencies [50]. However, a limitation of Power Spectral Density (PSD) is its inability to determine the precise timing of frequency-specific events, as it offers only overall power contributions in the frequency domain [44]. Additional metrics such as the energy and entropy of PSD, intensity-weighted mean frequency and bandwidth, edge frequency, and peak frequency are often extracted for analysis [77].

2.1.2.3 *Time-Frequency Characteristics*

A key limitation of the time domain analysis of EEG is its inability to reveal the underlying "frequency content" or detect changes in frequency patterns. Similarly, spectral analysis alone cannot accurately identify the time localization of specific frequency components [50]. To address these limitations, EEG signals are often analyzed jointly in both time and frequency domains, reflecting their transient and dynamic nature. This combination allows researchers to observe changes in the power of specific frequency bands over time, offering insights into transient neural processes [24]. Time-frequency analysis techniques, such as wavelet transforms (WT) and Short-Time Fourier Transform (STFT), decompose signals into smaller components within short time windows, enabling the simultaneous time and frequency analysis.

2.1.2.4 *Spatial Characteristics*

EEG signals can be recorded from almost any location on the scalp, thanks to the wide spatial arrangement of electrodes, which provides valuable insights

into regional brain activity. While time and frequency features primarily reflect neural activity within a single EEG channel, spatial features capture the interdependencies and functional connectivity between multiple channels [131].

These spatial features illustrate how different areas of the brain communicate and work together to support various cognitive and neural processes. Several neural activities involve distributed circuits rather than isolated brain regions [131], and disruptions in these intricate networks have been linked to cognitive deficits in disorders such as autism, schizophrenia, and major depressive disorder [134]. EEG's spatial structures enable the study of network properties and the functional organization of brain regions [63]. Key spatial features include correlation, coherence, phase synchronization, mutual information, and the topological structure of EEG channels.

2.1.2.5 *Noise and Artifacts*

EEG signals are often contaminated by noise and artifacts from various sources, which originate from non-cerebral activities [107]. Artifacts can be broadly categorized into subject-related and technique-related types. Subject-related artifacts arise from undesired physiological signals, including muscle activity (EMG), cardiac activity (ECG), eye movements, minor body movements, and sweating. Technique-related artifacts, on the other hand, are caused by external environmental factors, such as power line interference (50/60 Hz), impedance fluctuations, cable movements, broken wire contacts, excessive or dried electrode paste/jelly, and low battery levels [50]. Artifacts typically exhibit higher amplitudes and distinct morphologies compared to true brain activity. These can distort the actual neural signals, significantly affecting their analysis and interpretation of EEG data. As a result, artifact removal is a critical step in the preprocessing pipeline.

2.1.3 *Strengths and Limitations of EEG*

2.1.3.1 *Advantages*

The primary advantage of EEG is its highly precise temporal resolution. Most cognitive processes are fast, dynamic, and unfold in temporal sequences that occur within fractions of a second. The rapid propagation of electrical fields allows EEG to capture complex neural patterns at the millisecond level after a stimulus is triggered [24]. EEG also offers a multidimensional view of brain activity, capturing information across several dimensions: time, space, frequency, power (the strength of frequency-band-specific activity), and phase (the timing of the activity) [24]. This multidimensionality enables EEG to offer rich insights into brain function as well as dysfunction, supporting a variety of analytical approaches.

Additionally, EEG is a popular brain-imaging tool due to its ability to directly measure neural activity, its relatively low cost, non-invasive nature, and portability [3]. These advantages have led to its widespread application

across numerous domains, highlighting EEG's versatility and importance in both research and practical settings. In clinical practice, injuries or abnormalities in the brain can be detected using EEG. It serves as a first-line method for investigating sleep patterns or epilepsy, as well as for diagnosing a range of neurological and psychiatric disorders, such as attention deficit hyperactivity disorder (ADHD), Alzheimer's disease, brain tumors, Parkinson's disease, and schizophrenia [93, 109].

EEG is also instrumental in research involving neural engineering, neuroscience, psychology and biomedical engineering, as it enables the study of brain activity patterns associated with cognitive processes [13], emotional responses [88], and sensory perceptions [79]. Furthermore, an EEG system acts as a bridge between the brain and external devices, allowing the observation of brain's responses to specific stimuli events and the interpretation of certain aspects of a person's cognitive state [123]. Therefore, EEG is widely applied in brain-computer interface (BCI) as assistive technological solutions for individuals with disabilities and brain-controlled rehabilitation devices for patients with strokes and other neurological deficits [3].

2.1.3.2 *Disadvantages*

Although EEG is a powerful and informative brain-imaging technique, its effectiveness in analysis and classification tasks is constrained by certain limitations. First, EEG data suffer from low spatial resolution, providing only a coarse measure of brain activity [50]. This limitation arises from the non-invasive nature of EEG sensors, which measure neural activity at a distance from its sources. As the brain's electric fields pass through layers of tissue, such as the skull and scalp, they become distorted, and electrical potentials spread through the brain's conductive medium [93], [136]. As a result, signals from multiple brain regions are mixed together, making it difficult to determine the exact origin of brain activity or distinguish activity from closely spaced regions.

Secondly, EEG signals have very a low SNR, with weak amplitudes typically ranging from 10 to 300 μV, making them easily contaminated with various physiological and electrical noises [117]. To address this, advanced filtering and noise reduction techniques are adopted to remove artifacts, consequently reducing their impact and extract true brain activity from the recorded signals.

Thirdly, EEG signals are nonlinear and non-stationary, meaning their statistical properties vary over time [93]. This dynamic nature creates challenges in extracting consistent and generalizable features, often hindering the performance of models trained on such data, as the learned patterns may fail to predict signals recorded at a different time from the same individual. Most traditional methods struggle to capture the complexity, instability, and irregularity of EEG signals, as they implicitly assume stationarity in the data [128].

Finally, significant differences in brain activity patterns between individuals (inter-subject variability) and within the same individual under vary-

ing conditions (intra-subject variability) pose additional challenges for EEG recordings [97]. This high variability further limits the practical applicability of EEG, as well-trained models often fail to perform consistently across multiple subjects or even on the same subject over time.

Together, these factors increase the complexity of preprocessing, modeling, and extracting meaningful insights from EEG data, highlighting the need for advanced techniques to overcome these challenges.

2.2 DEEP LEARNING

2.2.1 *Fundamentals of Deep Neural Networks*

DL is a subset of ML that implements multilayered neural networks to learn the hierarchical and complex patterns from input data [40]. DL does not require any human-designed rules; instead, it leverages massive amounts of data to map the given inputs to specific labels [62]. Recent advancements in DL have greatly outperformed its predecessors. Conventional ML methods typically involve multiple sequential steps, including preprocessing, feature extraction, feature selection, learning, and classification. Errors or biases at any of these stages can adversely impact model performance. In contrast, DL automates the feature learning process with an aim to extract discriminative data representations with minimal human intervention or field knowledge. Moreover, it integrates feature learning and classification into a unified process, streamlining the workflow compared to traditional approaches [62].

2.2.1.1 *The Basic Architecture of DNN*

At its core, DL builds upon the concept of single-layer neural networks, also known as perceptrons [90]. Introduced in the 1950s, the *perceptron* is a simple computational model that functions as a binary classifier. It feeds a set of inputs, each associated with a weight, computes their weighted sum, and applies a threshold activation function to determine the output. A bias term is added to shift the decision boundary, providing greater flexibility to position it optimally within the input space [4].

Additionally, the perceptron utilizes a basic error-driven learning rule to minimize the misclassification error in prediction, quantified using a loss function. Softmax outputs with cross-entropy loss are commonly used for discrete predictions, while linear outputs with squared loss are preferred for real-valued predictions. During training, the parameters of the model, such as weights and biases, are iteratively adjusted using gradient descent, which minimizes the loss by updating the parameters in the direction of the steepest descent, thereby improving prediction accuracy.

By stacking multiple perceptrons into interconnected layers and introducing non-linear activation functions (e.g., ReLU, sign, sigmoid, tanh) [4], *artificial neural networks (ANNs)* were developed to model non-linear and more complex data structures. The additional intermediate layers between input and output are known as *hidden layers* because their computations are not

directly observable. The architecture of Artificial Neural Networks (ANNs) is typically *feedforward*, as information extracted from input data flows in a single direction through the hidden layers to the output layer. Each neuron in the network applies weights and biases to its input and transforms it using an activation function.[40].

ANNs with multiple hidden layers are referred to as Deep Neural Networks (DNNs). These networks are designed to learn high-level features with greater complexity and abstraction compared to shallower networks. Basically, DNNs consist of three main types of layers: an input layer, multiple hidden layers, and an output layer.

- Input layer: The first layer receives the initial raw data as input. Each neuron in this layer represents a feature or attribute of the input data.
- Hidden layer: Positioned between the input and output layers, hidden layers perform the primary computations. They transform the input data into increasingly abstract representations through weighted connections and activation functions [80].

  The term "deep" in DNNs refers to the network's depth, defined by the number of hidden layers it contains [42]. Adding more hidden layers increases the depth of the network, enabling it to capture higher-dimensional patterns and representations in the data. The hierarchical features extracted in these layers empower DNNs to handle sophisticated real-world problems, such as Natural Language Processing (NLP), image classification, or object detection.
- Output layer: The final layer generates predictions based on the transformed data, tailored to the specific task. For classification tasks, the output represents the probability of belonging to each class, while in regression tasks, it provides a continuous numerical value output. The number of neurons in the output layer corresponds to the number of target classes in classification problems.

2.2.1.2 *Training Process in DNN*

One of the defining features of DNNs is their capability of learning the dynamics embedded in data from the presentation of patterns. The iterative learning process in DNNs follows a systematic sequence of steps designed to adjust the model's parameters — weights and biases — so that it can minimize the error between the predicted output and the actual target output. This process repeats over multiple iterations, or epochs, until the model converges to an optimal solution. The process can be broken down into five key steps [23]:

1. **Initialization of Weights and Bias Parameters**

   At the start of the training process, the weights and bias parameters of the neural network are initialized. Common strategies include random initialization for weights and setting biases to zero or small non-zero

values [82]. Data-driven initialization, where the initialization point is derived from data statistics [65]. These initial values provide the starting point for the learning process, and the choice of initialization can significantly affect the speed and effectiveness of training [82].

2. **Forward Propagation**

   In this step, the input data passes through the network layer by layer, with computations propagating forward using the current set of weights, biases, and activation functions, finally generating predictions in the output layer [4]. Each layer performs the following operations:

   $$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)}, \quad a^{(l)} = \phi(z^{(l)})$$

   where:

   - $z^{(l)}$: Weighted input to the neuron.
   - $W^{(l)}$: Weight matrix for layer l.
   - $b^{(l)}$: Bias vector.
   - $a^{(l-1)}$: Activations from the previous layer.
   - $\phi(\cdot)$: Activation function.

   The final output layer produces predictions, denoted $\hat{y}$.

3. **Loss Function Computation**

   The loss function quantifies the error between the predicted output ($\hat{y}$) and the actual target (y_{true}). The choice of loss function depends on the task:

   - For regression: Mean Squared Error (MSE).
   - For classification: Cross-Entropy Loss.

   Mathematically:

   $$L(\hat{y}, y_{\text{true}}) = \frac{1}{N}\sum_{i=1}^{N} \text{Loss}(\hat{y}_i, y_{\text{true},i})$$

   where N is the number of samples. This loss value guides the adjustments needed in the weights and biases by computing its derivative with respect to the output.

4. **Backpropagation** In single-layer perceptrons, updating the parameters is straightforward with simple gradient computation because the loss function can be expressed as a direct function of the weights. However, in multi-layer networks, the loss is a complicated composition function involving the weights from previous layers [4], making gradient computation more challenging.

   The backpropagation algorithm efficiently addresses this problem by applying the chain rule of differential calculus to compute the gradients. Backpropagation is the process of propagating the error calculated by the loss function, backward through the network. The chain

rule calculates error gradients as summations of local gradient products along the various paths from a node to the output. Using this approach, the gradients of the loss function with respect to each weight ($W^{(l)}$) and bias ($b^{(l)}$) at layer l are determined as follows:

$$\frac{\partial L}{\partial W^{(l)}} = \frac{\partial L}{\partial a^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial z^{(l)}} \cdot \frac{\partial z^{(l)}}{\partial W^{(l)}}$$

$$\frac{\partial L}{\partial b^{(l)}} = \frac{\partial L}{\partial a^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial z^{(l)}} \cdot \frac{\partial z^{(l)}}{\partial b^{(l)}}$$

$$\frac{\partial L}{\partial a^{(l)}} = \frac{\partial L}{\partial z^{(l+1)}} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(l)}}$$

where:

- $z^{(l)}$: Weighted input to the neuron.
- $W^{(l)}$: Weight matrix for layer l.
- $b^{(l)}$: Bias vector.
- $a^{(l-1)}$: Activations from the previous layer.
- $\phi(\cdot)$: Activation function.

These gradients indicate the magnitude and direction in which each parameter should be adjusted to minimize the loss. Through repeated iterations over batches of training data across multiple epochs, backpropagation enables the effective updating of weights and biases until the neural network's performance reaches a satisfactory level or converges to a solution.

5. **Update of Weights and Bias Parameters**

   After each backpropagation step, the weights and biases are updated using an optimization algorithm, such as mini-batch gradient descent or Stochastic Gradient Descent (SGD) [110]. The update rule is:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial L}{\partial W^{(l)}}$$

$$b^{(l)} \leftarrow b^{(l)} - \eta \frac{\partial L}{\partial b^{(l)}}$$

   where η is the learning rate, controlling the step size of the updates. This step adjusts the parameters to reduce the loss in the next iteration.

Besides, the training process of DNNs also involves several other important steps, such as *dataset preparation*, where the data is usually split into three subsets to ensure robust evaluation. Training set is used to train the model

by adjusting weights and biases, while validation set is used during training to tune hyperparameters and monitor for overfitting. Last but not least, test dataset is used after training to evaluate the model's performance on unseen data.

Another key step is epochs and batches processing. An epoch is one complete pass through the training data. During each epoch, the data is divided into smaller subsets called batches. This enables efficient memory use and faster convergence.

2.2.1.3 *Regularization Techniques*

During the training process, the type and amount of input data directly affect the performance of DNNs. Insufficient training data can lead to issues such as *overfitting* or *underfitting* [83]. Overfitting occurs when a model achieves near-perfect predictions on the training data but fails to generalize to unseen test data [31]. This happens possibly because the model memorizes the training data and its corresponding outputs rather than learning the actual underlying trends. Overfitting is often observed as a notable gap between training and test data performance, especially in complex models trained on small datasets. Underfitting, on the other hand, arises when the model is unable to capture the data patterns, often due to an overly simplified model or poor-quality data, resulting in high error rates on both training and test sets [31].

Regularization techniques mitigate this issue by improving the model's generalization ability. They ensure balanced performance on both training and test datasets.

One common approach is *dropout*, which randomly deactivates some nodes and their connections during training, encouraging the model to learn independent features rather than relying too heavily on specific nodes [112]. A related method, *drop-weights*, deactivates connections between nodes instead of nodes themselves, promoting diverse learning pathways.

Data augmentation artificially increases the size of the training dataset by applying transformations such as rotations, flips, color adjustments, or sliding time windows. By simulating a larger dataset, this approach provides a broader representation of possible input variations and helps to reduce overfitting.

Batch normalization normalizes the output activations of each layer to follow a unit Gaussian distribution, minimizing internal covariance shift and improving training stability [11]. It also prevents the *vanishing gradient problem*, accelerates convergence, and reduces dependency on hyperparameter tuning.

2.2.2 *Common Architectures in Deep Learning*

DL has evolved with the development of various architectures tailored to different types of data and tasks. This section provides an overview of some widely-used DL architectures and their unique characteristics.

2.2.2.1 *Deep Belief Network*

Deep Belief Networks (DBNs) are probabilistic generative DL architectures that differ from traditional multilayer perceptrons (MLPs) in their weight initialization process [47]. Unlike Multi-Layer Perceptrons (MLPs), which initialize weights randomly, Deep Belief Networks (DBNs) utilize a greedy, layer-wise pre-training algorithm based on Restricted Boltzmann Machines (RBMs) to initialize their weights. Restricted Boltzmann Machines (RBMs) are stochastic neural networks that model the probability distribution of input data, using a two-layer architecture with visible and hidden units. DBNs are constructed by stacking multiple RBMs, where the output of one RBM serves as the input to the next.

Training a DBN involves two main steps: pre-training and fine-tuning [74]. During pre-training, an unsupervised learning approach is applied for each RBM in a bottom-up direction to extract features. This phase of DBNs enhances performance by initializing weights based on the input data structure, which brings them closer to the global optimum than random initialization. Once pre-training is complete, the DBN is fine-tuned using supervised learning with backpropagation in a top-down direction to further refine the network's parameters. DBNs are effective for processing unlabeled data and mitigating both overfitting and underfitting issues.

2.2.2.2 *Convolution Neural Networks*

Convolutional Neural Networks (CNNs) are a specialized class of DNNs designed to process data with a grid-like structure, such as images [66]. In the field of DL, Convolutional Neural Networks (CNNs) are among the most prominent algorithms, broadly recognized for their efficiency in pattern recognition and image processing tasks. Typically, a CNN architecture is composed of one or more stacked convolutional layers, each comprising three stages: the convolution stage, the detector stage, and the pooling stage [40].

The convolution stage involves applying convolutional filters to the original 2D input data, resulting in multiple feature maps. In detector stage, these output feature maps undergo a non-linear transformation using activation functions such as Sigmoid and Rectified Linear Unit (ReLU). This transformation introduces non-linearity, enabling the network to model complex patterns in the data. Finally, pooling operations such as Max Pooling or Average Pooling are applied to replace the output with summary statistics of nearby regions. Pooling introduces translational invariance to the representation and reduces the size of the input passed to the next convolutional layer or a fully connected layer, and preserves essential information.

According to [40], CNNs offer three key advantages: sparse interactions, parameter sharing, and equivariant representation. While traditional neural network layers use matrix multiplication, where each output unit interacts with every input unit, resulting in fully connected (FC) layers. In contrast,

CNNs employ smaller kernels than the input size, with only a limited number of weights connecting two adjacent layers.

Additionally, CNNs do not allocate separate weights for every neuron pair in neighboring layers. Instead, a single set of weights operates across all pixels of the input matrix, allowing the network to learn a unified set of parameters rather than distinct ones for each location. By leveraging shared weights and local connections, CNNs efficiently utilize the structure of 2D input data. This approach drastically reduces the number of weight parameters, thereby lowering the demand for parameter storage, simplifying the architecture, and accelerating the training process.

Another key property of CNNs is equivariance to translation, which refers to the ability of convolution to produce consistent feature maps that change predictably with transformations in the input. This property helps CNNs extract location-invariant features effectively.

2.2.2.3 *Recurrent Neural Networks*

Unlike feedforward neural networks, recurrent neural networks (RNNs) are characterized by their cyclic connections [105], where the output of a layer becomes the input to itself, forming a feedback loop. This structure enables Recurrent Neural Networks (RNNs) to have memory about previous states and integrates that into current computations. As a result, RNNs can process a sequence of inputs and generate a corresponding sequence of outputs, making them particularly well-suited for tasks that involve sequential and time-dependent data. They excel at capturing features and long-term dependencies in sequential data.

However, traditional RNNs often encounter the vanishing gradient problem during training, which limits their ability to learn long-term patterns effectively. Long Short-Term Memory (LSTM) networks [49] address this issue by incorporating multiplicative gates that regulate the flow of information. These gates enable constant error propagation through specialized units known as *memory cells*, allowing the network to retain important information over time. Stacking recurrent hidden layers further enhances the learning capacity of RNNs, enabling them to capture higher-level temporal patterns and complex dependencies.

2.2.2.4 *Auto-encoder Neural Networks*

Auto-encoder (AE) networks are widely used models for data compression and dimensionality reduction [85]. An Auto-encoder (AE) consists of two main components: an encoding network and a decoding network. The encoding network learns to represent the input data in a lower-dimensional latent space, capturing the most important features. The decoding network then reconstructs the original input data from this compressed representation. By minimizing reconstruction errors during training, AEs are designed to preserve the most salient features of the input data.

As an extension of traditional auto-encoder architectures, the Variational Autoencoder (VAE) [20] introduces a probabilistic approach to the latent representation. Instead of mapping input data to fixed encodings, Variational Auto-encoder (VAE) maps it to a probability distribution in the latent space, typically Gaussian. This probabilistic modeling enables the the generation of new data samples by sampling from the learned latent distribution, which makes AEs suitable for generative tasks and other applications requiring flexibility in latent space representation.

2.2.2.5 *Generative Adversarial Network*

As generative models, *generative adversarial networks* (GANs) are widely used for generating realistic data samples based on the learned probability distribution of the original dataset [95]. A Generative Adversarial Network (GAN) consists of two components: a generator and a discriminator. The generator aims to create data samples that replicate the distribution of the training data, while the discriminator's role is to distinguish between real data and generated data. These two models are trained adversarially, with the generator improving its ability to produce realistic samples as the discriminator refines its ability to differentiate real data from generated data.

