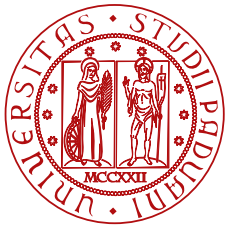

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

UNIVERSITY OF PADUA
DEPARTMENT OF INFORMATION ENGINEERING

Master Degree in ICT for Internet and Multimedia

A Preliminary Study on Open-Source EEG Datasets: Human and GPT-Based Review and Cross-Dataset Classification

Master Candidate

Betül Sena Anar

Student ID 2080328

Supervisor

Prof. Leonardo Badia

University of Padova

Co-supervisor

Prof. Giulia Cisotto

University of Trieste

Academic Year
2025/2026
Graduation date 15 April 2026

To my family, and to the presence within me that has already changed the meaning of this journey.

Abstract

Electroencephalography (EEG), a non-invasive signal recording method capable of measuring the electrical activity of the brain with high temporal resolution, is widely used in brain-computer interfaces and cognitive analysis studies. In recent years, with the development of machine learning and deep learning-based approaches, classification studies based on EEG signals have gained importance.However, there are significant differences in the reporting of methodological characteristics of EEG datasets in the literature, which can limit the comparability and reproducibility of experimental studies. This study proposes a multi-stage evaluation framework that combines systematic literature review, AI-assisted data extraction, and machine learning-based classification methods to allow for more systematic analysis of open-access EEG datasets.

In the first stage of the study, the methodological characteristics of the EEG datasets were systematically examined by a human panel of four evaluators representing different levels of expertise. Key parameters such as sampling frequency, number of channels, number of participants, data format, and experimental protocol were analyzed during the evaluation process. The findings revealed significant heterogeneity in the reporting of EEG datasets, particularly in terms of methodological details such as electrode configuration, reference electrode information, and session structure.

In the second phase of the study, the evaluation framework developed by the human panel was transferred to a GPT-based model, and the model's performance in automatically extracting methodological information from scientific articles was examined. The results showed that the GPT-based approach could provide high precision and consistency, especially when identifying the technical parameters explicitly reported in the text. However, limitations of the model were observed in identifying information requiring contextual interpretation or presented only in figures.

In the final phase of the study, the performance of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) based classification models was comparatively evaluated on different open-access motor imagery EEG

datasets. The experimental results showed that both models could perform the motor imagery EEG classification task. However, CNN-based architectures were observed to generally provide higher classification performance. Furthermore, the findings revealed that the classification performance depends not only on the model architecture but also on factors such as the participant population of the datasets, the experimental protocol, and the recording conditions.

In conclusion, this study presents a holistic research approach that combines systematic analysis of EEG datasets, data extraction of the AI-assisted literature, and machine learning-based classification methods. The proposed framework contributes to a more transparent and systematic evaluation of the methodological characteristics of EEG datasets and allows a more reliable comparison of classification studies performed on different datasets.

Sommario

Lelettroencefalografia (EEG), un metodo non invasivo di registrazione dei segnali capace di misurare lattività elettrica del cervello con elevata risoluzione temporale, è ampiamente utilizzata nelle interfacce cervello-computer e negli studi di analisi cognitiva. Negli ultimi anni, con lo sviluppo di approcci basati sul machine learning e sul deep learning, gli studi di classificazione basati su segnali EEG hanno acquisito crescente rilevanza. Tuttavia, nella letteratura esistono differenze significative nella modalità di riportare le caratteristiche metodologiche dei dataset EEG, il che può limitare la comparabilità e la riproducibilità degli studi sperimentali. Questo lavoro propone un framework di valutazione multi-fase che combina revisione sistematica della letteratura, estrazione dei dati assistita da intelligenza artificiale e metodi di classificazione basati su machine learning, al fine di consentire unanalisi più sistematica dei dataset EEG ad accesso aperto.

Nella prima fase dello studio, le caratteristiche metodologiche dei dataset EEG sono state esaminate sistematicamente da un panel umano composto da quattro valutatori con diversi livelli di competenza. Durante il processo di valutazione sono stati analizzati parametri chiave quali la frequenza di campionamento, il numero di canali, il numero di partecipanti, il formato dei dati e il protocollo sperimentale. I risultati hanno evidenziato una significativa eterogeneità nella descrizione dei dataset EEG, in particolare per quanto riguarda dettagli metodologici come la configurazione degli elettrodi, le informazioni sullelettrodo di riferimento e la struttura delle sessioni.

Nella seconda fase dello studio, il framework di valutazione sviluppato dal panel umano è stato trasferito a un modello basato su GPT, ed è stata analizzata la capacità del modello di estrarre automaticamente informazioni metodologiche da articoli scientifici. I risultati hanno mostrato che lapproccio basato su GPT è in grado di fornire elevata precisione e coerenza, in particolare nellidentificazione

dei parametri tecnici esplicitamente riportati nel testo. Tuttavia, sono emerse alcune limitazioni del modello nellindividuare informazioni che richiedono interpretazione contestuale o che sono presentate esclusivamente sotto forma di figure.

Nella fase finale dello studio, le prestazioni di modelli di classificazione basati su Support Vector Machine (SVM) e Convolutional Neural Network (CNN) sono state valutate comparativamente su diversi dataset EEG di motor imagery ad accesso aperto. I risultati sperimentali hanno mostrato che entrambi i modelli sono in grado di eseguire il compito di classificazione dei segnali EEG di motor imagery. Tuttavia, le architetture basate su CNN hanno generalmente mostrato prestazioni superiori. Inoltre, i risultati indicano che le prestazioni di classificazione dipendono non solo dallarchitettura del modello, ma anche da fattori quali la popolazione dei partecipanti nei dataset, il protocollo sperimentale e le condizioni di registrazione.