The training process uses a minimax loss function, where the generator minimizes its loss by generating realistic samples, and the discriminator maximizes its accuracy in distinguishing between real and generated data [41]. As training progresses, the generator produces increasingly realistic data samples, while the discriminator becomes more accurate in identifying them. GANs are well-suited for tasks such as data augmentation, image synthesis, and modeling complex data distributions.

2.3 EEG ANALYSIS METHODS

The analysis of EEG signals contributes significantly to understanding brain activity [6, 88], diagnosing neurological disorders [118], and developing BCI [122]. EEG processing methodology has evolved from basic techniques, such as statistical comparisons, to advanced algorithms, including ML and DL.

The conventional approach of EEG analysis involves visually inspecting raw signals, measuring frequency and amplitude based on simple rules to detect transient features or anomalies [17, 27]. This manual monitoring method is often considered subjective, time-consuming and labor-intensive, as it requires specialized knowledge and experience of experts who need to be well-trained in EEG interpretation to produce reliable results [118]. Furthermore, the absence of clearly defined criteria makes the evaluation subjective, potentially leading to inconsistencies among different evaluators.

The task becomes more challenging when it comes to evaluating large amounts of high-dimensional EEG data. To address this problem and achieve quicker and more accurate results with a higher degree of automation, computer-aided technologies are increasingly utilized in EEG signal processing and analysis. Automated classification also helps minimize human error while

providing consistent and objective assessment for research and practical applications.

In recent years, the rapid progress in data-driven technologies has led researchers to apply new and efficient ML and DL algorithms more frequently to EEG decoding, establishing them as transformative tools in automating EEG classification. These advancements build upon a structured EEG processing pipeline, which typically consists of three key steps: preprocessing, feature extraction, and classification. [102]

2.3.1 *Preprocessing*

Initially, raw EEG signals undergo *preprocessing* with an attempt to improve signal quality without changing any of the data. Several preprocessing steps are applied, such as artifact removal, filtering, re-referencing, segmenting, and feature scaling[50].

The process begins with *artifact removal*, where undesired signals, such as physiological artifacts and external noise are eliminated. This can be performed manually or automatically using techniques, such as regression-based methods and independent component analysis (ICA). Next, *filtering* is applied to minimize irrelevant or noisy frequency components. Common techniques include high-pass, low-pass, and bandpass filtering, which isolates signals within a specific frequency range (e.g., 4–45 Hz), and notch filtering, which removes power line interference at 50/60 Hz.

Re-referencing is the process of changing the reference channel(s), which serve as the baseline to compare signals from the other channels. EEG signals are inherently referential, meaning they are measured as the voltage difference between a recording electrode and a reference electrode [24]. However, the initial choice of reference can cause some biases, making certain brain regions appear more or less active than they truly are. Re-referencing to a different site or method, such as linked mastoids, bipolar referencing, or common average referencing, helps reduce this bias and provides a more balanced representation of scalp activity.

In the next step, continuous EEG recordings are divided into smaller *epochs* through segmentation. These segments are usually time-locked to the onset of specific events to identify changes in EEG activity associated with sensory stimulation or cognitive tasks [50]. Trials contaminated by artifacts and poorly functioning channels are excluded to maintain data quality.

Last but not least, *feature scaling* is applied to normalize the data and ensure consistency across channels and trials. This step prevents excessive fluctuations in the range of raw data values, reduces distortion caused by amplitude differences and improves the overall reliability of subsequent analyses.

2.3.2 *Feature Extraction*

The next phase is feature extraction, which plays a decisive role in the interpretation and classification of signals. Features represent distinctive compo-

nents extracted from signal segments, differentiating them from other patterns, as well as reflecting the underlying neural activity. Hence, the quality of feature extraction will have a direct effect on the accuracy of classification. However, extracting features from such complex and dynamic EEG data is a challenging task. Feature extraction techniques can be classified into *one-dimensional* approaches, which derive features from the time domain, frequency/spectral domain, or decomposition domain, and *multi-dimensional* approaches, which extract features from the time-frequency domain and spatial domain [106].

Time-domain feature extraction is a primitive technique, focusing on analyzing signals with respect to time and quantifying temporal changes. For frequency-domain analysis, the time-domain signals are transformed into frequency-domain representation using the Fast Fourier Transform (FFT) algorithm. Another key feature, the PSD, is obtained by using Welch's method [127].

The third method involves decomposing EEG signals into simpler components, each capturing specific aspects of the signal, such as frequency bands, spatial patterns, or temporal dynamics. This decomposition-domain approach functions both as a feature extraction and filtering technique by isolating relevant features while simultaneously removing undesirable components, such as noise or artifacts [106]. The Wavelet Transform (WT) is particularly popular due to its effectiveness with non-stationary signals. While Continuous Wavelet Transform (CWT) provides a highly detailed and continuous time-frequency representation of the signal, Discrete Wavelet Transform (DWT) uses discrete scales and positions to efficiently capture key features of the signal.

In joint time-frequency domain feature extraction, both spectral and temporal features are analyzed together. Time-frequency analysis techniques, including WT and Short-Time Fourier Transform (STFT), achieve this by decomposing signals into smaller components within short time windows, allowing for simultaneous examination of time and frequency characteristics.

Lastly, spatial-domain feature extraction method - also known as spatial filtering - processes EEG signals by focusing on the spatial relationships between the signals recorded from multiple electrodes. Common Spatial Pattern (CSP), a supervised spatial filter, transforms EEG signals into a unique spatial representation where the variance of one class (e.g., a specific task or mental state) is maximized while the variance of the other classes is minimized. This transformation creates features that are highly discriminative for classification tasks.

Following feature extraction, optional steps such as feature selection and dimensionality reduction can be applied to simplify the data and enhance model performance while reducing computational overhead [102]. Feature selection involves selecting the most relevant features from the high-dimensional data while eliminating irrelevant and redundant features [100]. Dimensionality reduction, on the other hand, focuses on transforming the original feature set into a smaller, lower-dimensional representation while preserv-

ing critical information. This helps minimize processing time and improve classification accuracy, especially for high-dimensional data like EEG signals. Linear techniques like Principal Component Analysis (PCA) focus on finding components with the greatest variance, while non-linear methods, such as Independent Component Analysis (ICA), identify components with the greatest statistical independence [102].

2.3.3 *Classification*

The final step is signal classification, where selected features are used to identify patterns or predict the class label of new data points in specific tasks, such as distinguishing between normal and abnormal brain activity [91] or detecting cognitive states [10]. ML uses data as a guide to identify patterns and make predictions on unseen and new data, enabling systems to learn and improve performance without being specifically programmed. Classification algorithms can be separated into two categories: conventional ML algorithms and DL algorithms.

2.3.3.1 *Conventional Classification Algorithms*

Traditional ML algorithms are built on statistical concepts and use handcrafted input features extracted from raw data to train the model. The algorithms then apply statistical methods or predefined rules to classify data into the output categories [96].

ML approaches can broadly be categorized into supervised and unsupervised learning. While supervised learning relies on labeled training data to predict outputs with well-defined categories, unsupervised learning works with unlabeled data, aiming to discover hidden patterns or clusters without predefined labels. Among these, supervised learning methods are more widely used for EEG data because they allow for precise mapping between the input signals and their corresponding labels. This is particularly important because EEG signals are highly complex and non-stationary, often requiring analysis in short-duration segments tied to localized events [106].

The most prevalent classification algorithms encompass Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Random Forest (RF) [96]. These methods are capable of building precise classification models based on input data. However, they also come with some limitations. Traditional machine learning approaches heavily depend on human expertise, prior domain knowledge of EEG data, and a clear understanding of the specific task to identify features deemed relevant for the classification.

Such reliance becomes problematic when dealing with real-world problems, as EEG data are highly complex and dynamic. Designing systems to explicitly capture all the nuances and variability in EEG signals is highly challenging, especially since handcrafted programming is often tailored to specific datasets or scenarios, limiting the ability to uncover hidden patterns or intricate interactions within the data [69]. Moreover, the need for

domain-specific processing pipelines reduces the flexibility and generalization capability of traditional algorithms, often leading to low classification performance when handling highly dynamic features [93].

These challenges have driven progress toward more flexible and automated approaches, such as DL. Unlike conventional methods, DL can automatically extract hierarchical features directly from raw data without relying on separate feature extraction steps. This capability not only simplifies the EEG processing pipeline by offering end-to-end learning and achieving competitive performance on the target task [96].

2.4 DEEP LEARNING IN EEG CLASSIFICATION

2.4.1 *Overview of EEG Classification with Deep Learning*

In the past decade, DL methods have made significant advancements in EEG analysis and classification, driven by their powerful ability to handle large datasets and decode complex patterns [39]. Numerous studies have shown high performance of Deep Neural Networks (DNNs) in classifying EEG signals rapidly and accurately. To fully leverage the advantages of DL, selecting an appropriate neural network architecture is crucial. A systematic literature review by [91] analyzed 154 papers on EEG classification using DL published between 2010 and 2018, revealing the prevalence of neural network architecture choices. The most frequently used framework was CNN, accounting for 40% of the papers. RNNs and AEs followed, each being chosen for about 13% of the works. Combinations of CNNs and RNNs, as well as DBNs, appeared in 7% of studies, while other architectures, including Boltzmann machines (RBMs), fully connected (FC) neural networks, and generative adversarial networks (GANs), were used in the remaining studies.

The widespread adoption of CNNs can be attributed to their ability to extract spatial features effectively, manage 2D and 3D representations, and exploit hierarchical structures in EEG signals. Similarly, RNNs are favored for their excellence at capturing temporal dependencies and sequential patterns in time-series EEG data, while AEs are known for their strength in unsupervised feature extraction. The advantage of a dual stream spatio-temporal neural network, CNN-RNN, over an independent CNN or RNN model lies in its stronger learning and memory capabilities [125]. The combination of CNNs and RNNs offers a powerful approach to handle the spatiotemporal complexity of EEG data, making full use of the strengths of each architecture to uncover both spatial and temporal patterns. In addition to these approaches, DBNs are effective for learning hierarchical features through unsupervised pretraining and handling noisy EEG data. GANs can help generate synthetic EEG data to address class imbalance and model complex distributions to enhance classification performance.

Among the diverse applications of DL in EEG analysis, 86% of state-of-the-art studies have focused on using DL for the classification tasks involving EEG data [91]. The remaining research explores methods to enhance processing

tools, including feature learning, artifact handling, and model visualization, and generating data from EEG signals using DL. Researchers have applied DL to a wide range of EEG classification tasks. According to the review by [26], EEG classification tasks are categorized into six major groups along with their respective application distributions: motor imagery (22%), emotion recognition (16%), mental workload (16%), seizure detection (14%), event-related potential (ERP) detection (10%), and sleep stage scoring (9%). Beyond these primary categories, 13% of studies explore various applications, such as Alzheimer's classification, bullying indices detection, depression diagnosis, gender classification, and the detection of abnormal EEG patterns.

2.4.2 *Deep Learning in various Types of EEG Classification Tasks*

This section presents an overview of the different types of EEG classification tasks as well as the DL methods applied to them. EEG classification tasks cover a broad range of domains, each with unique challenges and requirements. The majority of existing studies have concentrated on developing domain-specific DL models tailored to datasets from a particular domain.

2.4.2.1 *Motor Imagery Tasks*

Motor Imagery (MI) tasks are extensively studied for their ability to activate similar brain pathways as in actual movement execution. MI signals are of spontaneous type because they are generated solely through the user's imagination of performing specific movements (e.g., hand or foot movements) without any physical execution [6].

A wide range of DL architectures has been developed for MI datasets. According to Al-Saegh, Dawwd, and Abdul-Jabbar [6], the most predominant architecture type is CNN (73%), followed by hybrid-CNN (hCNN) and RNN (including LSTM and GRU). hCNNs integrate layers from other architectures, such as autoencoders or RNNs, with standard convolutional layers

Several well-known CNN models have been widely applied to MI tasks and even across other domains, including EEGNet [67], DeepConvNet, and ShallowConvNet [99]. Beyond standard CNNs, other advanced variants such as attention-based CNNs, residual-based CNNs, inception-based CNNs, DenseNets, and 3D-CNNs, have further enhanced MI classification, achieving accuracies of up to 90% [10]. For instance, a CNN using wavelet transform temporal-frequency image representations achieved 85.59% accuracy on BCIC-IV-2a dataset and 90% on BCIC-II-3 [132], while an end-to-end multi-branch multi-scale CNN achieved approximately 80% accuracy on BCIC-IV-2a [57].

Hybrid CNN architectures have also been effective. A hybrid CNN-Stacked Autoencoder (SAE) model, which processes 2D spectral images using a 1D convolutional layer followed by a six-layer SAE, attained accuracies of 90.0% and 77.6% on the BCIC II-3 and BCI-C IV-2b datasets, respectively [115]. Zhang et al. [135] introduced a deep convolutional generative adversarial network (DCGAN) consisting of four-layer GAN model for MI data augmen-

tation. The DCGAN outperformed traditional data augmentation techniques such as geometric transformations and autoencoders.

RNN/LSTM architectures have excelled at capturing temporal features in MI EEG data. A model combining LSTM with common spatial pattern (CSP) and support vector machine (SVM) achieved 82.52% accuracy on the BCIC-IV-1 dataset citekumar2019brain.

2.4.2.2 *Emotion Recognition Tasks*

Emotions are defined as complex mental states that influence physical behaviors and physiological activities [55]. They are vast, diverse and difficult to categorize due to their non-mutually exclusive nature, causing overlaps and ambiguities across categories. According to Russell [94], emotion can be classified using dimensional models into two dimensions: arousal and valence. The term 'valence' refers to the level of pleasure, ranging from unpleasant (e.g., sad, stressed) to pleasant (e.g., happy, elated); meanwhile, 'arousal' indicates the level of excitation, from inactive (e.g., uninterested, bored) to active (e.g., alert, excited).

EEG is considered a reliable and objective source for detecting true emotions [98]. It is found that emotion recognition requires capturing expression over long durations with contextual temporal dependencies. Therefore, emotion recognition tasks typically involve having subjects watch video clips that have been pre-labeled with specific emotions by experts. EEG was measured during these viewings and an emotion self-assessment typically followed [26].

Li et al. [70] introduced a CNN-RNN framework (C-RNN) using Continuous Wavelet Transform (CWT) to preprocess EEG signals into 2D spectral energy representations. The CNN layers extracted cross-channel correlations, while the LSTM layers modeled temporal dependencies across sequential frames. Evaluated on the DEAP dataset, the C-RNN achieved accuracies of 74.12% for valence and 72.06% for arousal.

Alhagry, Fahmy, and El-Khoribi [7] presented an end-to-end LSTM-RNN as a lightweight approach that outperformed feature-based methods with accuracies of 85.65% for arousal and 85.45% for valence. Similarly, Salama et al. [98] proposed a 3D-CNN to extract the spatiotemporal features from EEG signals. A 3D input representation of EEG data was generated by combining 2D spatial matrices of multiple EEG channels and temporal frames. The model achieved recognition accuracies of 87.44% for valence and 88.49% for arousal.

Song et al. [108] introduced a state-of-the-art method, Dynamical Graph Convolutional Neural Network (DGCNN). Unlike traditional GCNNs, DGCNN dynamically learned the adjacency matrix representing the intrinsic relationships among EEG channels during training. EEG signals were modeled as a graph, where each channel was a node, and graph convolution was applied to extract discriminative features across frequency bands. The model achieved accuracies of 86.23% for valence, 84.54% for arousal on the DREAMER.

2.4.2.3 *Mental Workload Tasks*

Mental workload (MWL) refers to the cognitive resources needed to complete a given set of tasks. It is affected by factors such as task complexity, the amount of information to be processed, and individual's cognitive and perceptual abilities [86]. To evaluate Mental Workload (MWL), subjects are typically asked to perform a series of cognitive tasks with different levels of task complexity. Common tasks include the *n-back* task, where the subject needs to recall stimuli from n steps back, with the difficulty increasing as n grows; the *visual search* task, which requires locating a target within a visual field; the *simultaneous capacity (SIMKAP)* task, which assessed multitasking and attentional capacity by measuring a person's ability to process multiple streams of information at the same time [86].

Parveen and Bhavsar [86] introduced an attention-based 1D-CNN model for classifying EEG-based MWL on the STEW dataset, achieving accuracies of 98% for binary classification (rest vs. task) and 79.98% for ternary classification (low, moderate, high workload). Similarly, Chakladar et al. [21] proposed a Bidirectional Long Short-Term Memory (BLSTM)-LSTM hybrid model optimized with Grey Wolf Optimization (GWO) for the same dataset, achieving accuracies of 86.33% for "No task" and 82.57% for "SIMKAP-based multitasking" tasks.

In another study, Sharma and Ahirwal [103] developed a cascaded end-to-end 1D-CNN-BLSTM model tested on the SIMKAP task dataset, achieving binary and ternary classification accuracies of 96.77% and 95.36%, respectively. Furthermore, the study of Ganguly et al. [35] applied a stacked LSTM on the EEGMAT database for mental arithmetic tasks, achieving 91.67% mean accuracy with spectral feature extraction.

2.4.2.4 *Sleep Staging Tasks*

Polysomnography (PSG) is a sleep study to objectively assess the quality of sleep, comprising EEG (brain activity), EOG (eye movement), EMG (muscle activity), and ECG (heart activity) [113]. Sleep is categorized into two primary types: non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. NREM sleep consists of four stages, which transition into the REM sleep stage [75]. Sleep stage classification has been among the least studied task in EEG analysis due to the large volume of overnight EEG recordings required [26]. These signals are typically scored by experts and classified into sleep stages 1, 2, 3, 4, and REM stage.

Supratak et al. [113] presented DeepSleepNet model, which combined CNNs and BLSTM for automatic sleep stage scoring using raw single-channel EEG. Eldele et al. [34] introduced AttnSleep, an attention-based deep learning architecture that used single-channel EEG signals. The model integrated a multi-resolution CNN to extract frequency-specific features, adaptive feature recalibration to enhance feature quality, and a temporal context encoder leveraging multi-head attention combined with causal convolutions to capture temporal dependencies. Similarly, Phan et al. [87] developed an

attention-based bidirectional RNN RNN with GRUs. EEG data were transformed into feature vectors, and smoothed with triangular or learned filter banks. Attention weights identified discriminative temporal features, and a linear SVM performed the final classification. All three models demonstrated strong performance across benchmark Sleep-EDF datasets.

2.4.2.5 *Seizure Detection Tasks*

Epilepsy is among the most prevalent neurological diseases worldwide. Seizures are defined as sudden changes in the brain's electrical activity, leading to altered behaviors including loss of consciousness, involuntary movements, temporary breathing difficulties, and memory loss [107]. These episodes primarily occur in the cortex, the brain's outermost layer, making EEG the most widely used signal for epileptic seizure detection [51]. When epilepsy is present, seizure activity will appear as rapid spiking waves on the EEG recording, often creating noticeable abnormalities that serves as a distinct signature of epileptic activity in EEG data. EEG data are typically recorded from epileptic patients during seizure and seizure-free periods, with some datasets including non-epileptic patients as a control group [26].

There are several approaches to epileptic seizure-related classification: epileptic seizure prediction to recognize the brain state before seizure event (preictal), seizure detection to distinguish between seizure (ictal) and non-seizure (interictal) events, and specific seizure type classification, such as identifying focal or non-focal seizures [121].

Truong et al. [120] utilized a CNN-based architecture for seizure prediction by transforming EEG data into Short-Time Fourier Transform (STFT) spectral images. This approach achieved high sensitivity and low false prediction rate (FPR) on both intracranial and scalp EEG datasets. Tsiouris et al. [121] developed a two-layer LSTM network and exploited a broad variety of features including time and frequency domains, cross-correlation between EEG channels, and graph-theoretic measures. Evaluated on the CHB-MIT Scalp EEG database, the method achieved 100% sensitivity and specificity with a very low FPR.

Hussein et al. [51] proposed an end-to-end LSTM network for seizure detection on the Bonn EEG dataset, achieving 100% accuracy under optimal conditions and remains robust against noise and artifacts such as muscle activity and eye-blinking. Additionally, Xu et al. [133] implemented a 1D-CNN-LSTM model on the UCI Epileptic Seizure Recognition Dataset. This method integrated a 1D-CNN for feature extraction with LSTM layers for capturing temporal patterns in EEG time-series data, achieving high recognition accuracies of 99.39% and 82.00% on the binary and five-class classification, respectively.

2.4.2.6 *Event-Related Potential (ERP) Tasks*

Event-Related Potentials (ERPs) are small brain voltages generated in response to specific sensory, motor or cognitive events or stimuli [114]. De-

tection of Event-related Potentials (ERPs) often involves recording EEG data while subjects perform visual presentation tasks, where they observe a rapid sequence of images or letters and focus their attention on specific targets [26].

The P300 wave, a type of ERPs detectable via EEG, is characterized by a positive voltage deflection occurring roughly 300 ms after the stimulus [19]. Its presence, amplitude, timing, and spatial distribution serve as metrics of cognitive function, particularly in decision-making processes. The P300 speller is based on the oddball paradigm, where infrequent and expected stimuli evoke a P300 response [89]. The main goal is to accurately and rapidly detect the P300 peaks in the EEG. In this paradigm, a 6×6 grid of alphanumeric characters is presented, with subjects focusing on specific target characters. Rows and columns were randomly intensified at 5.7 Hz, with 2 out of 12 intensifications containing the target character, eliciting P300 responses.