In conclusione, questo studio presenta un approccio di ricerca olistico che combina lanalisi sistematica dei dataset EEG, lestrazione dei dati assistita da intelligenza artificiale e metodi di classificazione basati su machine learning. Il framework proposto contribuisce a una valutazione più trasparente e sistematica delle caratteristiche metodologiche dei dataset EEG e consente un confronto più affidabile tra studi di classificazione condotti su diversi dataset.

Contents

Sommario **v**

List of Figures **xi**

List of Tables **xiii**

List of Acronyms **xv**

Introduction **1**

1 State of the Art **5**

2 Background **17**
2.1 Basic Characteristics of EEG Signals 17
2.2 Principles of Systematic Literature Review 22
2.3 GPT-Assisted Literature Review 24
2.4 Machine Learning and Deep Learning Models 26
2.5 Related Work 28

3 Materials and Methods **31**
3.1 Scientific Systematic Review 34
3.2 Dataset Overview and Access 37
3.3 Preprocessing and Feature Extraction 41
3.4 GPT Dataset Curator 45
3.5 Classification Models 48

4 Results and Discussion **55**
4.1 Systematic Review Results 56
4.2 GPT Performance 59
4.3 Cross-dataset EEG Statistical Analysis 63

4.4 Intra and Cross-dataset Classification 68
4.4.1 Results on Dataset 1 68
4.4.2 Results on Dataset 2 75
4.4.3 Results on Dataset 3 82
4.5 Discussion 94

Conclusions and Future Works **99**

References **103**

Declaration of Generative AI **115**

List of Figures

1 Number of publications on EEG-based classification and brain-computer interfaces between 2017 and 2025. The steady increase reflects the growing research interest in the field. Data retrieved from Dimensions.ai. 2

1.1 Standard electrode placement according to the international 10–20 EEG system [15] 6

2.1 Example of multichannel electroencephalography (EEG) signals recorded over time [59] 18
2.2 A representative EEG data acquisition setup for a motor imagery brain–computer interface study, illustrating the EEG cap, amplifier, data receiver, and host computer, along with the experimental trial structure for left- and right-hand motor imagery tasks. Modified from [64] under CC BY 4.0 license. 20
2.3 Workflow of the human-based systematic review process. The lead investigator defined the evaluation criteria and communicated them to four reviewers through a knowledge transfer process. Each reviewer independently examined dataset documentation and recorded observations using a structured form. Responses were consolidated in a shared spreadsheet for subsequent analysis. 24

3.1 Flowchart of the proposed multi stage evaluation and analysis framework 33
3.2 Structure of the Multi-Level Evaluation Team 35
3.3 Experimental Setup of Dataset 1 [101] 40
3.4 General pipeline of the CNN architecture employed in this study. Adapted from [29]. 50

4.1 Example of structured dataset annotation sheet used in the manual review process 57
4.2 Amplitude Distribution of Dataset 1 64
4.3 Amplitude Distribution of Dataset 2 65
4.4 Amplitude Distribution of Dataset 3 66
4.5 Amplitude Distribution Graph of All Dataset 67
4.6 Confusion Matrix results for Dataset 1 for SVM 69
4.7 Precision Recall Curve for Dataset 1 for SVM 70
4.8 ROC for Dataset 1 for SVM 71
4.9 Confusion Matrix results for Dataset 1 for CNN 72
4.10 Training and validation accuracy curves for Dataset 1 74
4.11 Training and validation loss curves for Dataset 1 74
4.12 Confusion Matrix results for Dataset 2 for SVM 76
4.13 Precision Recall Curve for Dataset 2 for SVM 77
4.14 ROC for Dataset 2 for SVM 77
4.15 Confusion Matrix results for Dataset 2 for CNN 79
4.16 Training and validation accuracy curves for Dataset 2 80
4.17 Training and validation loss curves for Dataset 2 80
4.18 Confusion Matrix results for Dataset 3 for SVM 83
4.19 Precision Recall Curve for Dataset 3 for SVM 84
4.20 ROC for Dataset 3 for SVM 84
4.21 Confusion Matrix results for Dataset 3 for CNN 86
4.22 Training and Validation Accuracy of Dataset 3 87
4.23 Train vs Validation Loss Graph of Dataset 3 87
4.24 Confuion Matrix for Combined Dataset for SVM 90
4.25 Precision Recall Curve for Combined Dataset for SVM 91
4.26 ROC for Combined Dataset for SVM 91
4.27 Confuion Matrix for Combined Dataset for CNN 92
4.28 Training and Validation Accuracy of Combined Dataset 93
4.29 Training and Validation Loss of Combined Dataset 93

List of Tables

2.1 Summary of key characteristics of commonly used open-source motor imagery EEG datasets. MI = Motor Imagery; L/R = Left-/Right hand. 21

3.1 Datasets 38
3.2 Comparison of the main characteristics of the EEG motor imagery datasets used in this study 41
3.3 Layer-wise architectural details of the CNN model. T denotes the number of time points and C the number of EEG channels in the input. 51
3.4 Hyperparameter configuration for CNN training across all datasets. 52

4.1 Error Analysis of GPT on EEG Dataset Articles 60
4.2 Classification Report of the SVM Model for Dataset 1 70
4.3 Classification Report of the CNN Model for Dataset 1 73
4.4 Classification Report of the SVM Model for Dataset 2 76
4.5 Classification Report of the CNN Model for Dataset 2 79
4.6 Classification Report of the SVM Model for Dataset 3 83
4.7 Classification Report of the CNN Model for Dataset 3 86