Deep learning models for accurate decoding of ERPs remain relatively scarce in the literature [2]. Cecotti and Graser [19] explored various CNN models, ranging from single classifiers (e.g., CNN-1, CNN-2a) to multi-classifier systems (e.g., MCNN-1, MCNN-3) and compared their performance. The best-performing method, MCNN-1, achieved a recognition rate of 95.5% for character detection without requiring channel selection. In another study, Maddula et al. [78] proposed a hybrid approach combining 3D-CNN, 2D-CNN, and LSTM to capture the spatiotemporal structure of ERPs. However, despite the efforts, the search for an optimal deep learning model for ERPs decoding is still ongoing.

Apart from the aforementioned tasks, EEG classification using DL approaches has also been explored in areas such as abnormalities detection, where EEG signals are analyzed to identify irregular brain activities. CNNs have also been applied to steady-state visually evoked potential (SSVEP) tasks, which detect brain responses to flickering visual stimuli at distinct frequencies. These responses help determine the user's focus and identify their intended selection [84]. Additionally, error-related potentials (ErrPs), which arise in response to errors during task execution, have gained attention in BCIs and cognitive monitoring for error correction and adaptive responses [111]. Recent studies have employed DL techniques, such as GAN-CNNs, to classify Error-related Potentials (ErrPs) with improved performance [36]. Furthermore, specialized tasks, such as fatigue detection, diagnosis of neurological and neurodegenerative diseases (e.g., dementia, Alzheimer's disease, brain tumors, strokes, Parkinson's disease) [107], have increasingly adopted DL methods to develop automated detection systems and enhance EEG interpretation, leading to improvements in accuracy, robustness, and automation in real-world applications.

2.4.3 *Generalizability in Deep Learning*

The concept of **generalizability** in deep learning has been interpreted in various ways across the literature, depending on the context in which it is

applied. This section provides a precise definition of generalizability within the scope of this thesis.

Conventionally, **generalizability** refers to the ability of a model to provide accurate predictions on new and unseen data [12]. This includes performing well not only on unseen samples from the same dataset it was trained on but also on new datasets that may represent different conditions or distributions.

In the study of Shahbazi and Aghajan [101], **generalizability** is characterized as the ability of a seizure prediction method to perform reliably and consistently across different patients and conditions. This challenge arises from the complexity and variability of EEG data, both among subjects and between seizures from the same patient. In this study, **generalizability** specifically refers to the transferability of a model. That is, the model should not be specifically tailored to each seizure or individual patient. Transferability is a significant challenge in EEG-based DL due to the high intra- and inter-subject variability in EEG data, together with the scarcity of labeled data. Typically, data from a single patient is insufficient to train and test robust models. Therefore, a model trained on one group of patients or seizure episodes should generalize effectively to unseen patients or new seizure episodes.

Transfer learning has recently been widely explored as a means to improve generalizability in EEG classification. In the review of transfer learning in EEG, Wan et al. [124] defines generalizability as the transferability across different datasets. This describes a model's ability to achieve strong performance on a different dataset after being trained on one, effectively transferring knowledge learned in one domain to another related domain with minimal retraining. A generalizable model should be able to extract transferable and domain-independent features.

Within the scope of this thesis, generalizability is defined as the ability of a model to perform effectively across multiple datasets and domains without requiring architectural adjustments. The model is trained and tested separately on each dataset, ensuring it learns dataset-specific patterns while avoiding overfitting. A model's generalizability reflects its capacity to learn relevant features from new data and accurately classify them, assuming sufficient data is available for training.

2.4.4 *State of the Art*

This section provides an overview of the current state of EEG-based deep learning, focusing on recent advancements towards achieving generalization. There is a limited number of articles directly addressing the generalizability of EEG-based deep learning methodologies, as well as a representative collection of EEG datasets.

A review conducted by Craik, He, and Contreras-Vidal [26] reveals the prevalence of task-specific deep learning strategies in EEG analysis. While tasks such as emotion recognition, motor imagery, and sleep stage scoring showed no preference for specific DL algorithms, seizure detection studies mostly leverage CNN's or RNN's, with RNN's being slightly more in use,

whereas studies on ERPs had a clear preference for CNNs. Additionally, the review underscores the challenge in comparing classification accuracy achieved by different architecture design choices across tasks and EEG datasets due to the variety in algorithm design or input processing methods. This lack of standardization makes evaluating the overall effectiveness of different DL approaches for EEG analysis difficult.

Heilmeyer et al. [46] proposed a novel framework for the large-scale evaluation of DL architectures on EEG datasets across various decoding problems of different difficulty, promoting generalizability beyond specific domains. The evaluation leveraged four well-established CNN architectures: Braincode Deep4 ConvNet, Braincode Shallow ConvNet and two versions of EEGNet. To ensure that the success or failure of the compared methods is not limited to a specific decoding domain, they selected a range of datasets representing various common BCI tasks including motor tasks, speech imagery, and error processing. This approach provides valuable insights into the performance and applicability of DL architectures in diverse EEG decoding scenarios.

EEGNet, an EEG-specific convolutional neural network introduced by Lawhern et al. [67], has proven its ability to generalize across different BCI paradigms: P300 visual-evoked potentials, Error-Related Negativity (ERN), Movement-Related Cortical Potentials (MRCP) and Sensory Motor Rhythms (SMR). It is probably the first work that has validated the efficacy of a single network architecture across multiple BCI datasets. The results showed that deep CNN (i.e., five convolutional layers) tended to perform better on the oscillatory BCI data set, while shallow CNN (i.e., two convolutional layers) achieved better performance on the event-related potential BCI data set [39]. EEGNet was further evaluated for emotion recognition in the study conducted by Wang et al. [126], albeit achieving lower accuracy compared to other tasks.

Although still in its early stages, the development of generalizable DL frameworks for EEG classification is gaining increasing attention. The end-to-end multi-branch multi-scale CNN of Jia et al. [57], initially developed on MI tasks, reached an outstanding accuracy compared to state-of-the-art models and delivered consistent performance across subjects and time. The authors now attempt to generalize this framework further by extending it to other EEG classification domains, such as the EEG-based emotion recognition.

3

METHODOLOGY

This chapter explains the methodology used to address the research focuses outlined in Chapter 1. It begins with an overview of the study design and research framework, followed by a detailed explanation of its implementation.

3.1 CONCEPTUAL FRAMEWORK

The framework consists of three key components: (i) a collection of EEG data, (ii) DL methods for classification, and (iii) an approach for subset selection. This chapter provides a detailed description of each component. The primary objective of this study is to identify a small yet representative subset of EEG datasets that capture the diversity of a broader dataset collection. The ultimate goal is to ensure that future researchers can evaluate algorithms and confidently infer their performance on the entire range of EEG datasets from various domains based on the subset. Additionally, this study explores which datasets can consistently achieve high classification accuracy across deep learning models, while also identifying datasets that remain particularly challenging to classify with current techniques.

3.1.1 *Dataset Selection*

We selected commonly-used and publicly available EEG-based experiment datasets as our primary data sources based on prominent literature review articles in the field. Datasets should represent a broad spectrum of common EEG-based classification paradigms, including emotion recognition, motor imagery [6], sleep stage scoring, seizure detection, mental workload, and ERPs.

3.1.1.1 *Data Preprocessing*

In conventional EEG classification pipelines, the initial step involves data preprocessing and feature extraction from raw EEG data. In this study, multiple end-to-end deep learning models are applied directly to raw EEG signals, automatically learning features from the data. Therefore, no additional feature extraction step is necessary. However, EEG signals still need preprocessing to eliminate or reduce the effects of artifacts caused by their low SNR [39]. To eliminate any discrepancies and ensure fair comparisons across datasets, all preprocessing steps are standardized.

As mentioned in Chapter 2, EEG data requires multiple preprocessing steps to achieve high-quality data for reliable analysis. However, the wide variety of available preprocessing techniques makes it impractical to exhaus-

tively apply all of them. In order to select appropriate preprocessing steps without negatively influencing the end results, we conducted a small experiment. We searched for a study on EEG classification that demonstrated high-performance results and provided publicly available code for both dataset preprocessing and model implementation, along with an open-access dataset. This paper served as a reference point for evaluating and refining preprocessing pipelines. The study of Ingolfsson et al. [53] was selected, with the corresponding code obtained from the associated GitHub repository. The model proposed in this study, EEG-TCNet, was evaluated on the 4-class motor-imagery BCI Competition IV-2a dataset, achieving classification accuracy of 77.35%. Various combinations of preprocessing steps (e.g., bandpass filtering, downsampling, artifacts removal, sliding window, standardization) were tested on the dataset and model to identify configurations that yield results comparable to the reference study. Based on these tests, a simple yet effective preprocessing pipeline was adopted, which includes downsampling, normalization, and generating sliding window samples.

For our experiments, EEG signals are downsampled to 128 Hz, then normalized by subtracting the mean and scaling each channel to unit variance. This downsampling procedure enables the model to coherently analyze EEG time-frequency patterns using a consistent encoder architecture [16]. Next, the EEG recordings are divided into overlapping short time frames using a sliding time window approach. Figure 3.1 illustrates a sliding window example with a 2-second duration and a 1-second overlap. This step not only facilitates signal processing by focusing on smaller and more manageable segments of data but also serves as a form of data augmentation, effectively increasing the size of the training dataset. Overlapping windows provide the model with more diverse input samples while preserving important temporal patterns. Finally, the resulting data shape is standardized across all datasets to facilitate seamless input into deep learning models, formatted as (number of segments, number of channels, segment length).

3.1.2 *Deep Learning Model Selection*

Widely-known DL models will be selected based on review papers in the field, citation frequency, performance on tasks, and reproducibility. In this study, we focused only on end-to-end DL models capable of decoding raw EEG signals without any preprocessing and exploiting hierarchical structure on the data.

As end-to-end DL has the capabilities to directly self-learn the hierarchical structure of datasets, it raises the question of whether these models can effectively learn data from diverse categories. These categories often have distinct structures and typically require feature extraction to be customized for each category. The DL models in this study were specifically designed to fit EEG data within certain categories. To evaluate their generalizability, we tested the models across data from various categories, assessing their performance beyond their intended domains. To ensure accurate results and

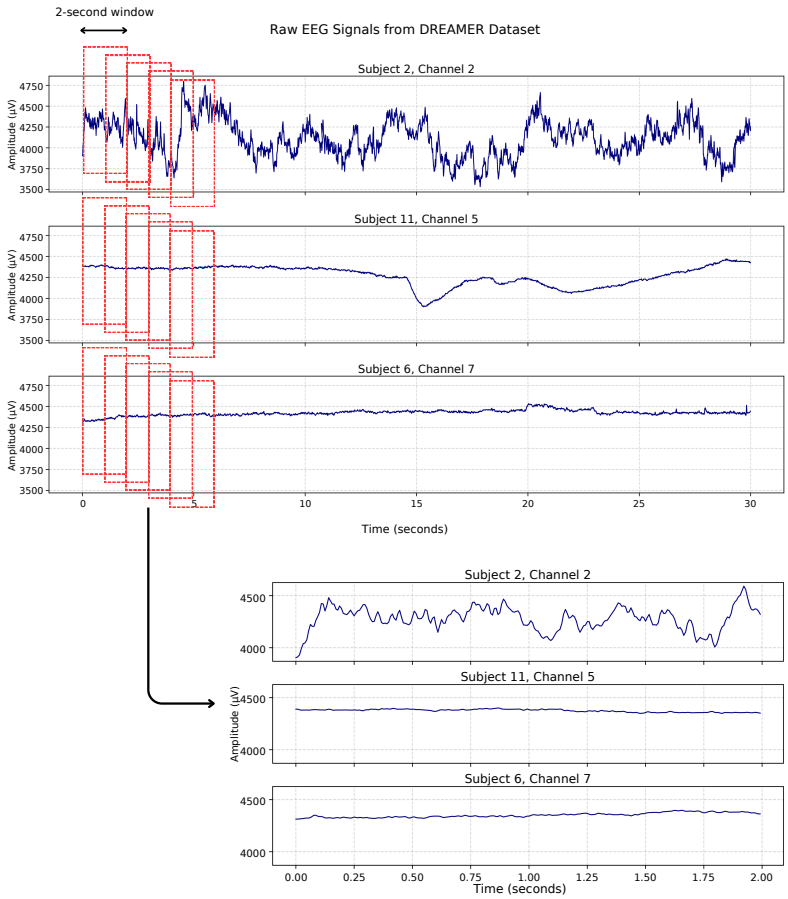

Figure 3.1: An Example of EEG Window Sampling

minimize errors, we replicated the original implementations of the authors for all decoding methods. For implementation details, refer to the respective publications.

Performance Evaluation

To ensure a fair comparison, we aimed to use consistent train-test splits and validation techniques across all models and datasets. However, due to considerable differences in dataset structures, achieving the ideal scenario was not feasible. Instead, we made an effort to apply methods that are as standardized as possible for all datasets, employing an appropriate train-test split and cross-validation approach for each dataset.

3.1.2.1 *Train-Test Splits*

The main goal is to evaluate the performance of existing, readily available models across various EEG datasets rather than to develop or fine-tune new models. Therefore, a validation dataset, typically used for hyperparameter tuning, is not required. Instead, each dataset is split into training and test sets to directly assess the models' performance. This approach ensures that the

study remains aligned with its main goal without introducing unnecessary complexity.

Because of the high inter-subject variability commonly seen in EEG data [97], most of the EEG classification methods are developed in a subject-specific scenario [52]. Thus, in most cases, the datasets are loaded so that models are trained and evaluated on each subject's data separately. Some datasets are already split into predefined train and test sets; for the remaining datasets, the data will be split trial-wise or segment-wise. However, for datasets with a large number of subjects (more than 20) or limited data per subject, a subject-independent classification approach with 3-fold cross-validation and three repetitions is applied. The decision to use 3 folds was made after careful consideration of the limited computational resources available.

Rather than combining data across all subjects and splitting it into k folds, trials from a single subject are kept together within the same fold to prevent data leakage. Splitting trials from the same subject across different folds increases the risk of data leakage, as patterns unique to a specific subject may appear in both training and testing sets, potentially inflating model performance [58]. To address this, we apply subject-level splits. A subset of randomly selected subjects is designated as the test set, while the remaining subjects form the training set. This process is repeated three times, with a different subset serving as the test set each time.

3.1.2.2 *Metrics*

Each of the DL models will be trained and tested on all EEG datasets to evaluate their ability to generalize across different types of data. The performance of the models will be assessed using a variety of metrics, including accuracy, $F1$-score, precision, and recall. These metrics are selected to address potential class imbalances and provide a comprehensive view of model performance. This multi-metric approach also ensures flexibility in identifying the most suitable metrics for the subsequent subset selection steps. Since each dataset includes multiple subjects or involves k-fold cross-validation, the final performance metrics will be calculated as the average across all subjects or repetitions. Additionally, confusion matrices and training history will be generated to verify the functionality of the model and dataset implementations by identifying potential issues in classification and training behavior.

Representative Subset Selection Procedure

Once all models have been successfully trained across all datasets, the expected result is a matrix where each cell contains the performance metric for a specific model-dataset pair. Building on this, we now shift focus to the main objective of this study: finding a small yet representative subset of EEG datasets.

This approach was inspired from the methodology outlined in the paper *Atari-5: Distilling the Arcade Learning Environment Down to Five Games*, which

proposed a novel strategy for selecting small but representative subsets of reinforcement learning (RL) environments. The study addressed the computational inefficiency of evaluating algorithms on the full 57-game dataset of the Arcade Learning Environment (ALE). By applying a systematic subset selection method, the authors introduced *Atari-5*, a five-game subset capable of approximating the median score estimates on the full dataset, but at less than one-tenth the cost. These results can accelerate the development of novel algorithms through faster iteration and also improve the reproducibility of results in Reinforcement Learning (RL) [5]. In this study, the prediction of an established summary score (target score) served as the guideline for subset selection. Specifically, the performance scores from various RL algorithms were used to evaluate all possible subsets of games of a given size. The subsets were ranked based on their ability to predict the overall median score, evaluated by R^2 values. As R^2 measures the percentage of variance in the target score explained, a high R^2 value indicates the subset's ability to capture the variability in performances of the full dataset.

Adapting this framework, we developed a customized algorithm specifically tailored for this study. We implemented a modified subset selection procedure using Ridge regression, combined with correlation analysis to account for multicollinearity between datasets. The goal is to identify a subset of datasets that minimizes the prediction error of a regression model trained to predict the summary metric while ensuring diversity without redundancy.

The computation of the target metric plays a crucial role in determining the optimal subset. Two approaches are considered for computing $t - k$:

1. Included Approach: The target metric t_k is computed as the median F_1-score across all EEG datasets, *including* the candidate subset. This approach ensures that the target metric reflects the performance diversity across the entire dataset but risks introducing bias through data leakage when information from the target value leaks into the predictors.

2. Excluded Approach: The target metric t_k is computed as the median F_1-score across all EEG datasets, *excluding* the candidate subset. This approach avoids potential data leakage when the candidate subset remains unseen during the computation of the target metric.

Additionally, to address multicollinearity, clustering and pruning steps were introduced, which removed highly correlated EEG datasets before the subset selection procedure.

Among the selected datasets, some originate from the same source but are categorized differently (e.g., datasets used for emotion recognition may analyze emotions based on distinct dimensions named valence and arousal [54]), while others represent extended versions of existing datasets. In such cases, these datasets may show similar performance tendency across all models, resulting in high correlations and potential multicollinearity issues. In regression analysis, multicollinearity occurs when two or more predictors are highly correlated. This is undesirable because it inflates the standard errors of coefficients and causes unreliable coefficient estimates [28].

It is crucial that the selected subset captures the diversity of the entire dataset collection rather than duplicating information and becoming redundant. To ensure this, a *correlation analysis* is conducted prior to regression modeling to reduce redundancy and handle multicollinearity. A pairwise correlation matrix of metrics across datasets is generated. Using hierarchical clustering, datasets with high correlations (e.g., correlation > 0.9) are grouped into clusters based on Ward's linkage method, which minimizes within-cluster variance [81]. Only one dataset, the most similar to the others in its cluster, is selected as the representative for each cluster. This is determined using a similarity-based criterion: the average similarity (or inverse distance) of each dataset to all other datasets in the cluster is calculated. The dataset with the highest average similarity is chosen as the representative for its cluster. The end result is a collection of datasets with low correlations and diverse characteristics.

Following the preprocessing, a brute force search, also known as exhaustive search, is conducted to thoroughly evaluate all possible combinations of datasets based on Mean Squared Error (MSE) through regression modeling. This approach guarantees that the selected subset is the optimal choice according to the evaluation criterion. For n datasets, there are $2^n - 1$ possible non-empty subsets to evaluate.

Regression with regularization is employed to address any remaining correlations among the datasets in the subset. Specifically, *Ridge regression* models are used to predict the target metric for models based on the subset of datasets. Ridge regression applies a penalty to large coefficient estimates, shrinking them toward zero, which mitigates correlations while maintaining all features—in this case, datasets [56]. It is important to note that the purpose of the regressions in this context is as a method of feature selection rather than predictive power. The primary focus is on identifying which datasets are selected through this process, rather than on their exact performance in predicting the target summary score.

Given m models $M_k : k = 1, \ldots, m$, with their individual evaluations $s_{ki} \in \mathbb{R}$ on datasets $i \in \{1, \ldots, n\}$, and a summary metric $t \in \mathbb{R}^m$, where each entry t_k represents the summary metric (e.g., the median $F1$-score) of model M_k across all datasets. The dataset for regression modeling is defined as:

$$D = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1n} & t_1 \\ s_{21} & s_{22} & \ldots & s_{2n} & t_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ s_{m1} & s_{m2} & \ldots & s_{mn} & t_m \end{bmatrix},$$

where s_{ki} is the performance of model M_k on dataset i.

Let $I \subseteq \{1, \ldots, n\}$ be a subset of columns (datasets), and $s_{k,I}$ represent the corresponding sub-vector of scores for model k. We aim to find a map-

ping $f_I^* : \mathbb{R}^{|I|} \to \mathbb{R}$ that best predicts t_k from $s_{k,I}$, minimizing the following objective:

$$f_I^* = \arg\min_f \sum_{k=1}^{m} (t_k - f(s_{k,I}))^2 + \lambda \|\theta\|^2,$$

where λ is the Ridge regression regularization parameter and θ are the regression coefficients.

Given $|I| = C$, we seek to find a subset I of fixed size C that minimizes the prediction error. Thus, we solve:

$$I^* = \arg\min_{|I|=C} \sum_{k=1}^{m} \left(t_k - f_I^*(s_{k,I})\right)^2,$$

where f_I^* is the optimal Ridge regression model trained on the subset I.

This process involves iterating over all possible subsets $I \subseteq \{1, \ldots, n\}$ of size C, training a Ridge regression model for each subset, and selecting the subset I^* that yields the lowest Mean Squared Error (MSE) via 5-fold cross-validation. The final selected subset I^* and its corresponding regression model f^* provide an efficient and interpretable solution to the representative subset selection task.

3.2 IMPLEMENTATION

3.2.1 *Dataset Selection*

3.1 provides a list of datasets along with their respective categories. The selected datasets are publicly available and have been widely used as benchmarks for state-of-the-art classification algorithms. A detailed description of each dataset is presented below, including preprocessing steps, data splitting strategies, cross-validation methods, and other relevant processing details.

Dataset	**Task Type**
STEW, EEGMAT	Mental Workload
SEED, SEED IV, DEAP, DREAMER	Emotion Recognition
BCIC-IV-2a, BCIC-IV-2a, High-Gamma, PhysioNetMI	
BCIC-III-2 Dataset	Event-Related Potential
CHB-MIT, Siena Scalp Dataset	Epilepsy Detection
The TUH-Abnormal EEG Corpus	Neurological Abnormality Detection
Sleep-EDF	Sleep stage scoring

Table 3.1: Selected EEG Datasets and Their Corresponding Task Types

3.2.1.1 *STEW Dataset*

The Simultaneous Task EEG Workload (STEW) dataset contains raw EEG data from 48 subjects who performed a multitasking mental workload activity using the Simultaneous Capacity (SIMKAP) test [71]. EEG signals were recorded from 14 channels at a sampling rate of 128 Hz. The experiment consisted of two tasks: a 'No task' condition and a 'SIMKAP task' condition. First, participants were asked to sit comfortably without performing any task for three minutes, which served as the resting condition. They then completed the SIMKAP test for three minutes to assess their ability in multitasking heavy occupations. To minimize transition effects, the first and last 15 seconds of each recording were excluded, resulting in 2.5 minutes of data for each condition. After each segment of the experiment, participants rated their perceived mental workload (MWL) on a 1-to-9 scale. These ratings were divided into three levels: low, medium, and high MWL. For this study, we focus on classifying EEG data to distinguish between states with and without mental workload. The signals were segmented into overlapping windows with a window size of 2 seconds and a 1-second overlap. Given the large number of subjects, a three-fold subject-level cross-validation scheme was applied for robust model evaluation.

3.2.1.2 *EEGMAT Dataset*

The EEGMAT Dataset contains 23-channel EEG recordings sampled at 500 Hz from 36 participants, collected before and during the performance of mental arithmetic tasks [38, 138]. The signals were preprocessed with artifact removal, a high-pass filter at 30 Hz, a 50 Hz power line notch filter, and segmented with a 60-second window. Each participant contributed two recording files: a baseline recording, during which they were not performing any task and were instructed to sit comfortably, and a recording captured while they performed the mental arithmetic task. For the arithmetic task, participants were asked to serially subtract two numbers — a 4-digit minuend and a 2-digit subtrahend — and communicate their results orally. In this study, we focused on classifying EEG recordings as either with or without mental arithmetic tasks. The EEG data were further epoched into one-second intervals with a 1-second overlap. A subject-level 3-fold cross-validation technique was applied for classification.

3.2.1.3 *DEAP Dataset*

The DEAP dataset [64] was collected from 32 healthy participants aged between 19 and 37. Their EEG and peripheral physiological signals were recorded as they watched a set of 40 one-minute long music videos. Afterward, participants were asked to complete a self-assessment to rate provide subjective ratings for each video based on affective characteristics, including arousal, valence, liking, dominance, and familiarity, using a discrete 9-point scale. These ratings provided subjective insights into emotional impact of the stimuli. For this study, only arousal and valence were selected for the

classification task, with the data categorized into low and high arousal or valence classes. The EEG signals were originally recorded at 512 Hz using a 48-channel setup. In addition to standard preprocessing steps, the data was segmented into 60-second trials, with a 3-second pre-trial baseline removed. Only 32 EEG channels were selected, and eye-movement artifacts were removed. Based on the result of Liu et al. [72], a sliding window size of 8 seconds was used for evaluating the networks. This means that for each 60-second trial, 14 segments were obtained using an 8-second window with a 4-second overlap. Each segment retained the label of the original sample. A three-fold cross-validation was used for this dataset.

3.2.1.4 *SEED Dataset*

The SJTU Emotion EEG Dataset (SEED), contributed by Duan, Zhu, and Lu [33] and Zheng and Lu [137], includes EEG recordings from 15 healthy participants watching Chinese film clips designed to elicit neutral, negative, and positive emotions. Each participant completed three experimental sessions on different days, with 15 trials for each session. n a single trial, participants were presented with a five-second movie hint, followed by a four-minute film clip, a 45-second self-assessment and 15-second resting period. The EEG signals were collected using a 62-channel EEG system at a sampling rate of 1000 Hz, which was further downsampled to 200 Hz with a bandpass frequency filter of 0-75 Hz. Class labels -1, 0, and 1 represent the negative, neutral, and positive emotional states, respectively. A subject-dependent approach was applied, where each participant's data was used individually to build and evaluate models for that specific subject. The 45 trials for each participant were divided into training and testing sets with an 8:2 ratio.

3.2.1.5 *SEED-IV Dataset*

The SEED-IV dataset is an expanded version of the original SEED dataset. The experimental setup is similar to that of SEED, but emotions are categorized into four distinct states: happiness, sorrow, neutrality, and fear. EEG data was collected from 15 participants, each completing three sessions with 24 trials per session. Each film clip lasted approximately two minutes, preceded by a 5-second baseline period and followed by a 45-second self-assessment time. For each participant, the 72 total trials were divided into training and testing sets in an 8:2 ratio.

3.2.1.6 *DREAMER Dataset*

In the DREAMER dataset, 23 participants watched 18 film clips and provided self-assessments of their affective states after each stimulus. Signals were recorded at a sampling rate of 128 Hz with 14 EEG channels. Each clip varied in length from 64 to 393 seconds and was designed to evoke one of nine emotions: amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, or surprise [60]. Participants rated their levels of arousal, valence, and dominance on a five-point scale. For this study, arousal and

valence were used in the classification task and were categorized into low and high states using a threshold of 3. The data were segmented into one-second sliding windows for training the models. A subject-dependent approach was initially considered for this dataset; however, due to imbalances in class distribution among some subjects, a 3-fold cross-validation approach was ultimately chosen to ensure more balanced and reliable model evaluation. For example, subject 6 had only one low-arousal trial compared to 17 high-arousal trials (Appendix A)

3.2.1.7 *BCI Competition IV-2a Dataset*

The BCI Competition IV-2a (BCIC-IV-2a) dataset [18] consists of EEG recordings from nine subjects performing imagined movements of four different body parts: the left hand, right hand, both feet, and tongue. Data collection was organized into two separate sessions, one for training and the other for testing, each containing a total of 288 trials with balanced class distribution.

EEG signals were recorded using 22 electrodes at a sampling frequency of 250 Hz, with a bandpass filter between 0.5 and 100 Hz, along with an additional 50 Hz notch filter. Each trial lasted 7.5 seconds, beginning with a 2-second preparation period, followed by a 1.25-second cue indicating the motor imagery (MI) task. Participants were then prompted to perform the specified MI task and continue it for 4 seconds, with a brief break following each trial.

Sliding input windows within each trial were generated based on the method in Schirrmeister et al. [99], using a window size of 2 seconds and a 1-second overlap. The first crop began 0.5 seconds before trial onset, and the final crop extended to 4 seconds after trial onset. Models were evaluated using a subject-dependent approach.

3.2.1.8 *BCI Competition IV-2b Dataset*

The BCI Competition IV-2b (BCIC-IV-2a) dataset [68] is designed for binary classification of motor imagery (MI) involving imagined left-hand and right-hand movements. Nine subjects participated in five sessions, each consisting of 120 trials.

The first two sessions provided training data without feedback, while the last three sessions included smiley feedback. In the cue-based screening sessions, each trial presented a visual cue (a left or right arrow), prompting the subject to imagine the indicated hand movement for four seconds. In the feedback sessions, a smiley on the screen provided performance feedback by turning green or red, with participants instructed to maintain motor imagery to keep the smiley on the correct side.

The timing scheme of this paradigm is similar to that of BCIC-IV-2a. Three bipolar recordings (C3, Cz, and C4) were sampled at 250 Hz and band-pass filtered between 0.5 Hz and 100 Hz and notch filtered at 50Hz. Within the trial, 2-second windows with a 1-second overlap were generated. For model evaluation, the screening sessions and the first feedback session were used

as training data, while the last two feedback sessions served as testing data. A subject-dependent approach was applied, where each model was trained and validated on individual participants' data.

3.2.1.9 *High-Gamma Dataset*

High-Gamma Dataset was acquired under controlled recording conditions to capture movement-related frequencies in the high-gamma range for motor imagery (MI) tasks with minimal noise [99]. The dataset includes recordings from 14 healthy subjects using 128 electrodes sampled at 500 Hz and consists of 880 trials in training set and 160 trials in the test set. Subjects were instructed via visual stimuli (arrow) to remain still, or repetitively tap the fingers of either left hand, right hand, or flex the toes of both feet for a duration of four seconds while the arrow was displayed. The task involved classifying movements in each trial as left hand, right hand, both feet, or rest. The EEG data was segmented using a 2-second sliding window. Models were trained individually for each subject, and the average performance was calculated to obtain the final results.

3.2.1.10 *PhysioNet MI-EEG Dataset*

The PhysioNet EEG Motor Movement/ Imagery Dataset [38] comprises 64-channel EEG signals from 105 individuals recorded by BCI2000 system with a sampling frequency of 160 Hz. Participants performed or imagined performing motor tasks involving the opening and closing of their hand and feet for a duration of 4 seconds each.

The experiment included three two-minute runs for four distinct motor imagery (MI) tasks: left fist, right fist, both fists, and both feet, with each MI task requiring 21 trials. A typical trial began with a 2-second relaxation period, starting at t = -2 seconds, during which the participant was instructed to relax. At t = 0s, a visual target appeared on the screen, indicating the task to be performed or imagined. The participant then executed the assigned MI task for 4 seconds. At t = 4s, the target disappeared, signaling the end of the trial, followed by a 2-second rest interval before the next trial began.

The dataset in this study was subjected to three-fold cross validation. 2-second windows with a 1-second overlap were generated within the trial.

3.2.1.11 *BCI Competition III-2 Dataset*

BCI Competition III Dataset II (BCIC-III-2) represents a record of P300 evoked potentials generated using the P3 Speller paradigm in BCI2000 [138]. EEG data were collected from two subjects across three sessions, using 64 electrodes with signals bandpass-filtered between 0.1–60 Hz and sampled at 240 Hz. The goal is to classify whether the post-stimulus EEG signal contains a P300 event-related potential (ERP).

In this study, continuous EEG data were segmented into target and non-target trials, each capturing data of 667 ms after stimulus onset, as this time window is sufficient for P300-based character recognition [89]. Each subject

contributed 85 training and 100 testing characters. In each trial, two P300 target responses and ten non-target responses were generated, resulting in an imbalanced dataset. To address this problem, oversampling was applied to equalize the class distribution by replicating minority class instances. Specifically, the P300 training samples were replicated four times to match the sample sizes of P300 and non-P300 data [73].

3.2.1.12 *CHB-MIT Dataset*

The CHB-MIT Scalp EEG dataset was acquired through a collaboration between the Massachusetts Institute of Technology (MIT) and Boston Children's Hospital [38, 104]. It contains EEG recordings from 22 pediatric patients with intractable seizures. Subjects were observed for many days after the withdrawal of antiseizure medication. The recordings can contain none, one or more seizures, with the onset and end of each seizure annotated. EEG signals were sampled at 256 Hz, and the number of channels varied between sessions, though most recordings used 23 channels. Each file contains between one and four hours of EEG data.

In this study, the classification task is seizure detection, distinguishing between seizure (ictal) and non-seizure (interictal) EEG segments. Seizures have shorter duration than non-seizures, resulting in an imbalanced class distribution. To address this, we extracted 5-second seizure segments with a 4-second overlapping, while non-seizure segments were extracted as non-overlapping 5-second windows [8].

A consistent set of 18 EEG channels was used across all 24 cases. Three epochs from subject 12 (files "chb12_27–29.edf") were excluded due to montage inconsistencies with the rest of the recordings [121]. For this data, our evaluation focused on patient-specific modeling. For training and testing, a trial-wise dataset split was applied, ensuring that one entire record (an .edf file) was reserved for testing while the remaining files were used for training.

3.2.1.13 *SIENA Dataset*

The Siena Scalp EEG Database contains EEG recordings from 14 patients collected at the Unit of Neurology and Neurophysiology, University of Siena [29, 38]. The data were recorded using Video EEG Monitoring at a sampling rate of 512 Hz, with most recordings using 29 electrodes. Seizures were classified by expert clinicians using International League Against Epilepsy criteria after reviewing clinical and electrophysiological data. The dataset includes 41 EEG recordings with 47 seizure intervals, and recording durations range from 1 to 13 hours. Start and end time of both recordings and seizures were also annotated.

Data were segmented with a window length of five seconds. As the number of seizure segments was significant lower than the non-seizure ones, a 4-second overlap was applied to seizure windows, while non-seizure windows were non-overlapping. Patient-specific modeling, which involves training a unique model using data from a single patient, was employed in this study.

3.2.1.14 *TUH-Abnormal Corpus*

The TUH EEG Corpus, developed by Temple University Hospital, is the world's largest publicly available collection of clinical EEG data, containing over 25,000 recordings from more than 14,000 patients [43]. For this study, we used a subset, the TUH Abnormal EEG Corpus (v3.0.1), in which EEG signals are labeled by experts as either normal or abnormal [30]. Abnormal EEG signals were recorded from patients diagnosed with various pathologies, such as epilepsy, strokes, depression, and Alzheimer's disease.

This dataset is demographically balanced with respect to patient gender and age and consists of 1,488 abnormal and 1,529 normal EEG sessions. It is further divided into a training set (1,361 abnormal and 1,379 normal samples) and a test set (127 abnormal and 150 normal samples). Each patient appears only once in the evaluation set, labeled as either normal or abnormal, while some patients may appear multiple times in the training set.

The recordings include signals from at least 21 standard electrode positions, sampled predominantly at 250 Hz. Each recording contains approximately 20 minutes of EEG data. For preprocessing, a subset of 21 electrode positions was selected, and the first minute of every recording was discarded to remove potential artifacts [37].

Previous studies [76, 92] suggest that neurologists can accurately classify an EEG session as normal or abnormal by examining just the first few minutes of the signal. Based on this observation, we initially extracted only the first 60 seconds of EEG recordings for both the training and test datasets, hypothesizing that deep neural networks could achieve similar classification performance with this reduced input.

However, this approach significantly reduces the amount of data available for training, potentially impacting model performance. To address this, we extended the training set by extracting the first 4 minutes of each recording and segmenting it into non-overlapping 1-minute epochs, thereby increasing the number of training samples without performance degradation, as demonstrated in the experiment by Roy, Kiral-Kornek, and Harrer [91].

For this dataset, a subject-specific approach is not feasible because patients in the training and evaluation sets are completely distinct, with no overlap between the two groups. Since the training and evaluation sets are already pre-defined, k-fold cross-validation cannot be applied. Instead, all data within each set is concatenated and used for training and testing, respectively.

3.2.1.15 *Sleep-EDF Database Expanded*

Sleep-EDF Expanded dataset is an extended version of Sleep-EDF dataset on the Physiobank [9, 38, 61]. The dataset contains 197 whole-night polysomnographic sleep recordings from both healthy subjects and individuals with with mild difficulty falling asleep. Each recording includes two-channel EEG signals, an EOG signal (horizontal), an EMG signal (chin), along with expert

annotations of sleep stages for every 30-second segment, based on AASM guidelines [59].

For this study, only the EEG data from healthy subjects were used. Sleep is categorized into five sleep stages with distinct patterns of electrical brain activity: Wake(W), Non-Rapid Eye Movement stage (NREM) which consists of three stages (N1, N2 and N3) and Rapid Eye Movement stage (REM). Each segment is manually classified into one of eight classes: W, N1, N2, N3, N4, REM, MOVEMENT, or UNKNOWN.

Following Supratak et al. [113], the N3 and N4 stages were merged into a single stage N3. At the beginning and end of each recording, there were long periods of wakefulness (stage W) when the subject was not sleeping. Only 30 minutes of these wake periods before and after sleep were included, as the primary focus was on sleep phases. Movement artifacts labeled as UNKNOWN or MOVEMENT were excluded, as they were not relevant to the five sleep stages. For this dataset, sliding windows were not applied due to the large size of the data (whole-night recordings) and limited computational resources. Instead, the data was segmented into standard 30-second epochs, as this is sufficient for sleep stage classification. The modes' performance were evaluated using subject-wise 3-fold cross-validation.

3.2.2 *Deep Learning Model Implementation*

In the following sections, we briefly introduce the neural networks evaluated in this study. 11 end-to-end DL models were selected. A total of 11 end-to-end deep learning (DL) models were selected for comparison. A more detailed description of each model is provided below.

3.2.2.1 *EEGNet*

EEGNet [67] is a compact convolutional neural network designed for EEG-based BCIs which can generalize across various BCI paradigms, including P300 visual-evoked potentials, error-related negativity (ERN), movement-related cortical potentials (MRCP), and sensory motor rhythms (SMR). The model uses depthwise and separable convolution to minimize the number of parameters while still encapsulating key EEG-specific feature extraction concepts like spatial filtering and filter-bank construction.

EEGNet processes EEG signals through three main blocks. First, the temporal convolution layer only extracts frequency-specific features. Next, the depthwise convolution operates on each feature map separately to extract frequency-specific spatial filters. Lastly, the separable convolution, a comprising a depthwise and pointwise convolution, combines temporal and spatial filtering to reduce parameters and integrate features effectively. Each convolution layer is followed by a softmax classification layer. Regularization techniques such as dropout and batch are applied at every layer to improve training stability and prevent overfitting.

3.2.2.2 *DeepConvNet*

DeepConvNet [99] is a deep convolutional neural network designed for decoding raw EEG signals. It excels at end-to-end learning, enabling automated feature extraction and classification without reliance on expert knowledge, while achieving competitive accuracies.

The network's architecture is built to extract a wide range of features using multiple convolutional and pooling layers. It consists of four connvolution-max-pooling blocks, with a first block designed to handle EEG input, followed by three standard convolution-max-pooling blocks and a dense softmax classification layer. The specialized first block is divided into two layers: a temporal convolution layer which extracts frequency-specific features by convolving across time, and a spatial filtering layer that applies spatial filters across electrodes to capture inter-channel relationships.

DeepConvNet employs Exponential Linear Units (ELUs) as activation functions, applies batch normalization to the outputs of convolutional layers before the nonlinearity, and uses dropout to randomly set some layer inputs to zero during each training update.

3.2.2.3 *ShallowConvNet*

ShallowConvNet [99] is a simpler convolutional neural network inspired by the *Filter Bank Common Spatial Patterns* (FBCSP) pipeline. It is tailored to extract band-power features specifically and is designed with fewer layers to optimize these transformations.

The network consists of two layers, a temporal convolution and a spatial filter, as in the deep ConvNet, analog to to the bandpass and CSP spatial filter steps in FBCSP. Compared to DeepConvNet, the temporal convolution of the shallow ConvNet had a larger kernel size to process a broader temporal context. After these layers, the output undergoes a squaring nonlinearity and a mean pooling layer, which aggregate the squared activations over time. This is followed by a logarithmic activation function, analogous to the log-variance computation step in FBCSP. These components allow the network to mimic key steps in traditional EEG feature extraction, embedding them directly within the neural network.

Unlike DeepConvNet, ShallowConvNet's architecture is specialized for band power decoding and thus includes fewer layers. However, it benefits from the ability to optimize all processing steps jointly through backpropagation, unlike the separate stages in traditional methods.

3.2.2.4 *CNN-FC*

The model developed by Dose et al. [32] combines CNNs for feature extraction and a Fully Connected (FC) layer for classification. It consists of two convolutional layers with 40 kernels per layer. The first convolutional layer operates only performs convolution along the time axis, serving as a linear pre-filtering for each EEG channel. Second convolutional layer operates spatial convolution along the EEG channel axis, effectively reducing the number

of channels to a single value per time step. The valid padding ensures no extrapolation beyond the data.

An average pooling layer reduces the dimensionality of the output from the convolutional layers. The resulting pooled features are flattened into a 1D vector, which is then passed to a fully connected layer with 80 neurons. ReLU is used in all layers except the output, which uses a softmax activation function for classification.

3.2.2.5 *CNN-LSTM*

The dual-stream spatio-temporal neural network processes EEG data with a hybrid architecture that combines convolutional layers for feature extraction and recurrent layers for temporal modeling [32]. Similar to the CNN-FC model, the network has two convolutional layers with 40 filters each, followed by average pooling and flattening operations.

After convolution, an average pooling layer reduces the temporal resolution and the output is flattened along the temporal dimension using a time-distributed flattening layer. Sequential features are then processed by a single LSTM layer with 40 units and a sigmoid activation function, capturing temporal dependencies. This layer outputs the full sequence instead of only the final timestep's features.

Finally, the model includes a dense layer with a softmax activation function, producing probabilities for classification. Regularization is incorporated using $L1$ regularization in the convolutional layers and dropout layers in the recurrent components. The model is optimized using categorical cross-entropy loss with the RMSprop optimizer.

3.2.2.6 *MMCNN*

Multi-branch Multi-scale Convolutional Neural Network (MMCNN) [57] is an end-to-end model designed for motor imagery classification. The model provides a solution for the problems of subject variability and time differences in EEG data by incorporating multi-branch and multi-scale convolutional structures.