List of Acronyms

EEG Electroencephalography

GPT Generative Pre-trained Transformer

CNN Convolutional Neural Network

BCI Brain-Computer Interface

SVM Support Vector Machine

k-NN k-Nearest Neighbors

LDA Linear Discriminant Analysis

RDN Residual Dense Network

ERD Event-Related Desynchronization

ERS Event-Related Synchronization

BIDS Brain Imaging Data Structure

LLM Large Language Model

RBF Radial Basis Function

MI Motor Imagery

TMS Transcranial Magnetic Stimulation

DBS Deep Brain Stimulation

LPF Low-Pass Filter

ECoG Electrocorticography

Hz Hertz

EOG Electrooculography

Cz Central Electrode Position (Cz)

Fpz Frontal Pole Electrode Position (Fpz)

CPz Central-Parietal Electrode Position (CPz)

Pz Parietal Electrode Position (Pz)

PSD Power Spectral Density

CAR Common Average Referencing

CSP Common Spatial Pattern

NLP Natural Language Processing

ELU Exponential Linear Unit

AUC Area Under the Curve

ROC Receiver Operating Characteristic

AP Average Precision

ECP Electrocorticography Potential

RNN Recurrent Neural Networks

Introduction

Electroencephalography (EEG) is a non-invasive signal recording method that allows for high-resolution temporal measurement of the brain's electrical activity. EEG signals have been used for many years to study cognitive processes, motor activities, and various neurophysiological conditions, and are considered a fundamental data source, particularly in brain-computer interfaces (BCIs) and EEG-based cognitive analysis applications [1, 2]. Its portability, relatively low cost, and repeatable measurement capabilities make EEG widely preferred in both clinical and experimental research [1].

Classification problems based on EEG signals inherently involve various methodological challenges. EEG signals exhibit low amplitude, potential noise, and high variability among individuals [3]. Furthermore, artifacts during measurement, environmental noise, and physiological effects can directly impact signal quality [4]. This is one of the key factors limiting the accuracy and reliability of developed classification models [3]. Therefore, EEG-based classification studies present a multi-dimensional problem area that needs to be addressed not only in terms of the algorithms used but also in terms of the structure, quality, and evaluation processes of the datasets.

Numerous machine learning and deep learning-based approaches for EEG classification have been proposed in the literature. A significant portion of these studies report high performance values under specific datasets or controlled experimental conditions. However, whether these models demonstrate similar success on different datasets is often not adequately evaluated. Significant differences in model performance can be observed, particularly when dealing with EEG datasets that have different experimental protocols, task types, channel numbers, and sampling frequencies [5]. This raises generalizability and reproducibility problems in EEG classification studies and limits the comparability of results reported in the literature [6].

The increase in the number of open-source Electroencephalography (EEG) datasets in recent years presents a significant opportunity to address these problems more systematically.The increase in the number of open-source EEG datasets in recent years presents a significant opportunity to address these problems more systematically (see Figure 1). Open-access datasets allow methods. Open-access datasets allow methods developed by different research groups to be tested on the same data, increasing the transparency of scientific studies. However, open-source EEG datasets often have a heterogeneous structure. Different data collection protocols, task descriptions, sampling frequencies, and data quality characteristics make it difficult to directly use or compare these datasets together. This heterogeneous structure further highlights the need for a standardized evaluation and classification approach.

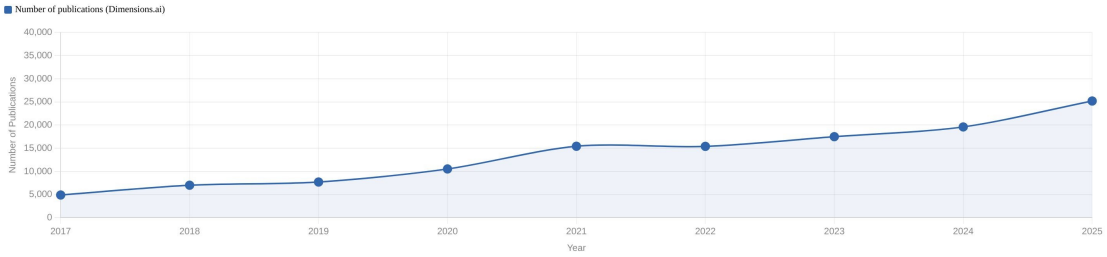

Figure 1: Number of publications on EEG-based classification and brain-computer interfaces between 2017 and 2025. The steady increase reflects the growing research interest in the field. Data retrieved from Dimensions.ai.

Among the various EEG-based paradigms, motor imagery (MI) has received particular attention in brain-computer interface research. MI refers to the mental simulation of a movement without actual physical execution, and the associated EEG patterns have been widely studied for their potential in rehabilitation and assistive technologies. Given this relevance, the present thesis focuses on MI-based EEG classification as its primary application domain.In addition to these sources of variability, the spatial configuration of EEG electrodes plays a critical role in the analysis of motor imagery signals.

Labeling processes used in the evaluation of EEG datasets are a critical element that directly affects classification performance. Traditionally, the evaluation of EEG data relies on manual labeling and classification processes performed by field experts. While manual evaluation is considered reliable due to its reliance on expert knowledge, it has significant limitations in terms of scalability due to its time-consuming nature and susceptibility to inconsisten-

cies among evaluators. Especially **studies requiring the analysis of multiple datasets or long-term EEG recordings** manual evaluation processes can create problems in terms of both practicality and consistency. Furthermore, **differences differences in the knowledge levels and experiences of evaluators can lead to subjectivity in classification results.** This situation necessitates a systematic analysis of human-based evaluation processes and their comparison with alternative approaches.

In this context, automated and semi-automated assessment approaches stand out as alternative or complementary methods to manual processes. Large language models developed in recent years offer new possibilities in data evaluation processes thanks to their ability to process complex contextual information and generate consistent classification decisions. Generative Pre-trained Transformer (GPT)-based models are particularly noteworthy for their capacity to analyze data structures containing textual descriptions and contextual information. The use of such models in the analysis of EEG-related literature allows for a systematic examination of the differences and similarities between human-based assessments and automated approaches [7, 8].Therefore, GPT-supported classification is considered in this study not as an independent predictive model, but as a methodological analysis tool compared with human-based assessments.

Modeling approaches used in EEG classification studies are also of great importance in terms of the interpretability and generalizability of the results. Classical machine learning methods have long been used in EEG analysis due to their lower computational costs and relatively high interpretability advantages. Methods such as Support Vector Machines (SVMs) define optimal decision boundaries between classes in high-dimensional feature spaces and have been widely applied in EEG classification tasks due to their strong generalization performance under limited data conditions [9]. On the other hand, deep learning-based approaches, especially convolutional neural networks (CNNs), have attracted great interest in recent years due to their ability to automatically extract features from raw EEG signals or transformed representations [10, 11]. However, the success of these models largely depends on the structure, integrity, and consistency of the datasets used. Systematically evaluating the performance of deep learning models on heterogeneous EEG datasets is therefore emerging as an important research topic.

This thesis pursues two interrelated objectives. First, it investigates the degree of agreement between manual EEG assessment performed by raters with

varying levels of expertise and GPT-assisted classification. Second, it examines the performance of classical machine learning (SVM) and deep learning (CNN) models trained on open-source EEG datasets, with the aim of evaluating how heterogeneous data sources affect automated classification outcomes. The resulting classification results are analyzed using both classical machine learning methods and deep learning-based models. Specifically, the effects of heterogeneous EEG datasets on classification performance are systematically evaluated through SVM and Convolutional Neural Network (CNN) models. This multi-stage approach allows for a detailed examination of the differences and similarities between human-based, AI-assisted, and model-based classification processes.

The main objective of this thesis is to reveal the strengths and weaknesses of different evaluation and modeling approaches used in the classification of heterogeneous open-source EEG datasets and to present a highly reproducible methodological framework. In this regard, the original contributions of this thesis can be summarized as: the development of a systematic evaluation protocol for open-source EEG datasets; a comparison of human-based manual classification and GPT-assisted classification approaches; an analysis of the effects of data integrity and dataset heterogeneity on performance through SSM and CNN-based model. These contributions aim to go beyond approaches that focus solely on model performance in EEG-based classification studies, revealing the decisive role of data structure and evaluation processes on the results.

This thesis consists of four main sections. *Introdcution* presents the motivation, purpose, and scope of the study. Following the Introduction, *State of the Art* section, which discusses the relevant literature and existing studies, examines EEG-based classification in detail. *Background* section explains the basic characteristics of EEG signals, the datasets used, and the methods that constitute the technical infrastructure of the study. *Materials and Methods* section presents the manual evaluation, GPT-supported classification, and modeling processes in detail and describes the applied methods. Finally, *Results and Discussion* section summarizes the findings, provides a general evaluation, and offers suggestions for future studies.This thesis aims to provide methodological integrity in EEG-based classification studies by combining manual evaluation, large language models, and machine learning approaches. The results obtained are expected to contribute to future EEG-based machine learning and brain-computer interface studies in terms of comparability, generalizability, and reproducibility.

1

State of the Art

The main purpose of this State of the Art chapter is to comprehensively present the work done to date in EEG-based classification studies, analyze the methods used, highlight strengths and weaknesses, and identify existing research gaps. EEG signal analysis, as a low-cost and high-temporal-resolution neurophysiological measurement tool, has become a central focus of neuroscience, machine learning, and human-computer interaction research in recent years. The EEG classification problems at the heart of this study are applied in a range of different fields, including brain-computer interfaces Brain-Computer Interface (BCI), cognitive load assessment, emotion recognition, epileptic seizure detection, and other clinical applications [12, 13, 14].EEG recordings are commonly obtained using standardized electrode placement systems, such as the international 10–20 system, which ensures consistent coverage of cortical regions across subjects and studies.

As illustrated in Figure 1.1 [15], electrodes are positioned based on anatomical landmarks, providing a systematic representation of brain activity. However, variations in electrode density, placement schemes (e.g., 10–20 vs. 10–10 systems), and channel selection across datasets may influence the quality and interpretability of the recorded signals. These differences further contribute to the heterogeneity of EEG datasets and may affect the comparability of results across studies. In this figure electrode positions are defined based on anatomical landmarks such as the nasion, inion, and preauricular points, ensuring consistent spatial sampling of brain activity across subjects.