The model comprises five parallel branches, each implemented as an EEG Inception Network (EIN). Each EIN is composed of three key components: an EEG Inception Block (EIB), a Residual Block, and a Squeeze and Excitation (SE) Block.

The EIB uses multi-scale convolution kernels of increasing sizes to capture features at different frequency scales. The Exponential Linear Unit (ELU) activation function is employed to mitigate gradient vanishing issues and improve robustness to noise.

The Residual Block is designed to prevent the degradation problem in deeper networks by introducing shortcut connections. It consists of two branches: executes a sequence of operations interleaved with 1D convolutional layers and Batch Normalization (BN) layers; and the other branch includes a direct shortcut connection to bypass certain layers.

The Squeeze-and-Excitation (SE) Block enhances the model's attention to important features by adaptively re-weighting feature maps. It comprises two components: the *squeeze* operation, which captures channel dependencies, and the *excitation* operation, which learns sample-specific activations for each channel through a channel-dependent self-gating mechanism.

3.2.2.7 *ChronoNet*

ChronoNet [92] is a deep recurrent neural network developed for abnormal EEG classification, combining the features from both convolutional and recurrent architectures. The network is constructed by stacking multiple 1D convolutional layers (Conv1D) followed by multiple Gated Recurrent Unit (GRU) layers.

Each inception-style Conv1D layer has multiple filters with exponentially varying lengths to capture features over multiple temporal scales. These convolutional layers are followed by stacked GRU layers, which model both short- and long-term dependencies in EEG data. The GRU layers are densely connected in a feed-forward manner, meaning the output of each GRU layer is passed as input to all subsequent GRU layers. This dense connectivity strengthens feature propagation, facilitates feature reuse, as well as helps prevent issues such as vanishing or exploding gradients, which can degrade training accuracy.

To aggregate features from varying temporal resolutions, a Filter-Concat layer concatenates the outputs of Conv1D layers with different filter lengths along the depth axis. The network concludes with a softmax layer for classification tasks.

3.2.2.8 *Attention-based-1D-CNN*

The Attention-based-1D-CNN architecture [86] is specifically designed for mental workload classification using EEG signals. It combines CNNs and attention mechanisms to extract temporal and spatial features while emphasizing relevant information within EEG data.

The network consists of four 1D convolutional layers, applied alternately on time and channel axes. Temporal features are extracted using horizontal convolutional filters along the time axis, whereas spatial features are captured using vertical convolutional filters along the channel axis. After each 1D-CNN block, feature-level self-attention blocks are applied to learn attention weights for each time step and EEG channel. With an attempt to focus on critical features while suppressing irrelevant information, these blocks assign weights to different parts of the input by computing a weighted sum and passing it to the next layer.

The model employs LeakyReLU activation function with $\alpha = 0.001$ and global average pooling after each convolutional layer. The pooled features are then passed to a dense layer with 100 neurons and a sigmoid activation function to compute the contribution of particular input parts to the learned features. These 1D-CNN blocks and attention layers are repeated four times.

Following the 1D-CNN layers, fully connected layers are utilized with different number of neurons. To prevent overfitting, L1 and L2 regularization are applied to penalize large weights, while Dropout is incorporated to randomly deactivate neurons during training. The final layer applies a softmax activation function for classification probabilities.

3.2.2.9 *EEGTCNet*

EEG-TCNet [53] is a temporal convolutional network which can achieve high classification accuracy with few trainable parameters. It combines the efficient feature extraction capabilities of EEGNet with temporal modeling strengths of Temporal Convolutional Networks (TCNs).

The network starts with 2D temporal convolution layer to extract frequency features, followed by a depthwise convolution layer to learn frequency-specific spatial features. The separable convolution is then applied to summarize temporal features for each spatial filter and to mix feature maps across channels. Dropout and average pooling are used after depthwise and separable convolutions for regularization and feature downsampling.

Following the EEGNet-inspired layers, the Temporal Convolutional Network (TCN) module is applied to process the remaining temporal features. The TCNs model sequential data by combining causal and dilated convolutions. *Causal convolutions* produce outputs of the same length as the inputs by using 1D fully-convolutional networks with zero-padding. They ensure that outputs depend only on current and past inputs, preserving temporal order and preventing future information from influencing past data. *Dilated Convolutions* expand the receptive field exponentially with network depth by introducing gaps between kernel elements, enabling the network to capture long-term dependencies without increasing depth or kernel size.

TCNs are structured as stacks of *residual blocks*, each containing two layers of causal dilated convolutions with batch normalization, ELU activation functions, and dropout applied between layers. Skip connections within residual blocks enhance gradient flow and allow deeper architectures.

Finally, the TCN output is fed into a fully connected dense layer with a softmax activation function for classification.

3.2.2.10 *BLSTM-LSTM*

The BLSTM-LSTM model [21] is designed for EEG-based mental workload classification. It combines BLSTM and LSTM networks to capture both past and future dependencies in sequential EEG data.

The architecture contains a single BLSTM layer with 256 neurons and two stacked LSTM layers with 128 and 64 neurons, respectively. The BLSTM layer reads the input in both forward and backward directions to capture contextual information from both the past and future. The stacked LSTM layers further refine sequential dependencies from the BLSTM output. Dropout layers with a rate of 0.2 and batch normalization layers are applied after the BLSTM and each LSTM layer.

Following the recurrent layers, the model incorporates two fully connected (dense) layers with 32 and 3 neurons, respectively. The final dense layer uses a softmax activation function for classification.

3.2.2.11 *DeepSleepNet*

DeepSleepNet [113] is a deep learning model specifically designed for automatic sleep stage scoring using raw single-channel EEG data. The model integrates CNNs to extract time-invariant features and bidirectional-LSTMs to learn transition rules among sleep stages. It consists of two main parts: *Representation Learning* and *Sequence Residual Learning*.

The *Representation Learning* component employs two parallel CNNs with different filter sizes at the first layers. The small filter focuses on capturing temporal patterns, while the larger filter extracts frequency components. Each CNN comprises four convolutional layers and two max-pooling layers, with operations including 1D convolution, batch normalization, and ReLU activation. The outputs of the two CNNs are concatenated to form feature representations. For the *Sequence Residual Learning* part, two layers of bidirectional-LSTMs are used to model sleep stage transition rules. A shortcut connection with a FC to add LSTM outputs into CNNs features.

In order to effectively train model end-to-end via backpropagation and prevent class imbalance problem in large sleep datasets, a two-step training algorithm is applied. The first step is to perform a supervised pre-training on the representation learning part of the model using a balanced dataset to address class imbalance. In the fine-tuning step, the entire model, including the sequence learning component, is trained on sequential EEG data. To prevent overfitting, two regularization techniques are utilized: Dropout layers with a probability of 0.5 in every layer and L2 weight decay in the initial CNN layers.

3.2.3 *Experimental Setup*

All models were trained on a VX-3 computational server, equipped with 20 CPU cores, 80GB RAM, and an NVIDIA Tesla V100 GPU with 32GB VRAM. The server operates on Ubuntu 22.04.4 LTS and features a 120GB root partition along with a 1TB SSD for data storage. Training was conducted using TensorFlow 2.15.0 [1] with the Keras API, configured for compatibility with CUDA 12.0 and NVIDIA Driver Version 550.54.15 to enable GPU-accelerated computations.

The training hyperparameters, including the optimization method and learning rate, were configured based on the settings provided in the original publications for each model. Most models utilized the Adam optimizer, while the CNN-LSTM model employed RMSprop. The learning rates alternated between 0.0001 and 0.001. Additionally, the batch size was set to 16, and training was conducted for 100 epochs. The training time varied significantly depending on the model-dataset pair.

3.2.4 *Subset Selection*

We decided on a subset size of 5, meaning all combinations of 5 variables from the 17 datasets were evaluated based on their performance across 11 deep learning (DL) models.

To define an appropriate target metric for evaluation, the median was chosen over the mean for its robustness against extreme values and its stability in the presence of high variability across datasets. Moreover, since the datasets included both balanced and unbalanced class distributions, $F1$-score was selected as a more reliable metric than accuracy. Ultimately, the target metric was defined as the median $F1$-score for each model across all datasets.

The target metric y and feature matrix X are organized as follows, adhering to a tabular format. Each row corresponds to a specific model, while each column represents a feature or the target value:

Model	X_1	X_2	X_3	X_4	X_5	y
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	y_1
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	y_2
3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	y_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	X_{111}	X_{112}	X_{113}	X_{114}	X_{115}	y_{11}

Here:

- The columns X_1, X_2, X_3, X_4, X_5 represent the features of the models.
- The last column y contains the target values corresponding to each model.

The resulting table has a shape of $(11, 7)$, where 11 represents the number of models, and 7 represents the total number of variables (5 features and 1 target).

To evaluate the effectiveness of the proposed subset selection procedure, We conducted an experiment with 2×2 factorial design to explore the impact of the target metric computation (include vs. exclude datasets in the candidate subset) and preprocessing step of dataset pool (with vs. without clustering and pruning) on the subset selection procedure. This design examines the combined effects of two independent factors, each with two levels:

1. Target Metric Computation
    - Inclusion: The target metric is computed as the median $F1$-score across all datasets, including the subset being evaluated.
    - Exclusion: The target metric is computed as the median $F1$-score across all datasets, excluding the subset being evaluated.

2. Dataset Corpus Preprocessing
    - Without Clustering/Pruning: The subset selection procedure is applied directly to the full dataset pool without any preprocessing.
    - A correlation analysis is performed to group highly correlated datasets into clusters, and only representative datasets are retained before applying the subset selection procedure.

The above factorial design resulted in four following experimental conditions:

1. Full Dataset Corpus + Inclusion: Evaluating subsets using target metric including subsets without preprocessing the dataset corpus.
2. Full Dataset Corpus + Exclusion: Evaluating subsets using target metric excluding subsets without preprocessing the dataset corpus.
3. Pruned Dataset Corpus + Inclusion: Evaluating subsets using target metric including subsets with clustering and pruning the dataset corpus.
4. Pruned Dataset Corpus + Exclusion: Evaluating subsets using target metric including subsets with clustering and pruning the dataset corpus.

4

RESULTS

This chapter provides an objective presentation of the study results. First, an overview of the performance of various neural network models across the datasets is presented. Next, the generalizability of DL architectures is examined across diverse domains, evaluating their performance on tasks outside their original domains. Generalizability is assessed based on lower variance or standard deviation of F_1-scores, which indicates more consistent performance across datasets. All decoding metrics mentioned in this section were caculated as mean F_1-score.

Additionally, the study explores the compatibility of different DL architectures with various types of EEG data, giving insights into the most effective pairings of architectures with specific EEG signal domains. Finally, the results of the subset search on the EEG dataset collection are presented, using the median F_1-score as the target metric for evaluation.

4.1 COMPREHENSIVE PERFORMANCE EVALUATION

In total, the performance of 11 models was evaluated across 17 datasets. Figure 4.1 presents a heatmap displaying the average F_1-score for each model-dataset pair. Table 4.1 and 4.2 provide key summary statistics for 11 DL models and 17 EEG datasets, respectively. The statistics include the *Mean*, *Standard Deviation*,*Median* of performance of each particular model and dataset.

4.1.1 *Performance Evaluation of Models across Datasets*

Figure 4.2 provides an overview of model performance based on their mean performance and variability. What stands out the most is that half of selected models demonstrate consistently comparable performance, as seen by their clustering in the bottom-left corner of the plot, with low variability and moderate mean F_1-score. Only a few models are scattered more broadly across the rest of the plot.

The top-performing models include EEGNet and ShallowConvNet, which achieve the highest average mean values ($\overline{M} = 0.69$). By comparison, BLSTM-LSTM and DeepSleepNet models are marked by instability, with weaker performance and more scattered results, ($\overline{M} = 0.54, SD = 0.18$) and ($\overline{M} = 0.56, SD = 0.18$), respectively. The median values align closely with mean values, reflecting symmetric distributions in the model metrics.

As mentioned earlier, it is intriguing to analyze how well the selected models generalize, which can be partly assessed through the standard deviation of their performance across datasets. A model demonstrates better generalizability when the variability in metrics between datasets is low. However,

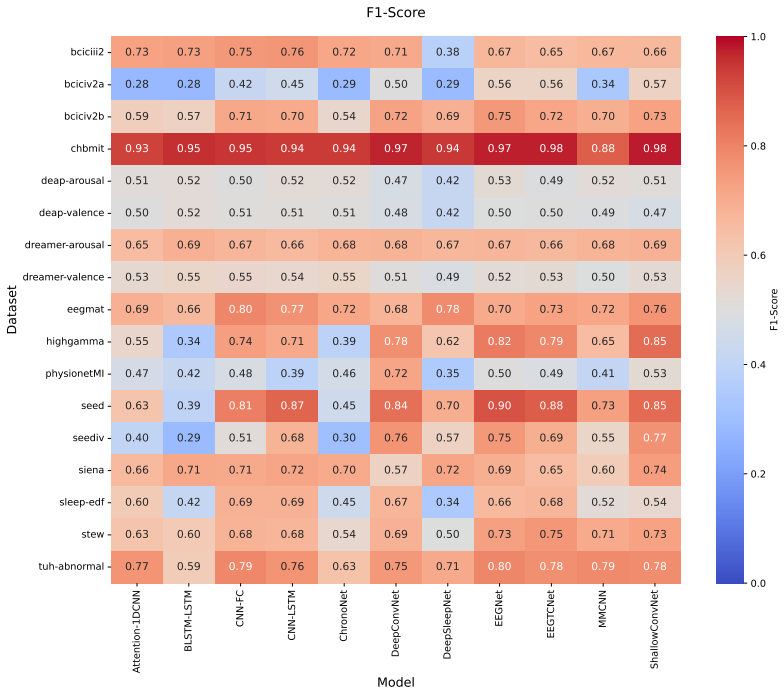

Figure 4.1: Heatmap of F_1-scores Across All Datasets and Models

the analysis shows that no single model significantly outperforms the others in this regard, as the standard deviation ranges narrowly between 0.14 and 0.18. Figure 4.1 also clearly illustrates that no model exhibits a single-tone color representation but instead a strong mix of colors, reflecting significant variability in performance across datasets.

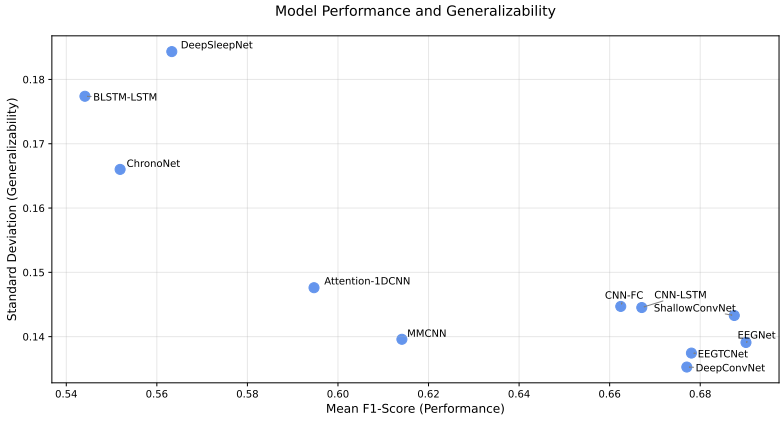

Figure 4.2: Evaluation of Models: Performance and Generalization Ability

4.1.2 *Performance Evaluation of Datasets across Models*

Figure 4.3 illustrates the distribution of F_1-score for each dataset using boxplot, summarizing how the selected models perform on the datasets. A diversity in central tendency, box lengths, whiskers, and distribution of outliers

Table 4.1: Summary of Model Performance Across Datasets (F_1-score)

Model	Mean	Std	Median
Attention-1DCNN	0.59	0.15	0.60
BLSTM-LSTM	0.54	0.18	0.55
CNN-FC	0.66	0.14	0.69
CNN-LSTM	0.67	0.14	0.69
ChronoNet	0.55	0.17	0.54
DeepConvNet	0.68	0.14	0.69
DeepSleepNet	0.56	0.18	0.57
EEGNet	0.69	0.14	0.69
EEGTCNet	0.68	0.14	0.68
MMCNN	0.61	0.14	0.65
ShallowConvNet	0.69	0.14	0.73

between datasets is observed. Some boxplots such as CHB-MIT, DREAMER and DEAP display compact distributions with narrow interquartile ranges (IQRs), suggesting stable performance across models. In contrast, other datasets such as SEED, SEEDIV and BCIC-IV-2a show wider IQRs and extended whiskers, reflecting more dispered results, where some models achieve high scores while others score much lower. Outliers are also visible in certain datasets, such as STEW, Sleep-EDF, and High-Gamma, where specific models perform much better or worse compared to the majority.

It is noticeable that the CHB-MIT dataset shows the highest average performance and a low variability ($\overline{M} = 0.95, SD = 0.03$), indicating consistently high results across models. In contrast, BCIC-IV-2a has the lowest mean F_1-score and moderate variability ($\overline{M} = 0.41, SD = 0.12$), which is an unexpected result given its frequent use as a benchmark dataset in ML and DL studies for EEG data.

While the two dimensions of the DEAP dataset share similar statistical properties, with comparable mean values ($\overline{M}_{valence} = 0.49$, $\overline{M}_{arousal} = 0.5$) and identical standard deviations ($SD = 0.03$), the corresponding dimensions of the DREAMER dataset exhibit distinct statistical patterns, with mean values of 0.53 for valence and 0.67 for arousal.

The DREAMER Dataset achieves relatively consistent performance across models, with the lowest variability for both dimensions ($SD_{valence} = 0.02$, $SD_{arousal} = 0.01$). On the other hand, although the SEED and SEED-IV datasets have high average performance, considerable variations in model performances are observed through the highest variability ($SD_{valence}$ = $SD_{arousal}$ = 0.18).

The median F_1-score are typically aligned with the mean value, with a few exceptions. For instance, High-Gamma ($Median = 0.71$) and SEED ($Median = 0.81$) show higher medians than their respective means, probably due to skewed distribution.

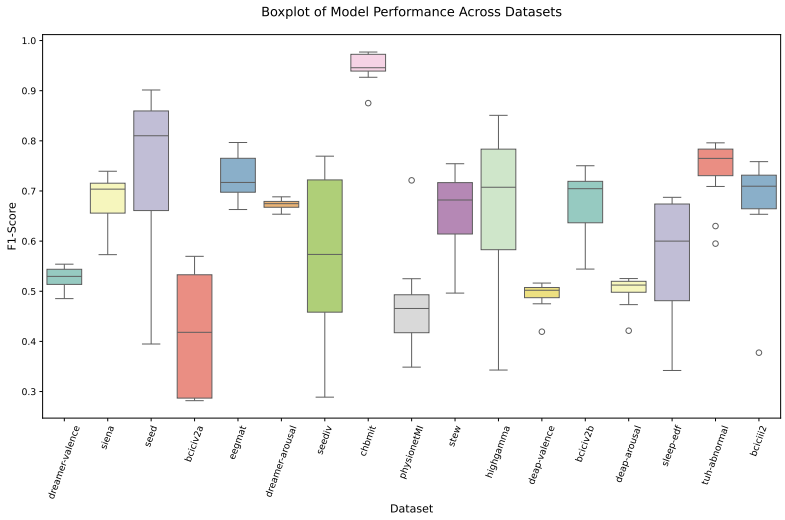

Figure 4.3: Boxplot of Model Performance Across Datasets

4.1.3 *Model-Dataset Compatibility Analysis*

In the original studies, each DL model was specifically tailored for a distinct category of EEG data. Within the scope of this study, of exploratory interest is how these models perform on their corresponding datasets, whether they can also classify datasets from other categories effectively, as well as how frequently a particular model achieves the best performance across datasets.

Table 4.1 highlights the compatibility between different DL architectures and various types of EEG data, that is to say, which model performs best for specific EEG datasets. Interestingly, none of the models achieve their best performance on the benchmark datasets for which they were originally developed.

From the table, it is evident that ShallowConvNet emerges as the most compatible model, outperforming other models on five datasets, three of which belong to the motor imagery category. Beyond that, ShallowConvNet also performs well in other domains, including epilepsy detection (CHB-MIT, SIENA) and emotion recognition (SEED-IV). EEGNet follows closely, delivering the best performance in four datasets from diverse domains. These results suggest that both models are relatively effective at capturing relevant features and are well-suited for handling EEG datasets across multiple domains. As such, they can be considered strong candidates for general-purpose EEG analysis tasks, especially when dataset characteristics are unknown.

Models with moderate compatibility account for BLSTM-LSTM and CNN-FC, which achieve the best performance on 3 and 2 datasets, respectively. CNN-FC performs better than expected, surpassing the random baseline on several datasets. In contrast, BLSTM-LSTM exhibits high inconsistency, excelling on DEAP and DREAMER datasets but ending up at the bottom in performance on five other datasets, with results even below the random baseline.