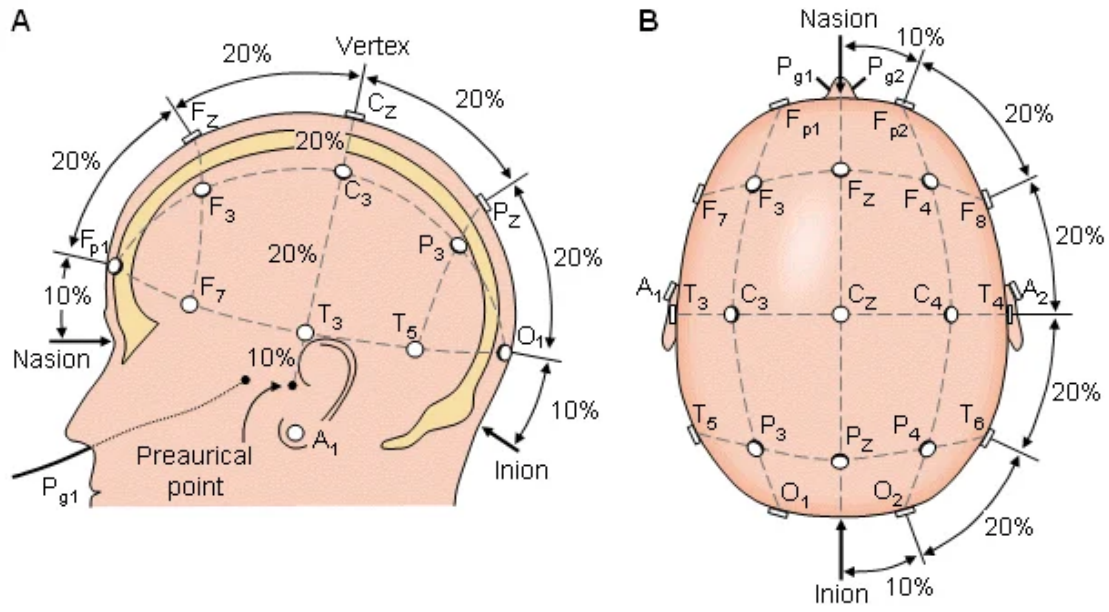

Figure 1.1: Standard electrode placement according to the international 10–20 EEG system [15]

Traditional machine learning techniques have been the fundamental approach in EEG classification studies for many years. These approaches generally rely on features extracted manually from signals; features such as bandwidth, frequency spectrum, entropy, and statistical properties are combined with classifiers such as Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Linear Discriminant Analysis (LDA) [16, 17].

While these methods have been successfully applied to certain task types (e.g., motor imagery or resting state), they have disadvantages such as their reliance on feature extraction, limited generalization capabilities, and limitations in handling complex sample structures.

On the other hand, deep learning techniques have shown significant growth in the field of EEG classification over the last decade. Deep learning approaches can automatically extract features from raw or minimally processed signals, thus providing higher performance compared to traditional methods. In particular, CNN, Recurrent Neural Networks (RNN), and various hybrid architectures have gained widespread acceptance in the field of EEG classification [13].

However, these models require large and high-quality datasets to function successfully, and performance degradation is observed in small or heterogeneous datasets.

The current literature addresses both the advantages and limitations of deep

learning methods; these include, in particular, issues such as the weak explainability of the model, high computational requirements for training, and generalizability across different datasets [12, 13] .Furthermore, recent studies emphasize that the performance of deep learning models is highly sensitive to data preprocessing, enhancement, and architectural choices [18, 19] .Studies reviewed in the EEG classification literature are not limited to motor imagery tasks but encompass a wide range of applications, including emotional state recognition, cognitive load assessment, and clinical event detection [20, 21] .Furthermore, efforts to improve classification success by combining multimodal EEG data with multiple signal types are increasing [22].

The literature on EEG classification varies according to different applications and task types, and this diversity is considered an important factor in both methodological and applied analyses.In the literature on EEG-based classification, studies based on motor imagery tasks hold a central position due to both their methodological maturity and the breadth of their application areas. Motor imagery refers to the brain activity that occurs when an individual mentally imagines a specific movement without actually performing it. The EEG signals generated during this process exhibit distinctive patterns, particularly in regions associated with the motor cortex, and these patterns offer a suitable framework for classification problems. The relatively clear definition of motor imagery tasks and the ease with which experimental protocols can be standardized compared to other cognitive tasks are among the main reasons for the extensive research in this field in the literature [23, 24].

In the early stages of motor imagery-based EEG classification studies, handcrafted features extracted from signals were adopted as the basic approach. Statistical features such as mean amplitude, variance, and signal strength in the time domain, and measures such as band strength and power spectral density in the frequency domain, were frequently used. Event-related desynchronization and synchronization ERD/ERS patterns observed particularly in the mu (8-12 Hz) and beta (13-30 Hz) bands have been strongly associated with motor imagery tasks [25, 26]. These features have been evaluated in conjunction with classical machine learning algorithms such as support vector machines, linear discriminant analysis, and logistic regression.

Although these approaches have reported high classification accuracies in specific motor imagery tasks, the generalizability of these results has frequently been a subject of debate in the literature. Many studies have been conducted on

datasets with a limited number of participants and homogeneous experimental protocols. There are also studies showing that the performance of the same methods decreases significantly when applied to different datasets or different participant profiles [27] . This highlights that dataset specificity is a significant problem, even in the motor imagery literature.

With the increasing use of deep learning-based approaches in motor imagery EEG classification studies in recent years, methodological discussions in this field have gained a new dimension. CNN perform automated feature extraction by being trained on raw EEG signals or time-frequency representations. Schirrmeister et al. showed that deep CNN architectures can compete with, and in some cases even outperform, traditional hand-crafted feature-based approaches on motor imagery EEG data [28]. Similarly, the EEGNet architecture proposed by Lawhern et al. produced effective results in various motor imagery tasks with a low number of parameters [29].

However, the literature emphasizes that the high performance values reported in deep learning-based motor imagery studies are largely dependent on the datasets used. CNN-based models can achieve high accuracy when trained on a specific dataset; however, they can experience significant performance degradation when transferred to different datasets [30, 31] . This demonstrates that the generalizability of deep learning models is limited, even for motor imagery tasks, and highlights the importance of comparative analyses performed on heterogeneous datasets.

In recent years, in addition to motor imagery studies, the literature on cognitive task-based EEG classification has also expanded significantly. Cognitive states such as attention, working memory, mental load, decision-making, and perceptual processes are represented in EEG signals with complex and multidimensional patterns. Therefore, cognitive EEG classification studies are considered a more methodologically challenging problem area [32, 33] .