Table 4.2: Summary Statistics of F_1-scores by Dataset

Dataset	Mean	Std	Median
BCIC-III-2	0.68	0.11	0.71
BCIC-IV-2a	0.41	0.12	0.42
BCIC-IV-2b	0.67	0.07	0.70
CHB-MIT	0.95	0.03	0.95
DEAP (arousal)	0.50	0.03	0.51
DEAP (valence)	0.49	0.03	0.50
DREAMER (arousal)	0.67	0.01	0.67
DREAMER (valence)	0.53	0.02	0.53
EEGMAT	0.73	0.04	0.72
High-Gamma	0.66	0.17	0.71
PhysionetMI	0.47	0.10	0.47
SEED	0.73	0.18	0.81
SEEDIV	0.57	0.18	0.57
SIENA	0.68	0.05	0.70
SLEEP-EDF	0.57	0.12	0.60
STEW	0.66	0.08	0.68
TUH-Abnormal	0.74	0.07	0.77

Meanwhile, with outstanding performance on only one dataset, CNN-LSTM, DeepConvNet, and EEG-TCNet show potential for a particular type of datasets but lack broad compatibility. Unexpectedly, they do not perform best on the tasks they were originally designed for. It is also worth highlighting that DeepSleepNet performs the worst across seven datasets, indicating a limited ability to classify EEG signals effectively.

4.2 SUBSET SELECTION

Before subset selection procedure, correlation analysis was conducted to check multi correlation between datasets to avoid multilcollinearity problem.

4.2.1 *Correlation Analysis*

Figure 4.4 visualizes the correlation matrix across datasets. The correlation matrix visually represents the pairwise relationships between datasets, with correlation value ranging from -1 (strong negative correlation) to 1 (strong positive correlation).

At first glance, what immediately strikes the eye are the dark warm-colored blocks, indicating that several datasets highly correlate with one another. An-

Table 4.3: Best Performing Models for Each EEG Dataset

Dataset	**Model**	F_1**-score**
BCIC-III-2	CNN-LSTM	0.76
BCIC-IV-2a	ShallowConvNet	0.57
BCIC-IV-2b	EEGNet	0.75
CHB-MIT	ShallowConvNet	0.98
DEAP (arousal)	EEGNet	0.53
DEAP (valence)	BLSTM-LSTM	0.52
DREAMER (arousal)	BLSTM-LSTM	0.69
DREAMER (valence)	BLSTM-LSTM	0.55
EEGMAT	CNN-FC	0.80
High-Gamma	ShallowConvNet	0.85
PhysionetMI	DeepConvNet	0.72
SEED	EEGNet	0.90
SEEDIV	ShallowConvNet	0.77
SIENA	ShallowConvNet	0.74
SLEEP-EDF	CNN-FC	0.69
STEW	EEGTCNet	0.75
TUH-Abnormal	EEGNet	0.80

other noteworthy observation is that the dark-shaded blocks (strong correlation) are not evenly spread across the matrix but are instead concentrated in specific rows and columns, forming distinct clusters where datasets share highly similar behaviors. For instance, datasets such as SEED, SEEDIV, BCIC-IV-2b and High-Gamma exhibit extremely high multicorrelation (greater than 0.9), collectively reflecting overlapping trends in model performance. In other words, models performing well on one of these datasets are likely to generalize effectively to the others. On the other hand, certain dataset pairs, such as CHB-MIT with EEGMAT and BCIC-III-2 with TUH-Abnormal, show weak or slightly negative correlations with others, underscoring their unique characteristics and the potential need for specialized approaches. It further draws attention to the relationship within the DREAMER dataset, where two dimensions show no correlation with each other, in contrast to the high correlation observed between these dimensions in the DEAP dataset.

Figure 4.5 presents the hierarchical clustering dendrogram, revealing patterns of similarity and grouping based on model performance tendencies. The y-axis represents the distance or dissimilarity between clusters, with lower heights indicating stronger similarity between datasets. The dendrogram splits the datasets into two primary clusters. While Cluster 1 (orange) show relative strong similarity, as evidenced by their clustering at lower heights, the Cluster 2 (green) display more internal variability, as it subclusters are formed at greater heights.

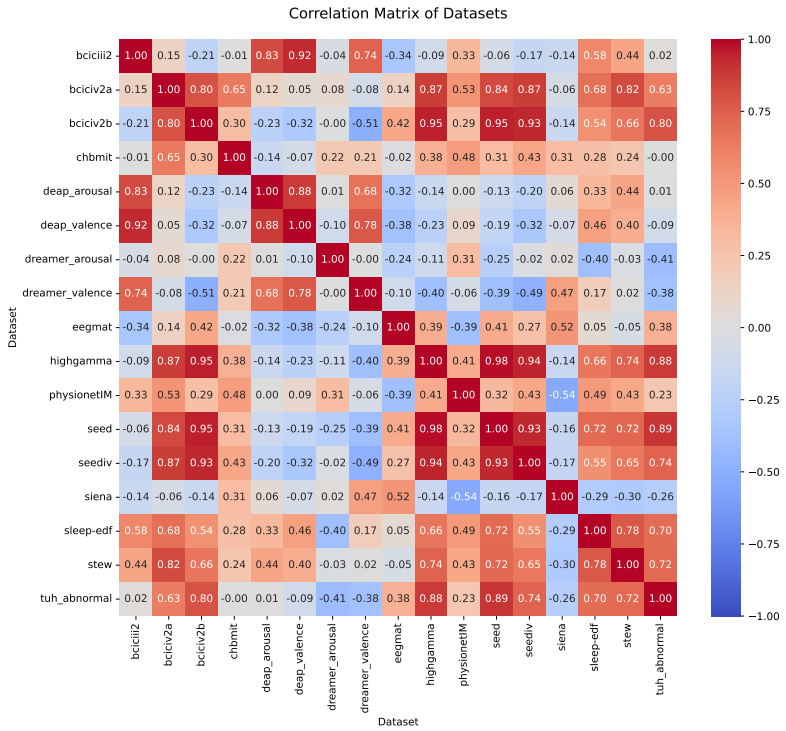

Figure 4.4: Correlation Matrix of Datasets Based on F_1-score

After clustering, a representative dataset for each cluster was selected based on the highest average similarity within that cluster. The list of low-correlated datasets includes SEED, STEW, BCIC-III-2, DEAP (arousal), DEAP (valence), EEGMAT, SIENA, DREAMER (arousal), DREAMER (valence), CHB-MIT, and PhysionetMI. In Cluster 1, the datasets in the left branch are closely linked, leading to the selection of SEED as the representative dataset for this group. For the right branch of Cluster 1, STEW was chosen as the representative dataset. In Cluster 2, the datasets display greater dissimilarity, reflecting more distinct characteristics. As a result, more representative datasets were selected to capture the diversity within the cluster.

4.2.2 *Evaluation of Subset Selection Procedures*

The subset selection procedure was proceeded under the four experimental conditions outlined earlier in Chapter 3. The subsets generated under each condition were assessed in terms of prediction error (Mean Squared Error, MSE) and the diversity of selected datasets, as evidenced by pairwise correlations among datasets.

Table 4.4 summarizes the Mean Squared Error (MSE) obtained from the representative subsets of EEG datasets selected under each experimental condition. As shown, the subset generated using the *full dataset corpus* while applying the *exclusion* approach achieves the lowest MSE (0.001), resulting in the most accurate prediction performance. When using the *full dataset corpus* for the selection process, the *exclusion* condition obtains a slightly

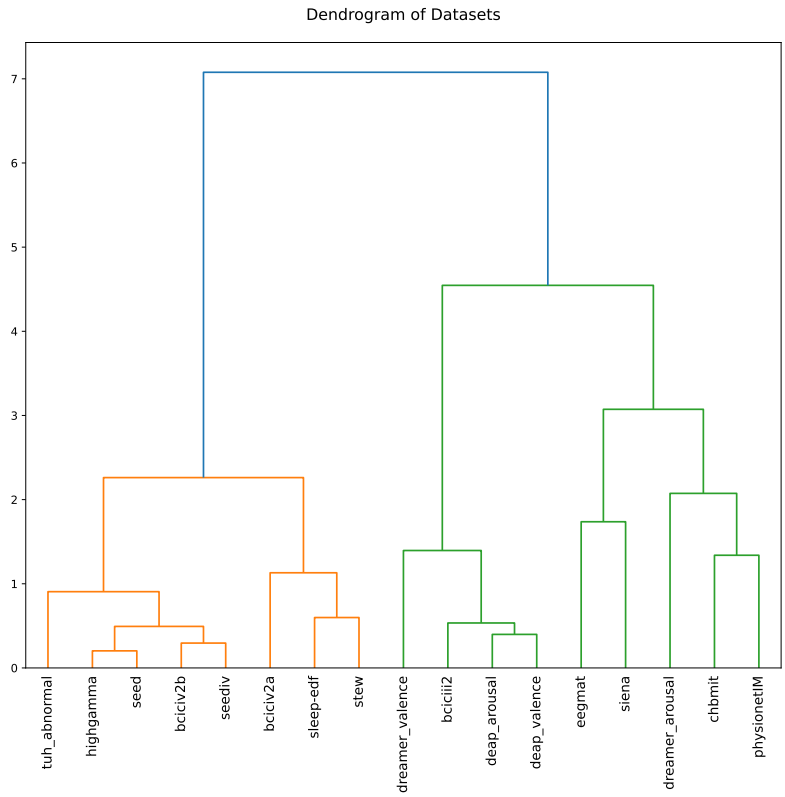

Figure 4.5: Hierarchical Clustering Dendrogram of Datasets Based on Similarity in Model Performance

lower MSE than the *inclusion* condition, 0.0001 and 0.0002 respectively. The same tendency is observed in the *pruned dataset corpus* with a lower MSE for the *exclusion* condition. It can be concluded that across both dataset corpora, the *exclusion* approach consistently yields lower MSE values compared to the *inclusion* condition. Using the *full dataset corpus* also results in lower MSE values for both approaches. For each condition, there is an overlap in the EEG datasets selected in the representative subsets, such as BCIC-III-2, SEED, EEGMAT, DEAP and DREAMER.

To further evaluate the diversity and redundancy of the selected subsets under each condition, we analyzed their correlation matrices. Weak correlations among datasets imply that each dataset in the subset has unique characteristics, contributing to the diversity of the subset. Conversely, strong correlations suggest potential redundancy, where datasets likely share similar information, thereby reducing the overall diversity of the subset. Figure 4.6, 4.7, 4.8 and 4.9 visualize the correlation matrices for the selected subsets under the four experimental conditions: *Full Dataset Corpus + Inclusion*, *Full Dataset Corpus + Exclusion*, *Pruned Dataset Corpus + Inclusion*, and *Pruned Dataset Corpus + Exclusion*, respectively. It can be observed that some datasets within the subset are highly correlated with each other. The presence of highly correlated datasets may reduce the interpretability of the subset's prediction performance.

Table 4.4: Results of Four Experimental Conditions for Representative Subset Selection

Condition	**Representative Subset**	**Mean Squared Error (MSE)**
Full Dataset Corpus + Inclusion	BCIC-III-2, BCIC-IV-2b, DEAP (valence), EEGMAT, STEW	0.0002
Full Dataset Corpus + Exclusion	BCIC-III-2, BCIC-IV-2a, DREAMER (arousal), High-Gamma, SIENA	0.0001
Pruned Dataset Corpus + Inclusion	SEED, DREAMER (valence), EEGMAT, CHB-MIT, DREAMER (arousal)	0.001
Pruned Dataset Corpus + Exclusion	SEED, BCIC-IV-2a, DEAP (valence), DREAMER (valence), PhysionetMI	0.0004

The subset under the *Full Dataset Corpus + Inclusion* condition includes a mix of datasets with diverse characteristics, as evidenced by weakly correlated datasets. However, it suffers from redundancy due to extremely high correlation between BCIC-III-2 and DEAP (valence) (0.92).

Similarly, the correlation matrix under the *Full Dataset Corpus + Exclusion* condition highlights a relatively diverse subset of EEG datasets, except for the broad overlap in behavior between BCIC-IV-2a and High-Gamma (0.87).

In contrast to the above two conditions, four out of five datasets under the *Pruned Dataset Corpus + Exclusion* condition exhibit strong correlations with others, indicating limited diversity despite the preprocessing.

Remarkably, the *Pruned Dataset Corpus + Inclusion* condition stands out as the only configuration with no high correlations among datasets, underscoring its effectiveness in selecting datasets with strong diversity.

This brings forth an intriguing question: Why the inclusion approach performs so distinctly better than the exclusion approach in the presence of clustering and pruning?

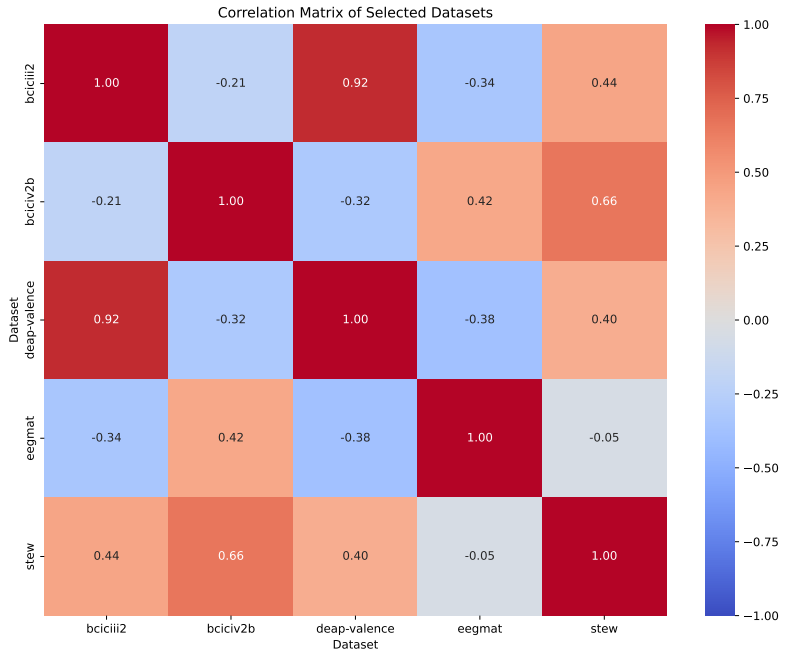

Figure 4.6: Correlation Matrix of Datasets in the Selected Subsets under Full Dataset Corpus + Inclusion Condition

Figure 4.7: Correlation Matrix of Datasets in the Selected Subsets under Full Dataset Corpus + Exclusion Condition

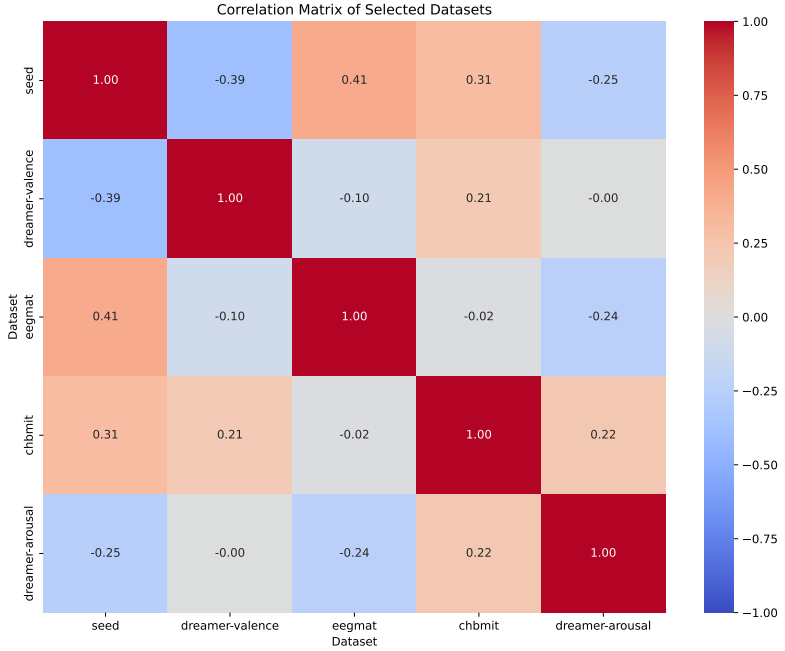

Figure 4.8: Correlation Matrix of Datasets in the Selected Subsets under Pruned Dataset Corpus + Inclusion Condition

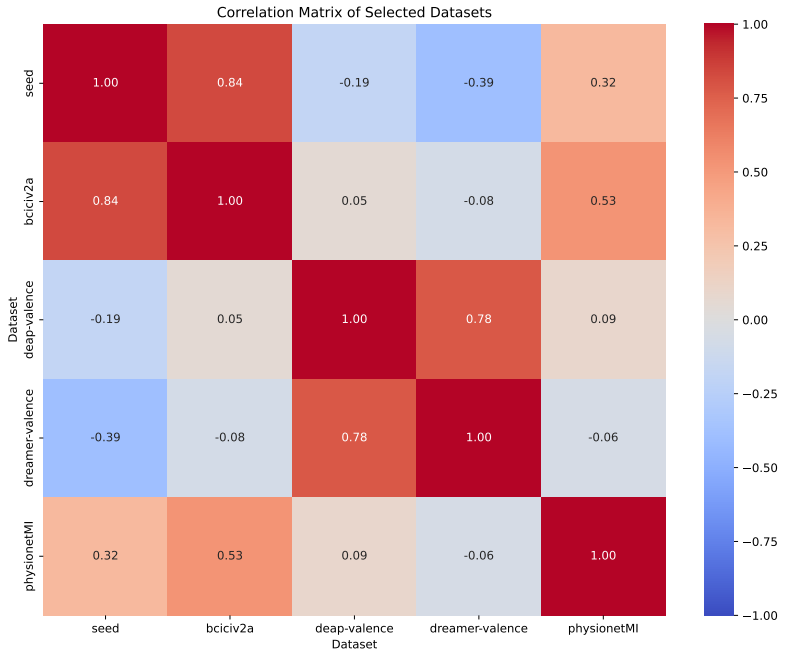

Figure 4.9: Correlation Matrix of Datasets in the Selected Subsets under Pruned Dataset Corpus + Exclusion Condition

5 DISCUSSION

This chapter discusses the results presented in Chapter 4 in relation to the study's objectives and research questions, evaluates their significance, discuss about limitations, and identifies potential directions for future research.

5.1 INTERPRETATION OF RESULTS

5.1.1 *Generalizability of DL Models*

The study revealed that no single DL model demonstrated outstanding generalizability across all EEG datasets, as evidenced by the narrow range of standard deviation values between 0.14 and 0.18 in model performance.

Models such as EEGNet and ShallowConvNet demonstrated low variability in F_1-scores and consistent ranking in performance across EEG datasets, highlighting their potential as general-purpose models for diverse EEG tasks. While MMCNN was proposed as a general framework for EEG classification and has shown competitive performance in MI tasks, our findings suggest its adaptability to other EEG domains remains limited.

Even the best-performing models displayed variability when applied to datasets outside their original domain, indicating a limitation in the generalization capabilities of existing architectures. This emphasizes the complex and dynamic nature of EEG signals, where unique characteristics such as noise, task-specific variability, and subject dependency challenge the robustness of DL models.

5.1.2 *Dataset-Specific Challenges*

The variability in model performance across datasets also revealed some dataset-specific characteristics. For instance, the CHB-MIT dataset exhibited the highest average performance with low variability ($\overline{M} = 0.95, SD = 0.03$), which might imply that its features are easier for models to decode consistently, although this interpretation remains speculative. Conversely, the BCIC-IV-2a dataset, despite being a common benchmark in EEG research, showed unexpectedly poor results ($\overline{M} = 0.41, SD = 0.12$). This finding challenges the assumption that frequently used datasets are necessarily well-suited for general DL evaluations.

The differing statistical patterns between the valence and arousal dimensions in datasets like DEAP and DREAMER further illustrate the variability in EEG datasets within the same category. While DEAP displayed similar performance trends across both dimensions, DREAMER showed greater per-

formance stability in the arousal dimension. Further analysis is necessary to better understand these differences and their implications.

5.1.3 *Compatibility of DL Models with EEG Domains*

The compatibility analysis revealed that EEGNet and ShallowConvNet consistently outperformed other models across multiple datasets, particularly in MI, epilepsy detection, and emotion recognition tasks. Their versatility suggests that these architectures effectively capture diverse EEG signal features.

Interestingly, most models did not perform best on the datasets for which they were originally developed. This finding raises questions about the transferability of domain-specific architectures. For instance, ShallowConvNet excelled on datasets beyond MI, such as epilepsy detection (CHB-MIT) and emotion recognition (SEED-IV). In contrast, DeepSleepNet demonstrated the lowest compatibility across datasets, likely due to its design being specifically tailored for sleep stage classification.

The inconsistency in BLSTM-LSTM's results, outperforming in DREAMER and DEAP datasets but underperforming in the others, reflects the influence of task complexity and dataset variability on model performance.

5.1.4 *Subset Selection Procedure*

The correlation analysis and hierarchical clustering revealed distinct clusters of highly correlated datasets, such as SEED, SEED-IV, and BCICIV2B. While these clusters indicate overlapping trends in model performance, their inclusion in the same subset risks redundancy, which could undermine the diversity required for robust model evaluation.

The selection of representative datasets from clusters balances diversity and representativeness. The observation that CHB-MIT and PhysionetMI exhibit weak correlations with other datasets underscores their potential as unique additions to the subset, ensuring broader coverage of EEG characteristics.

The evaluation of subset selection procedures confirmed that the **Pruned Dataset Corpus + Inclusion** condition achieved the highest diversity, with no high correlations among the selected datasets, although it did not yield the lowest MSE. This outcome is somewhat unexpected, as the Exclusion approach was anticipated to produce more robust and unbiased results. By ensuring that the target metrics remain unseen from the subset, the Exclusion approach is designed to avoid overly optimistic estimates, prevent data leakage, and provide a more rigorous evaluation.

However, the Exclusion approach appeared to behave inconsistently under clustering and pruning compared to the Inclusion approach. Without pruning, the Exclusion approach resulted in subsets with fewer correlated datasets than the Inclusion approach. Conversely, after pruning, the Exclusion approach produced subsets with higher correlations among the selected

datasets, while the Inclusion approach resulted in subsets with no correlations at all.

This inconsistency can be attributed to the interaction between pruning and the Exclusion approach. Pruning reduces the dataset pool to a smaller, more "representative" subset of clusters. However, excluding all datasets in the candidate subset may disrupt the balance and relationships between datasets established by pruning. The Exclusion approach may accidentally amplify existing correlations among the remaining datasets, as the excluded datasets may have played a role in balancing similarities or mitigating redundancy within the cluster structure. As a result, the selected subset includes more correlated datasets than expected.

Additionally, in the Exclusion approach, the regression model is optimized on a modified target metric that excludes candidate subset datasets. This partial target metric can alters the subset selection dynamics by over-prioritizing datasets that perform well on the remaining datasets, even if they are correlated with each other.

In contrast, the Inclusion approach keeps all datasets for target metric computation, preserving the diversity and relationships established during pruning. This ensures that the subset selection process considers the entire dataset pool.

The selected subset in this condition included datasets such as SEED, DREAMER-Valence, EEGMAT, CHB-MIT, and DREAMER-Arousal, which exhibited weak pairwise correlations, highlighting the subset's diversity.

5.2 LIMITATIONS

While this study provides valuable insights into the interaction between GANs models and EEG datasets, as well as proposes a systematic subset selection framework, several limitations must be acknowledged.

The relatively small scope of the study, along with limited time and available resources, restricts its potential to provide a comprehensive overview of EEG diversity and fully evaluate the performance of more diverse DL architectures. It is challenging to establish consistent benchmarks for evaluating deep learning (DL) models across diverse EEG datasets. External factors, including preprocessing differences, variations in dataset structures, and challenges in reproducibility, may have introduced variability, thereby impacting the accuracy of the results.

5.2.1 *EEG Datasets*

Several challenges arose during the dataset loading and preprocessing steps, which impacted the study results. Each dataset had its own unique structure and loading requirements, making it difficult to standardize the preprocessing pipeline. This lack of standardization likely influenced the reliability and comparability of the results.

Some datasets exhibited highly imbalanced class distributions, which could distort performance metrics, such as accuracy, making them less meaningful when comparing results across datasets with varying class distributions. As a result, conclusions drawn from these comparisons may be influenced by the underlying class distributions rather than the true generalizability of the models.

Take CHB-MIT as an example. This dataset exhibited exceptionally high performance metrics, but this does not imply that data is easier to classify. Instead, the imbalance in the dataset — where non-epileptic segments significantly outnumber epileptic ones — contributes to inflated metrics. Epileptic events typically last only about one minute, while non-epileptic recordings span approximately 30 minutes. Epilepsy detection, being a form of anomaly detection, skews results toward the majority class (non-epileptic), leading to overhyped accuracy. This imbalance complicates comparisons between DL models and makes it difficult to draw meaningful interpretations about CHB-MIT's actual classification difficulty.

The labor-intensive effort involved in extracting features for each dataset limited the ability to experiment with non-end-to-end deep learning models or other machine learning approaches that require explicit feature extraction before training. As a result, the study primarily focused on end-to-end DL models, leaving other potentially valuable methodologies unexplored.

5.2.2 *Deep Learning Models*

The selection of DL algorithms in this study, while diverse within the category of end-to-end deep learning models, lacks representation from other types of approaches. Applying additional machine learning and deep learning models that require feature extraction steps before training could offer greater understanding of model performance trends across datasets.

Moreover, more than half of the models included in this study were specifically designed for motor imagery tasks, as many of the EEG-based well-known and high-performing models are tailored for motor imagery classification. This may introduce a potential bias, as these models may perform disproportionately well on motor imagery datasets compared to other domains. Such a bias could increase the likelihood that motor imagery datasets are overrepresented in the selected subset.

The study also face reproducibility challenges, particularly with DL architectures that were too complex to implement. This limitation restricted the range of models included in the study, as some architectures were excluded due to their high implementation difficulty. Furthermore, the reproducibility of model performance was affected by differences in preprocessing techniques applied to the datasets. These variations may result in discrepancies between this study's findings and the results reported in the original papers.

5.2.3 *Performance Metrics*

Performance scores play a crucial role in the subset selection procedure, particularly for Ridge regression modeling. A key consideration is determining which metric should be used for modeling and evaluation, especially when datasets vary in characteristics such as the number of classes or degree of class imbalance.

To address the bias introduced by imbalanced datasets, where a classifier might achieve high accuracy by favoring the majority class, this study used the F_1-score instead of accuracy as the primary performance metric. However, during the evaluation, it was observed that the F_1-score did not fully mitigate the impact of imbalanced class distributions.

As observed in Chapter 4, while the two dimensions of the DEAP dataset displayed a similar trend with high correlation coefficients, the valence and arousal dimensions in the DREAMER dataset showed completely different behavior. Even under the *Pruned Dataset Corpus + Inclusion* condition, both dimensions were chosen as part of the representative subset. A small investigation into the results of the DREAMER dataset revealed that the arousal dimension suffers from significant class imbalance, which may explain its weak correlation with the valence dimension.

To ensure equal importance for all classes, regardless of their size, the macro F_1-score was used for evaluation. This metric computes the F_1-score for each class independently and then averages them, ensuring that performance on smaller classes is not overshadowed by dominant classes. By reflecting a model's ability to correctly classify both minority and majority classes, the macro F_1-score provides a balanced evaluation.

Figure 5.1 illustrates the performance of selected models on EEG datasets using macro-averaged F_1-scores. Notably, the valence and arousal dimensions of DREAMER dataset exhibited similar behavior and even correlated highly with those of the DEAP dataset. Furthermore, CHB-MIT, which previously showed high accuracy, appeared less dominant under macro-averaging. The use of macro-averaged F_1-scores resulted in a remarkable change in correlations and clustering among datasets (Figure 5.2). The datasets appeared to share more common patterns, as evidenced by the lower linkage distances, tighter clustering and fewer datasets at distinctly higher distances.

Additionally, comparing datasets with different numbers of classes presents challenges because the number of classes directly impacts task complexity and performance metrics. With more classes, the classification task becomes more difficult as the classifier has a greater chance to make incorrect predictions. For instance, an accuracy of 50% in a two-class dataset reflects only random performance, but the same accuracy in a four-class dataset indicates performance above the random threshold. Aggregated metrics like F_1-scores alone may not fully capture differences in task difficulty across datasets, potentially leading to unfair comparisons.

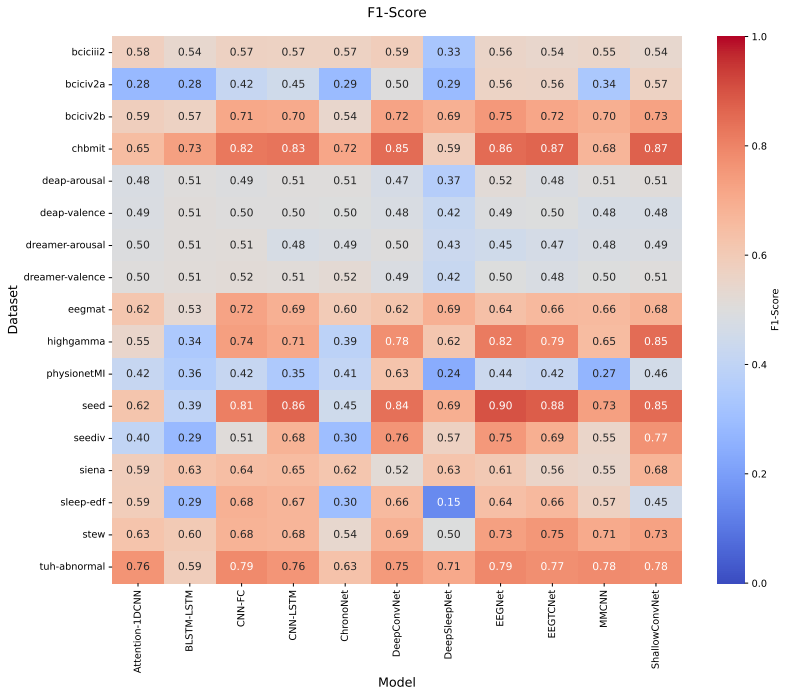

Figure 5.1: Heatmap of Macro average F_1-score Across All Datasets and Models

5.2.4 *Subset Validation*

This study primarily focuses on exploring subset selection procedures to support the development of generalizable deep learning (DL) models. However, as the subsets themselves are still in the early stages of development, no concrete models or frameworks have yet been built to fully utilize these subsets. It is currently difficult to test their validity or evaluate their effectiveness in representing the diversity of EEG datasets. Additionally, due to time constraints, there was no opportunity to design or implement a proper procedure to assess their validity. As a result, while the subsets hold theoretical potential, their practical utility and robustness remain untested, leaving room for future work to explore their impact on the development of generalizable DL models.

Despite these limitations, this study serves as an initial step toward understanding the generalizability of DL models for EEG analysis and highlights the importance of systematic subset selection in improving model evaluation frameworks. Addressing these limitations in future research could enhance the robustness and applicability of the findings.

5.3 FUTURE WORK

This study leaves several areas open for further exploration and improvement.

A critical avenue for future research is the validation of the effectiveness of the selected subsets. One potential approach could involve testing novel, un-

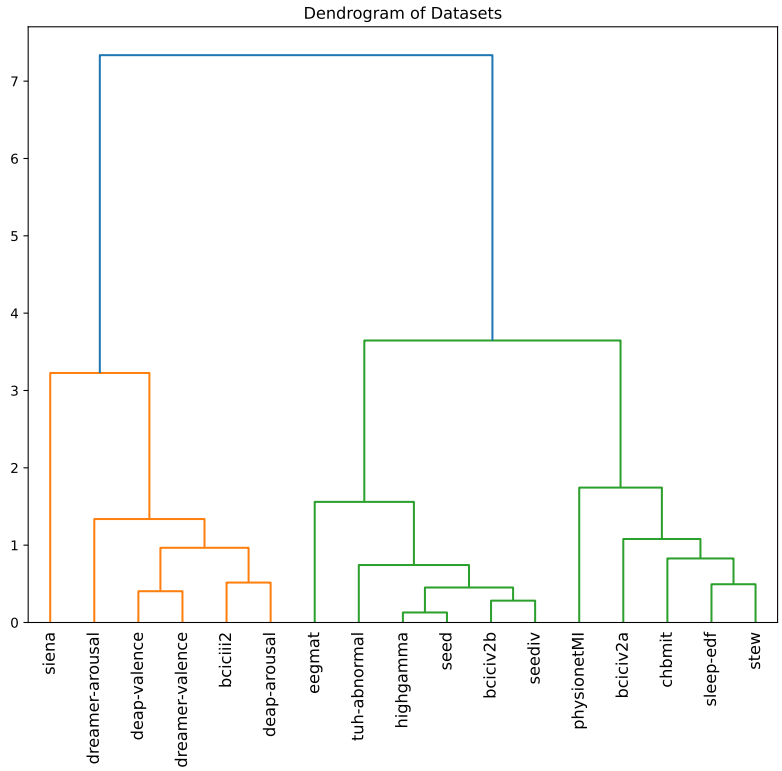

Figure 5.2: Hierarchical Clustering Dendrogram of Datasets Based on Macro average F_1-score

seen DL models on each dataset in the full collection, calculating the median performance, and comparing it against their performance on the selected subset. This would help determine if the results on the subset align with the median performance across all datasets. Alternatively, DL models could be trained on the subsets and evaluated on unseen EEG datasets to assess the subsets' ability to capture essential features and patterns that contribute to cross-domain generalization.

The scope of this study was constrained by limitations in time and resources, which affected the ability to implement a larger number of datasets and DL models, leaving room for expanding the dataset coverage and including a broader variety of models. Future work should aim to encompass a broader range of datasets from additional domains to strengthen the representativeness of the subset selection framework.

Similarly, expanding the variety of architectures beyond end-to-end DNN could also enhance evaluations. While this study focused on widely-used DNN such as CNNs and RNNs, future efforts could explore state-of-the-art architectures like Transformers and attention-based models. Additionally, conventional classification algorithms that require feature extraction could be implemented for comparative analysis. Techniques that extract spatial features by transforming EEG activities into a sequence of topology-preserving multi-spectral images could be explored [15].

To address potential biases toward motor imagery tasks, future studies should also consider applying deep learning models originally designed for datasets spanning multiple domains. Incorporating hybrid architectures or transfer learning techniques could further improve the evaluation of generalization across EEG domains, offering richer information about dataset performance.

Selecting an appropriate metric for evaluating and comparing performance across datasets remains a critical challenge. Future studies should focus on identifying performance metrics that enable reliable subset selection and ensure fair comparisons across datasets with varying characteristics. The alternative metrics should be able to balance class distributions and account for differences in task complexity.

Experimenting with metrics such as normalized accuracy or weighted F_1-scores could help improve the subset selection methodology. A normalized approach that considers the deviation of model performance relative to random guessing (e.g., a baseline accuracy of $\frac{1}{n}$, where n is the number of classes) could enhance interpretability by highlighting whether a model performs better or worse than random chance.

Another important direction is is the development of generalizable DL models specifically designed to work across diverse EEG tasks and domains. This could involve testing novel architectures that prioritize cross-task adaptability, employing pre-training strategies or transfer learning techniques, and comparing the performance of models trained on the selected subsets versus those trained on the full datasets. Such efforts would quantify the subsets' impact on enabling robust and efficient model training.

By addressing these directions, future research can enhance subset selection procedures for EEG datasets and establish a robust validation pipeline for the selected subsets, paving the way for the development of more robust, generalizable, and impactful DL models.

5.4 CONCLUSION

This study attempts to tackle one of the challenges of generalizability in EEG-based deep learning by proposing a systematic approach to identify a representative subset of EEG datasets. By minimizing redundancy while preserving diversity of EEG signals, this work demonstrates the feasibility of constructing a subset that captures the variability across multiple EEG classification tasks, paving the way for more efficient and scalable solutions. The study also provides broader insights into the generalizability and compatibility of state-of-the-art end-to-end deep learning models across multiple EEG datasets spanning different domains.

The study investigated the subset selection procedure for EEG datasets, aiming to identify diverse and representative subsets that minimize prediction error in regression models. Key findings demonstrate that the Pruned Dataset Corpus + Inclusion approach offers the best trade-off between diversity and redundancy, while the Full Dataset Corpus + Exclusion achieves the

lowest prediction error but with higher redundancy. Preprocessing methods like clustering and pruning play a critical role in subset selection, affecting diversity and prediction accuracy.

Despite limitations, such as the absence of standardized pipelines for loading, preprocessing, and training DL models across EEG datasets, as well as the lack of suitable validation methods for the selected subset, this work provides a foundational framework for developing generalizable EEG classification models. Future research should focus on exploring alternative subset selection strategies and expanding the study with more datasets and models to further improve diversity and predictive performance.

Part II

APPENDIX

A

ADDITIONAL INFORMATION OF EEG DATASETS

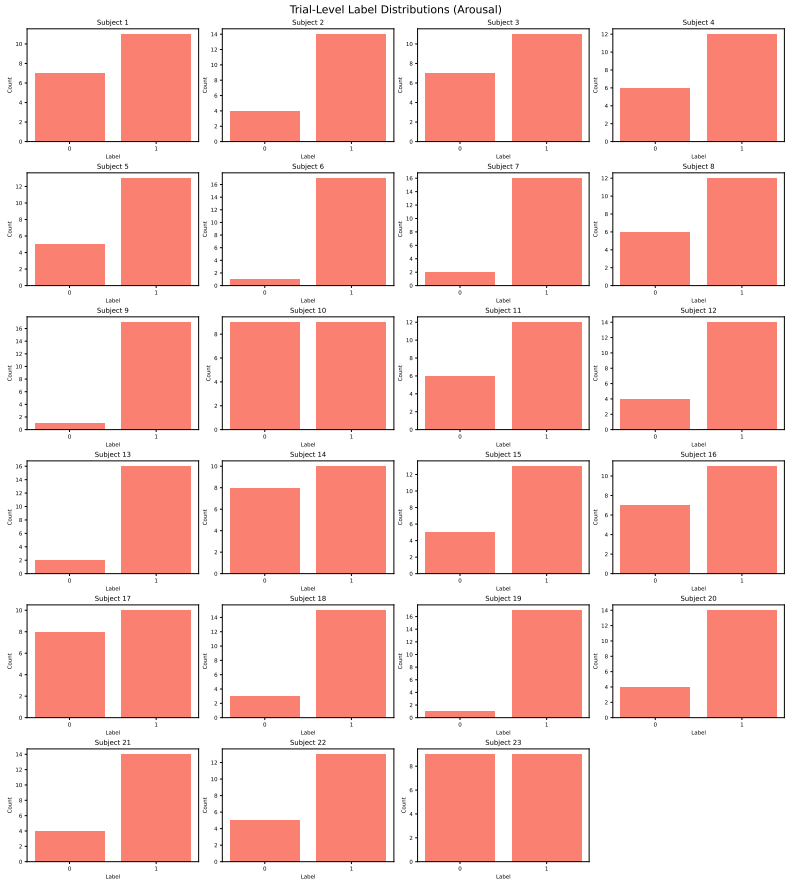

Figure A.1: Trial-Level Label Distributions Across Subjects for the DREAMER (arousal) Dataset

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "{TensorFlow}: a system for {Large-Scale} machine learning." In: *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016, pp. 265–283.

[2] Berdakh Abibullaev and Amin Zollanvari. "A systematic deep learning model selection for P300-based brain–computer interfaces." In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.5 (2021), pp. 2744–2756.

[3] Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. "A comprehensive review of EEG-based brain–computer interface paradigms." In: *Journal of neural engineering* 16.1 (2019), p. 011001.

[4] Charu C Aggarwal et al. *Neural networks and deep learning*. Vol. 10. 978. Springer, 2018.

[5] Matthew Aitchison, Penny Sweetser, and Marcus Hutter. "Atari-5: Distilling the arcade learning environment down to five games." In: *International Conference on Machine Learning*. PMLR. 2023, pp. 421–438.

[6] Ali Al-Saegh, Shefa A Dawwd, and Jassim M Abdul-Jabbar. "Deep learning for motor imagery EEG-based classification: A review." In: *Biomedical Signal Processing and Control* 63 (2021), p. 102172.

[7] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. "Emotion recognition based on EEG using LSTM recurrent neural network." In: *International Journal of Advanced Computer Science and Applications* 8.10 (2017).

[8] Emran Ali, Maia Angelova, and Chandan Karmakar. "Epileptic seizure detection using CHB-MIT dataset: The overlooked perspectives." In: *Royal Society Open Science* 11.6 (2024), p. 230601.

[9] Haifa Almutairi, Ghulam Mubashar Hassan, and Amitava Datta. "Classification of sleep stages from EEG, EOG and EMG signals by SSNet." In: *arXiv preprint arXiv:2307.05373* (2023).

[10] Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review." In: *Neural Computing and Applications* 35.20 (2023), pp. 14681–14722.

[11] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions." In: *Journal of big Data* 8 (2021), pp. 1–74.

[12] Pietro Barbiero, Giovanni Squillero, and Alberto Tonda. "Modeling generalization in machine learning: A methodological and computational study." In: *arXiv preprint arXiv:2006.15680* (2020).

[13] Erol Başar. "A review of alpha activity in integrative brain function: fundamental physiology, sensory coding, cognition and pathology." In: *International Journal of Psychophysiology* 86.1 (2012), pp. 1–24.

[14] Erol Başar. "Brain oscillations in neuropsychiatric disease." In: *Dialogues in clinical neuroscience* 15.3 (2013), pp. 291–300.

[15] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. "Learning representations from EEG with deep recurrent-convolutional neural networks." In: *arXiv preprint arXiv:1511.06448* (2015).

[16] David Bethge, Philipp Hallgarten, Ozan Özdenizci, Ralf Mikut, Albrecht Schmidt, and Tobias Grosse-Puppendahl. "Exploiting multiple EEG data domains with adversarial learning." In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2022, pp. 3154–3158.

[17] Katarzyna Blinowska and Piotr Durka. "Electroencephalography (eeg)." In: *Wiley encyclopedia of biomedical engineering* 10 (2006), p. 9780471740360.

[18] Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. "BCI Competition 2008–Graz data set A." In: *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology* 16 (2008), pp. 1–6.

[19] Hubert Cecotti and Axel Graser. "Convolutional neural networks for P300 detection with application to brain-computer interfaces." In: *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2010), pp. 433–445.

[20] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. "The autoencoding variational autoencoder." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15077–15087.

[21] Debashis Das Chakladar, Shubhashis Dey, Partha Pratim Roy, and Debi Prosad Dogra. "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm." In: *Biomedical Signal Processing and Control* 60 (2020), p. 101989.