Time-frequency analyses, wavelet transforms, and functional connectivity measures are frequently used in cognitive task-based EEG studies. In particular, power variations in the theta and alpha bands have been associated with mental and attention load processes [34] . These features have been analyzed using classical machine learning algorithms as well as models based on deep learning. However, due to the nature of cognitive tasks, labeling them is a more subjective process compared to motor imagery tasks.

The literature shows that the experimental protocols and evaluation criteria

used in labeling cognitive tasks vary significantly from study to study. This makes it difficult to directly compare the classification results [35]. Furthermore, a large proportion of cognitive EEG datasets have been collected in controlled laboratory settings with a limited number of participants. This leads to limited generalizability in real-world conditions.

Some recent studies propose multitasking classification approaches that consider motor imagery and cognitive tasks together. These studies aim to analyze the common and distinctive aspects of EEG patterns belonging to different task types [36, 37]. However, the number of such studies is limited and they are generally restricted to specific datasets. Systematic studies that evaluate heterogeneous task types and different datasets together are still rare in the literature.

Overall, while the literature on motor imagery and cognitive task-based EEG classification is quite rich in terms of methodological diversity, it struggles to provide a framework due to dataset heterogeneity, subjectivity in labeling processes, and generalizability problems. Most studies focus on a specific task type or a single dataset, addressing methodological commonalities and limitations between different task types in a limited way. This situation indicates that comparative and systematic analyses performed on heterogeneous EEG datasets could fill a significant gap in the literature.

The use of open-source datasets in EEG-based classification studies has become an increasingly important topic in the literature in recent years. Open-access datasets increase the transparency of scientific studies, allow methods developed by different research groups to be tested on the same data, and strengthen the comparability of results. In this respect, open-source EEG datasets play a central role in implementing the principle of reproducibility [38, 39] .

Many EEG classification studies in the literature have been performed on closed or privately collected datasets. While such studies are valuable in demonstrating the success of the proposed method on a specific dataset, it is often not possible for independent researchers to verify the results and compare them with different methods. This raises significant questions about the generalizability and reliability of results reported in EEG-based machine learning studies [40] .

The proliferation of open-source EEG datasets is considered a significant step towards overcoming these limitations. Numerous open-access EEG datasets encompassing different task types, such as motor imagery, cognitive tasks, emo-

tional state recognition, and resting state, have been added to the literature [41, 42, 43] . These datasets allow for the evaluation of different classification approaches under the same conditions, making inter-method comparisons more meaningful.

However, these advantages offered by open-source EEG datasets also bring significant methodological challenges. Open-source datasets are often collected by different research groups using different experimental protocols, hardware systems, and preprocessing approaches. Basic parameters such as channel number, electrode placement, sampling frequency, and recording duration can vary greatly between datasets [44] .This leads to a heterogeneous structure of open-source EEG datasets.

Dataset heterogeneity is one of the most common problems in the EEG-based classification literature. Heterogeneous datasets make it difficult for the same classification model to exhibit consistent performance across different datasets. Numerous studies in the literature show that models trained on a particular dataset experience significant performance degradation on a different one [27, 30]. This indicates that model success is closely related not only to algorithmic choices but also to the structural characteristics of the dataset.

The heterogeneity of open-source EEG datasets is not limited to technical parameters. Data quality also varies significantly between datasets. EEG recordings are highly susceptible to artifacts and can be affected by many factors such as blinking, muscle movements, environmental noise, and electrode contact problems. The extent to which artifact cleaning and preprocessing steps are applied in open-source datasets is not always explicitly reported. This situation makes it difficult to assess data quality and make fair comparisons between datasets [45] .

Some studies in the literature emphasize the need for systematic quality evaluation of open-source EEG datasets. It is stated that datasets should be analyzed not only in terms of task description and label information, but also in terms of signal-to-noise ratio, artifact density, and recording conditions [46]. However, the number of such quality-focused evaluation studies is relatively limited in the literature.

Another important issue in the context of open-source EEG datasets is **data standardization**. Presenting different datasets in different formats makes data sharing and reuse difficult. In this context, the Brain Imaging Data Structure Brain Imaging Data Structure (BIDS) standard aims to provide a common struc-

ture for EEG and other brain imaging data [39]. BIDS aims to make datasets more transparent and reusable by establishing standards in many areas, from naming data files to metadata definitions.

The EEG-BIDS format is considered a significant development in the standardization of open-source EEG datasets and is increasingly adopted in the literature [47] . BIDS-compliant datasets allow different research groups to understand, process, and compare data more easily. However, there are still many open-source EEG datasets in the literature that are not fully compliant with the BIDS format. This indicates that standardization has not yet been adopted by the entire community.

The lack of data standardization affects not only data sharing but also the comparability of classification results. Applying different preprocessing steps to different datasets directly impacts model performance and makes fair comparisons between methods difficult. Some studies in the literature highlight that the lack of standardized preprocessing and evaluation protocols is one of the fundamental problems in the field of EEG classification.

Another important aspect in evaluating open-source EEG datasets is the **quality of the labeling and metadata.** The accuracy of the labels in the datasets directly affects the success of the classification models. In open-source datasets, how labeling processes are performed is not always reported in detail. This creates uncertainties about the reliability and consistency of the labels [48].

Some studies in the literature argue that open-source EEG datasets should be evaluated not only in terms of signal content, but also in terms of metadata richness. Metadata such as participant number, demographic information, experimental conditions, and ethical consent information are critical to the reuse of datasets [38] . However, many open-source EEG datasets have incomplete or limited information on this subject.

In general, open-source EEG datasets offer significant opportunities in EEG-based classification studies, but also present fundamental challenges such as heterogeneity, data quality, and standardization. A large portion of the literature focuses on proposing specific methods or models using these datasets, neglecting to systematically address their structural characteristics and limitations. This hinders the full utilization of the potential offered by open-source EEG datasets.

In this context, a comparative evaluation of open-source EEG datasets, an analysis of differences in quality and standardization, and the identification

of the impact of these differences on classification performance constitute a significant research gap in the literature. To truly appreciate the advantages of transparency and reproducibility offered by open-source data that prioritize dataset heterogeneity are needed.

Data labeling and evaluation processes in EEG-based classification studies have long been discussed in the literature as a critical component directly affecting model performance. Since EEG signals are inherently complex, noisy, and exhibit high inter-individual variability, accurately labeling these signals constitutes one of the most challenging aspects of classification problems. Much of the early EEG research in the literature relied on manual evaluation and labeling processes performed by field experts [49, 48].

Manual evaluation has long been considered the gold standard due to its reliance on expert knowledge and experience. In particular in clinical EEG analyzes and experimental studies, experts' ability to distinguish artifacts of the signal, identify patterns related to tasks, and interpret contextual information offers significant advantages [50]. However, the reliability and consistency of **manual assessment processes are increasingly being questioned in the literature.**

Many studies show that there can be significant inconsistencies between EEG assessments performed by different experts. Inter-rater variability has been reported, particularly in areas such as artifact detection, interpretation of borderline cases, and identification of task-related patterns[51, 52].This highlights the subjective aspects of manual labeling processes and suggests that labels based on the assessment of a single expert may be problematic when considered as absolute accuracy.

Another significant limitation of manual evaluation is **scalability. Manually evaluating large-scale EEG datasets or long-term recordings incurs significant costs in terms of time and human resources.** With the widespread availability of open-source EEG datasets, the need to analyze EEG data from hundreds or even thousands of participants has emerged. In this context, the literature frequently emphasizes that manual evaluation processes alone do not offer a sustainable solution [40, 38].

These limitations have increased interest in automated and semi-automated labeling approaches. Machine learning-based methods offer the potential for automatic classification and labeling of EEG signals based on specific characteristics. The use of automated approaches has become widespread, particularly

in areas such as artifact detection, signal quality assessment, and basic task classifications [45, 44].

The main advantage of automated labeling approaches is that they provide consistency and speed. Applying the same criteria to different datasets with the same algorithm has the potential to reduce evaluator-induced subjectivity. However, the literature emphasizes that the accuracy of automated labeling methods largely depends on the training data used and the quality of the labels. Models trained with incorrect or inconsistent labels inevitably produce erroneous classifications [53] .

At this point, it is noteworthy that the relationship between human-based evaluation and automated approaches has not been sufficiently addressed in depth in the literature. Most studies evaluate the performance of automated methods through model accuracy or similar metrics, rather than comparing it to manual labels. **Studies that systematically analyze the agreement, differences, and potential deviations between human evaluations and automated classifications are limited in the literature** [9].

The rapid advancements in artificial intelligence in recent years have led to the emergence of new approaches in data evaluation processes. Large Language Model (LLM), in particular, have begun to be used in various fields due to their ability to process contextual information and model complex relationships. A new research direction in the literature has emerged: using these models not directly to process EEG signals, but rather to analyze descriptions, metadata, and evaluation criteria related to EEG datasets [54].

GPT-based models, being trained on textual data, have the potential to analyze information such as technical descriptions, experimental protocols, and task descriptions of EEG datasets. Studies in the literature suggest that large language models can be used in areas such as data curation, label consistency analysis, and supporting evaluation processes [55, 56] . These approaches are positioned not as tools to completely eliminate manual evaluations, but as tools that support and complement human evaluations.

However, the literature on the use of large language models in data evaluation processes is still in its early stages. The reliability, consistency, and potential for bias of these models are among the important issues discussed in the literature. In particular, the possibility that LLMs may reflect biases stemming from training data necessitates the careful use of these models in scientific evaluation processes [57].A related effort in this direction is the work of Koksa [58], who developed

a reproducible NLP-based information retrieval pipeline specifically tailored to EEG-related literature on arXiv. That system combines keyword-based semantic filtering with synonym expansion and supervised classification of arXiv subject categories, demonstrating how transparent, deterministic text-mining tools can be effectively applied to monitor the evolution of EEG research at scale.

In the context of EEG, one of the most important contributions offered by large language models is that they allow for the analysis of differences between human-based assessment and automated methods. Textual descriptions of assessments made by different experts can be analyzed through GPT-based models to measure consistency. This approach makes it possible to examine evaluation processes not only through results but also through the logic of decision-making.

The number of studies in the literature that address manual assessment, automated labeling, and large language model-based approaches together is quite limited. Most studies focus on only one of these approaches and exclude the others. However, the complex nature of EEG-based classification problems makes it difficult for a single assessment approach to solve all problems. This situation highlights the need for multi-layered and comparative assessment strategies.

In general, the EEG classification literature contains significant methodological gaps in terms of data labeling and evaluation processes. The subjectivity of manual assessments, the data dependence of automated methods, and the still immature application areas of large language models are among the main reasons for these gaps. In this context, studies that comparatively address human-based assessment, automated classification, and LLM-based approaches stand out as a significant need in the literature.

When the EEG-based classification literature is evaluated, it is seen that the field has reached a significant level of maturity in terms of both methodological diversity and breadth of its application areas. Studies conducted in different contexts such as motor imagery, cognitive tasks, resting state, and clinical applications have shown that EEG signals can be analyzed with machine learning and deep learning approaches and meaningful classification results can be produced [24, 28, 13] . However, it is also clear that the successes presented in these studies have been largely achieved under certain assumptions and limitations.A significant portion of the studies in the literature evaluate classification performance through singular metrics and generally report the results using accuracy or similar measures. However, this approach is far from fully reflecting model success, given the complex nature of EEG data and structural differences be-

tween datasets [9]. The literature emphasizes that evaluations based on a single performance metric can be misleading, especially when dealing with heterogeneous datasets.