[22] Di Chen, Haiyun Huang, Xiaoyu Bao, Jiahui Pan, and Yuanqing Li. "An EEG-based attention recognition method: fusion of time domain, frequency domain, and non-linear dynamics features." In: *Frontiers in Neuroscience* 17 (2023), p. 1194554.

[23] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[24] MX Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[25] Michael X Cohen. "Where does EEG come from and what does it mean?" In: *Trends in neurosciences* 40.4 (2017), pp. 208–218.

[26] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. "Deep learning for electroencephalogram (EEG) classification tasks: a review." In: *Journal of neural engineering* 16.3 (2019), p. 031001.

[27] F Lopes Da Silva. "EEG analysis: theory and practice." In: *Electroencephalography: basic principles, clinical applications and related fields* (1999), pp. 1125–1159.

[28] Jamal I Daoud. "Multicollinearity and regression analysis." In: *Journal of Physics: Conference Series*. Vol. 949. 1. IOP Publishing. 2017, p. 012009.

[29] Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. "EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings." In: *Processes* 8.7 (2020), p. 846.

[30] Silvia López de Diego. *Automated interpretation of abnormal adult electroencephalograms*. Temple University, 2017.

[31] Tom Dietterich. "Overfitting and undercomputing in machine learning." In: *ACM computing surveys (CSUR)* 27.3 (1995), pp. 326–327.

[32] 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." In: *Expert Systems with Applications* 114 (2018), pp. 532–542. URL: https://github.com/hauke-d/cnn-eeg/tree/master.

[33] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. "Differential entropy feature for EEG-based emotion classification." In: *2013 6th international IEEE/EMBS conference on neural engineering (NER)*. IEEE. 2013, pp. 81–84.

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

[35] Biswarup Ganguly, Arpan Chatterjee, Waqar Mehdi, Soumyadip Sharma, and Soumyadeep Garai. "EEG based mental arithmetic task classification using a stacked long short term memory network for brain-computer interfacing." In: *2020 IEEE VLSI DEVICE CIRCUIT AND SYSTEM (VLSI DCS)*. IEEE. 2020, pp. 89–94.

[36] Chenguang Gao, Zhao Li, Hiroki Ora, and Yoshihiro Miyake. "Improving error related potential classification by using generative adversarial networks and deep convolutional neural networks." In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 2468–2476.

[37] Lukas AW Gemein, Robin T Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. "Machine-learning-based diagnostics of EEG pathology." In: *NeuroImage* 220 (2020), p. 117021.

[38] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." In: *circulation* 101.23 (2000), e215–e220.

[39] Shu Gong, Kaibo Xing, Andrzej Cichocki, and Junhua Li. "Deep learning in EEG: Advance of the last ten-year critical period." In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2021), pp. 348–365.

[40] Ian Goodfellow. *Deep learning*. 2016.

[41] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[42] Mohamed M Hammad. "Artificial Neural Network and Deep Learning: Fundamentals and Theory." In: *arXiv preprint arXiv:2408.16002* (2024).

[43] Amir Harati, Silvia Lopez, I Obeid, J Picone, MP Jacobson, and S Tobochnik. "The TUH EEG CORPUS: A big data resource for automated EEG interpretation." In: *2014 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE. 2014, pp. 1–5.

[44] Varsha K Harpale and Vinayak K Bairagi. "Time and frequency domain analysis of EEG signals for seizure detection: A review." In: *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*. IEEE. 2016, pp. 1–6.

[45] Bin He, Abbas Sohrabpour, Emery Brown, and Zhongming Liu. "Electrophysiological source imaging: a noninvasive window to brain dynamics." In: *Annual review of biomedical engineering* 20.1 (2018), pp. 171–196.

[46] Felix A Heilmeyer, Robin T Schirrmeister, Lukas DJ Fiederer, Martin Volker, Joos Behncke, and Tonio Ball. "A large-scale evaluation framework for EEG deep learning architectures." In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2018, pp. 1039–1045.

[47] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[48] Bo Hjorth. "EEG analysis based on time domain properties." In: *Electroencephalography and clinical neurophysiology* 29.3 (1970), pp. 306–310.

[49] S Hochreiter. "Long Short-term Memory." In: *Neural Computation MIT-Press* (1997).

[50] Li Hu and Zhiguo Zhang, eds. *EEG Signal Processing and Feature Extraction*. Springer Singapore, 2019. ISBN: 978-981-13-9113-2. DOI: 10.1007/978-981-13-9113-2. URL: https://doi.org/10.1007/978-981-13-9113-2.

[51] Ramy Hussein, Hamid Palangi, Rabab Ward, and Z Jane Wang. "Epileptic seizure detection: A deep learning approach." In: *arXiv preprint arXiv:1803.09848* (2018).

[52] Sunhee Hwang, Minsong Ki, Kibeom Hong, and Hyeran Byun. "Subject-independent EEG-based emotion recognition using adversarial learning." In: *2020 8th international winter conference on brain-computer interface (BCI)*. IEEE. 2020, pp. 1–4.

[53] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces." In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2020, pp. 2958–2965. URL: https://github.com/iis-eth-zurich/eeg-tcnet/tree/master.

[54] Md Rabiul Islam, Mohammad Ali Moni, Md Milon Islam, Md Rashed-Al-Mahfuz, Md Saiful Islam, Md Kamrul Hasan, Md Sabir Hossain, Mohiuddin Ahmad, Shahadat Uddin, Akm Azad, et al. "Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques." In: *IEEE Access* 9 (2021), pp. 94601–94624.

[55] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Sara Bagherzadeh, Ahmad Shalbaf, David López García, Juan M Gorriz, and U Rajendra Acharya. "Emotion recognition in EEG signals using deep learning methods: A review." In: *Computers in Biology and Medicine* (2023), p. 107450.

[56] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. ISBN: 1461471370.

[57] Ziyu Jia, Youfang Lin, Jing Wang, Kaixin Yang, Tianhang Liu, and Xinwang Zhang. "MMCNN: A multi-branch multi-scale convolutional neural network for motor imagery classification." In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer. 2021, pp. 736–751. URL: https://github.com/jingwang2020/ECML-PKDD_MMCNN.

[58] Sayash Kapoor and Arvind Narayanan. "Leakage and the reproducibility crisis in ML-based science." In: *arXiv preprint arXiv:2207.07048* (2022).

[59] Vishesh K Kapur, Dennis H Auckley, Susmita Chowdhuri, David C Kuhlmann, Reena Mehra, Kannan Ramar, and Christopher G Harrod. "Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline." In: *Journal of clinical sleep medicine* 13.3 (2017), pp. 479–504.

[60] Stamos Katsigiannis and Naeem Ramzan. "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices." In: *IEEE journal of biomedical and health informatics* 22.1 (2017), pp. 98–107.

[61] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG." In: *IEEE Transactions on Biomedical Engineering* 47.9 (2000), pp. 1185–1194.

[62] Aditya Khamparia and Karan Mehtab Singh. "A systematic review on deep learning architectures and applications." In: *Expert Systems* 36.3 (2019), e12400.

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

[64] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "Deap: A database for emotion analysis; using physiological signals." In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[65] Saiprasad Koturwar and Shabbir Merchant. "Weight initialization of deep neural networks (DNNs) using data statistics." In: *arXiv preprint arXiv:1710.10570* (2017).

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25 (2012).

[67] 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." In: *Journal of neural engineering* 15.5 (2018), p. 056013. URL: https://github.com/vlawhern/arl-eegmodels.

[68] Robert Leeb, Clemens Brunner, G Müller-Putz, A Schlögl, and GJGUOT Pfurtscheller. "BCI Competition 2008–Graz data set B." In: *Graz University of Technology, Austria* 16 (2008), pp. 1–6.

[69] Gen Li, Chang Ha Lee, Jason J Jung, Young Chul Youn, and David Camacho. "Deep learning for EEG data analytics: A survey." In: *Concurrency and Computation: Practice and Experience* 32.18 (2020), e5199.

[70] Xiang Li, Dawei Song, Peng Zhang, Guangliang Yu, Yuexian Hou, and Bin Hu. "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network." In: *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2016, pp. 352–359.

[71] Wei Lun Lim, Olga Sourina, and Lipo P Wang. "STEW: Simultaneous task EEG workload data set." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.11 (2018), pp. 2106–2114.

[72] Junxiu Liu, Guopei Wu, Yuling Luo, Senhui Qiu, Su Yang, Wei Li, and Yifei Bi. "EEG-based emotion classification using a deep neural network and sparse autoencoder." In: *Frontiers in Systems Neuroscience* 14 (2020), p. 43.

[73] Mingfei Liu, Wei Wu, Zhenghui Gu, Zhuliang Yu, FeiFei Qi, and Yuanqing Li. "Deep learning based on batch normalization for P300 signal detection." In: *Neurocomputing* 275 (2018), pp. 288–297.

[74] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. "A survey of deep neural network architectures and their applications." In: *Neurocomputing* 234 (2017), pp. 11–26.

[75] Hui Wen Loh, Chui Ping Ooi, Jahmunah Vicnesh, Shu Lih Oh, Oliver Faust, Arkadiusz Gertych, and U Rajendra Acharya. "Automated detection of sleep stages using deep learning techniques: A systematic review of the last decade (2010–2020)." In: *Applied Sciences* 10.24 (2020), p. 8963.

[76] Silvia López, I Obeid, and J Picone. "Automated interpretation of abnormal adult electroencephalograms." In: *MS Thesis, Temple University* (2017).

[77] Debiao Ma, Junteng Zheng, and Lizhi Peng. "Performance evaluation of epileptic seizure prediction using time, frequency, and time–frequency domain measures." In: *Processes* 9.4 (2021), p. 682.

[78] Ramesh Maddula, Joshua Stivers, Mahta Mousavi, Sriram Ravindran, and Virginia de Sa. "Deep Recurrent convolutional neural networks for classifying P300 BCI signals." In: *GBCIC* 201 (2017), pp. 18–22.

[79] Andrew Melnik, W David Hairston, Daniel P Ferris, and Peter König. "EEG correlates of sensorimotor processing: independent components involved in sensory and motor processing." In: *Scientific Reports* 7.1 (2017), p. 4461.

[80] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. "Fundamentals of artificial neural networks and deep learning." In: *Multivariate statistical machine learning methods for genomic prediction*. Springer, 2022, pp. 379–425.

[81] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97.

[82] Meenal V Narkhede, Prashant P Bartakke, and Mukul S Sutaone. "A review on weight initialization strategies for neural networks." In: *Artificial intelligence review* 55.1 (2022), pp. 291–322.

[83] Ismoilov Nusrat and Sung-Bong Jang. "A comparison of regularization techniques in deep neural networks." In: *Symmetry* 10.11 (2018), p. 648.

[84] Natasha Padfield, Jaime Zabalza, Huimin Zhao, Valentin Masero, and Jinchang Ren. "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges." In: *Sensors* 19.6 (2019), p. 1423.

[85] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. "Deep learning for anomaly detection: A review." In: *ACM computing surveys (CSUR)* 54.2 (2021), pp. 1–38.

[86] Fiza Parveen and Arnav Bhavsar. "Attention based 1D-CNN for Mental Workload Classification using EEG." In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. 2023, pp. 739–745.

[87] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks." In: *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2018, pp. 1452–1455.

[88] Md Mustafizur Rahman, Ajay Krishno Sarkar, Md Amzad Hossain, Md Selim Hossain, Md Rabiul Islam, Md Biplob Hossain, Julian MW Quinn, and Mohammad Ali Moni. "Recognition of human emotions using EEG signals: A review." In: *Computers in biology and medicine* 136 (2021), p. 104696.

[89] Alain Rakotomamonjy and Vincent Guigue. "BCI competition III: dataset II-ensemble of SVMs for BCI P300 speller." In: *IEEE transactions on biomedical engineering* 55.3 (2008), pp. 1147–1154.

[90] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[91] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. "Deep learning enabled automatic abnormal EEG identification." In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 2756–2759.

[92] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. "ChronoNet: A deep recurrent neural network for abnormal EEG identification." In: *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer. 2019, pp. 47–56.

[93] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. "Deep learning-based electroencephalography analysis: a systematic review." In: *Journal of neural engineering* 16.5 (2019), p. 051001.

[94] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[95] Mikael Sabuhi, Ming Zhou, Cor-Paul Bezemer, and Petr Musilek. "Applications of generative adversarial networks in anomaly detection: a systematic literature review." In: *Ieee Access* 9 (2021), pp. 161003–161029.

[96] Maham Saeidi, Waldemar Karwowski, Farzad V Farahani, Krzysztof Fiok, Redha Taiar, Peter A Hancock, and Awad Al-Juaid. "Neural decoding of EEG signals with machine learning: A systematic review." In: *Brain Sciences* 11.11 (2021), p. 1525.

[97] Simanto Saha and Mathias Baumert. "Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review." In: *Frontiers in computational neuroscience* 13 (2020), p. 87.

[98] Elham S Salama, Reda A El-Khoribi, Mahmoud E Shoman, and Mohamed A Wahby Shalaby. "EEG-based emotion recognition using 3D convolutional neural networks." In: *International Journal of Advanced Computer Science and Applications* 9.8 (2018).

[99] 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." In: *Human brain mapping* 38.11 (2017), pp. 5391–5420. URL: https://github.com/robintibor/braindecode/blob/master/braindecode/models.

[100] Baha Şen, Musa Peker, Abdullah Çavuşoğlu, and Fatih V Çelebi. "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms." In: *Journal of medical systems* 38 (2014), pp. 1–21.

[101] Mohamad Shahbazi and Hamid Aghajan. "A generalizable model for seizure prediction based on deep learning using CNN-LSTM architecture." In: *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2018, pp. 469–473.

[102] Ramnivas Sharma and Hemant Kumar Meena. "Emerging Trends in EEG Signal Processing: A Systematic Review." In: *SN Computer Science* 5.4 (2024), pp. 1–14.

[103] Vipul Sharma and Mitul Kumar Ahirwal. "Quantification of Mental Workload Using a Cascaded Deep One-dimensional Convolution Neural Network and Bi-directional Long Short-Term Memory Model." In: *Authorea Preprints* (2023).

[104] Ali H. Shoeb. "Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment." Doctoral Dissertation. Massachusetts Institute of Technology, 2009. URL: https://dspace.mit.edu/handle/1721.1/54669.

[105] Ajay Shrestha and Ausif Mahmood. "Review of deep learning algorithms and architectures." In: *IEEE access* 7 (2019), pp. 53040–53065.

[106] Anupreet Kaur Singh and Sridhar Krishnan. "Trends in EEG signal feature extraction applications." In: *Frontiers in Artificial Intelligence* 5 (2023), p. 1072801.

[107] Siuly Siuly, Yan Li, and Yanchun Zhang. *EEG Signal Analysis and Classification: Techniques and Applications*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-47653-7. DOI: 10.1007/978-3-319-47653-7. URL: https://doi.org/10.1007/978-3-319-47653-7.

[108] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. "EEG emotion recognition using dynamical graph convolutional neural networks." In: *IEEE Transactions on Affective Computing* 11.3 (2018), pp. 532–541.

[109] Mahsa Soufineyestani, Dale Dowling, and Arshia Khan. "Electroencephalography (EEG) technology applications and available devices." In: *Applied Sciences* 10.21 (2020), p. 7453.

[110] Derya Soydaner. "A comparison of optimization algorithms for deep learning." In: *International Journal of Pattern Recognition and Artificial Intelligence* 34.13 (2020), p. 2052013.

[111] Martin Spüler and Christian Niethammer. "Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity." In: *Frontiers in human neuroscience* 9 (2015), p. 155.

[112] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[113] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.11 (2017), pp. 1998–2008. URL: https://github.com/kedarps/DeepSleepNet/blob/master/deepSleepNet.py.

[114] Shravani Sur and Vinod Kumar Sinha. "Event-related potential: An overview." In: *Industrial psychiatry journal* 18.1 (2009), pp. 70–73.

[115] Yousef Rezaei Tabar and Ugur Halici. "A novel deep learning approach for classification of EEG motor imagery signals." In: *Journal of neural engineering* 14.1 (2016), p. 016003.

[116] Michal Teplan et al. "Fundamentals of EEG measurement." In: *Measurement science review* 2.2 (2002), pp. 1–11.

[117] Nitish V Thakor and Shanbao Tong. "Advances in quantitative electroencephalogram analysis methods." In: *Annu. Rev. Biomed. Eng.* 6.1 (2004), pp. 453–495.

[118] Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. "Deep learning with convolutional neural networks for decoding and visualization of eeg pathology." In: *arXiv e-prints* (2017), arXiv–1708.

[119] Tomaton124. *21 electrodes of International 10-20 system for EEG*. Accessed: 2024-06-17, Public Domain. 2010. URL: https://commons.wikimedia.org/wiki/File:21_electrodes_of_International_10-20_system_for_EEG.svg.

[120] Nhan Duy Truong, Anh Duy Nguyen, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, and Omid Kavehei. "A generalised seizure prediction with convolutional neural networks for intracranial and scalp electroencephalogram data analysis." In: *arXiv preprint arXiv:1707.01976* (2017).

[121] Kostas M Tsiouris, Vasileios C Pezoulas, Michalis Zervakis, Spiros Konitsiotis, Dimitrios D Koutsouris, and Dimitrios I Fotiadis. "A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals." In: *Computers in biology and medicine* 99 (2018), pp. 24–37.

[122] Swati Vaid, Preeti Singh, and Chamandeep Kaur. "EEG signal analysis for BCI interface: A review." In: *2015 fifth international conference on advanced computing & communication technologies*. IEEE. 2015, pp. 143–147.

[123] Kaido Värbu, Naveed Muhammad, and Yar Muhammad. "Past, present, and future of EEG-based BCI applications." In: *Sensors* 22.9 (2022), p. 3331.

[124] Zitong Wan, Rui Yang, Mengjie Huang, Nianyin Zeng, and Xiaohui Liu. "A review on transfer learning in EEG signal analysis." In: *Neurocomputing* 421 (2021), pp. 1–14.

[125] Xiashuang Wang, Yinglei Wang, Dunwei Liu, Ying Wang, and Zhengjun Wang. "Automated recognition of epilepsy from EEG signals using a combining space–time algorithm of CNN-LSTM." In: *Scientific Reports* 13.1 (2023), p. 14876.

[126] Yi Wang, Zhiyi Huang, Brendan McCane, and Phoebe Neo. "EmotioNet: A 3-D convolutional neural network for EEG-based emotion recognition." In: *2018 international joint conference on neural networks (IJCNN)*. IEEE. 2018, pp. 1–7.

[127] Peter Welch. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." In: *IEEE Transactions on audio and electroacoustics* 15.2 (1967), pp. 70–73.

[128] Wojciech Wojcikiewicz, Carmen Vidaurre, and Motoaki Kawanabe. "Stationary common spatial patterns: towards robust classification of non-stationary eeg signals." In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, pp. 577–580.

[129] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. "Brain–computer interfaces for communication and control." In: *Clinical neurophysiology* 113.6 (2002), pp. 767–791.

[130] Jonathan R Wolpaw, Dennis J McFarland, Gregory W Neat, and Catherine A Forneris. "An EEG-based brain-computer interface for cursor control." In: *Electroencephalography and clinical neurophysiology* 78.3 (1991), pp. 252–259.

[131] Xun Wu, Wei-Long Zheng, and Bao-Liang Lu. "Identifying functional brain connectivity patterns for EEG-based emotion recognition." In: *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2019, pp. 235–238.

[132] Baoguo Xu, Linlin Zhang, Aiguo Song, Changcheng Wu, Wenlong Li, Dalin Zhang, Guozheng Xu, Huijun Li, and Hong Zeng. "Wavelet transform time-frequency image and convolutional network-based motor imagery EEG classification." In: *Ieee Access* 7 (2018), pp. 6084–6093.

[133] Gaowei Xu, Tianhe Ren, Yu Chen, and Wenliang Che. "A one-dimensional CNN-LSTM model for epileptic seizure recognition using EEG signal analysis." In: *Frontiers in neuroscience* 14 (2020), p. 578126.

[134] Yang Zhan, Rosa C Paolicelli, Francesco Sforazzini, Laetitia Weinhard, Giulia Bolasco, Francesca Pagani, Alexei L Vyssotski, Angelo Bifone, Alessandro Gozzi, Davide Ragozzino, et al. "Deficient neuron-microglia signaling results in impaired functional brain connectivity and social behavior." In: *Nature neuroscience* 17.3 (2014), pp. 400–406.

[135] Kai Zhang, Guanghua Xu, Zezhen Han, Kaiquan Ma, Xiaowei Zheng, Longting Chen, Nan Duan, and Sicong Zhang. "Data augmentation for motor imagery signal classification based on a hybrid neural network." In: *Sensors* 20.16 (2020), p. 4485.

[136] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, David Mcalpine, and Yu Zhang. "A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers." In: *Journal of neural engineering* 18.3 (2021), p. 031002.

[137] Wei-Long Zheng and Bao-Liang Lu. "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks." In: *IEEE Transactions on autonomous mental development* 7.3 (2015), pp. 162–175.

[138] Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. "Electroencephalograms during mental arithmetic task performance." In: *Data* 4.1 (2019), p. 14.