Machine learning-based approaches have long been preferred in EEG classification studies due to their low computational costs and relatively high interpretability. These methods have yielded successful results, particularly in early studies using hand-crafted feature-based representation approaches [26]. However, the performance of these approaches is largely dependent on the quality of the feature extraction process and the suitability of the selected features for the task. This dependence on the feature extraction process stands out as a significant factor that limits the generalizability of these methods across different datasets. Deep learning-based approaches have garnered significant attention in the literature due to their potential to overcome these limitations. CNN and derivative architectures offer a more flexible representation compared to traditional methods due to their ability to automatically extract features from raw or minimally preprocessed EEG signals [29]. However, it is also clear that the success of deep learning models is largely dependent on the size, quality, and homogeneity of the datasets. The literature contains numerous findings indicating that deep learning models trained on small or heterogeneous datasets experience generalizability problems [30, 31].

In this context, data set heterogeneity emerges as one of the most fundamental limitations in the literature. Although the increase in open-source EEG datasets offers significant opportunities for reproducibility, structural differences between these datasets make a direct comparison of classification results difficult [39, 47]. Differences in channel numbers, sampling frequencies, task descriptions, and preprocessing steps can prevent the same model from exhibiting consistent performance on different datasets [59].Data quality and artifact management are another limitation frequently highlighted in the literature. The noise-prone nature of EEG signals directly affects classification performance.**The number of studies that systematically address the impact of artifact cleaning and preprocessing steps on classification results is limited [46].** This constitutes a significant obstacle to fair and consistent evaluations in comparisons between datasets.

Labeling and evaluation processes constitute another critical dimension of the EEG classification literature. Although manual evaluations have been considered the gold standard for many years, **inter-rater inconsistencies and scala-**

bility issues have revealed the limitations of this approach [51, 52]. Although automated labeling approaches have the potential to address these problems, the precision and reliability of these methods depend largely on the quality of the training data [53].The recent emergence of large language models has opened a new research direction in data evaluation processes. LLM-based approaches are used not to directly classify EEG signals, but to analyze descriptions, metadata, and evaluation criteria of datasets [54]. Studies in the literature argue that such models can be used as tools to support human-based evaluations [56]. However, the extent to which these approaches are reliable and consistent in scientific evaluation processes remains an open research question.Another notable deficiency in the literature is the lack of systematic comparisons of different assessment approaches. Most studies focus on manual assessments or on automated classification results; studies that methodologically address the relationship between these two approaches remain limited [9]. Similarly, **studies comparing human-based assessments with LLM-based approaches are still in their infancy in the literature.**

In general, the current literature largely addresses EEG-based classification problems within an algorithmic framework; it relegates the effects of data evaluation processes, dataset heterogeneity, and lack of standardization on classification results to a secondary role. However, findings in the literature show that model success is strongly dependent not only on the algorithms used, but also on data integrity, labeling processes, and evaluation protocols [38].In this context, the need for comparative, multi-layered, and reproducible evaluation approaches in EEG-based classification studies is becoming increasingly evident. Systematic analysis of heterogeneous datasets, the integration of manual and automated evaluation approaches, and the examination of the effects of data integrity on model performance are among the significant research gaps in the literature. **The limitations revealed by current studies indicate the need for the development of more systematic frameworks in the field of EEG classification.**

This synthesis reveals the strengths and limitations of the existing knowledge in the literature; it also offers important clues about the directions in which EEG-based classification problems can be addressed in the future. These open problems identified in the literature require the development of more systematic, comparative, and reproducible approaches to the evaluation and classification of EEG data.

2

Background

This section details the fundamental concepts that form the methodological and conceptual framework of this thesis. The main objective of the Background section is to provide the necessary theoretical framework for understanding the analysis and modeling stages that will be presented in subsequent sections. Accordingly, the basic characteristics of electroencephalography (EEG) signals, the working principles of EEG-based brain-computer interfaces (BCIs), the general structure of the open-source EEG datasets used, manual evaluation and labeling processes, and the theoretical foundations of artificial intelligence and machine learning-based classification approaches are systematically explained.

The Background section summarizes widely accepted concepts and methods in the literature, revealing the theoretical foundation upon which the approaches used in this thesis are based. This structure allows the methods and experimental findings presented in the Analysis section to be evaluated within their context and in a consistent framework.

2.1 Basic Characteristics of EEG Signals

Electroencephalography (EEG) is a non-invasive neurophysiological technique used to measure electrical activity generated by neuronal populations in the brain [60] . It records voltage fluctuations resulting from ionic currents within cortical neurons, typically captured through electrodes placed on the scalp. Due to its high temporal resolution, EEG provides valuable insights into dynamic brain processes, making it widely used in cognitive neuroscience,

clinical diagnostics, and brain–computer interface (BCI) applications.

EEG signals are typically recorded from multiple channels over time, reflecting the dynamic nature of brain activity, as illustrated in Figure 2.1 [59]. In recent years, increased computing power and advancements in artificial intelligence have enabled more complex analysis methods of EEG data, significantly expanding the scope of research in this field.

EEG signals are typically recorded from multiple channels over time, reflecting the dynamic nature of brain activity, as illustrated in Figure 1 [59]. In recent years, increased computing power and advancements in artificial intelligence have enabled more complex analysis methods of EEG data, significantly expanding the scope of research in this field.

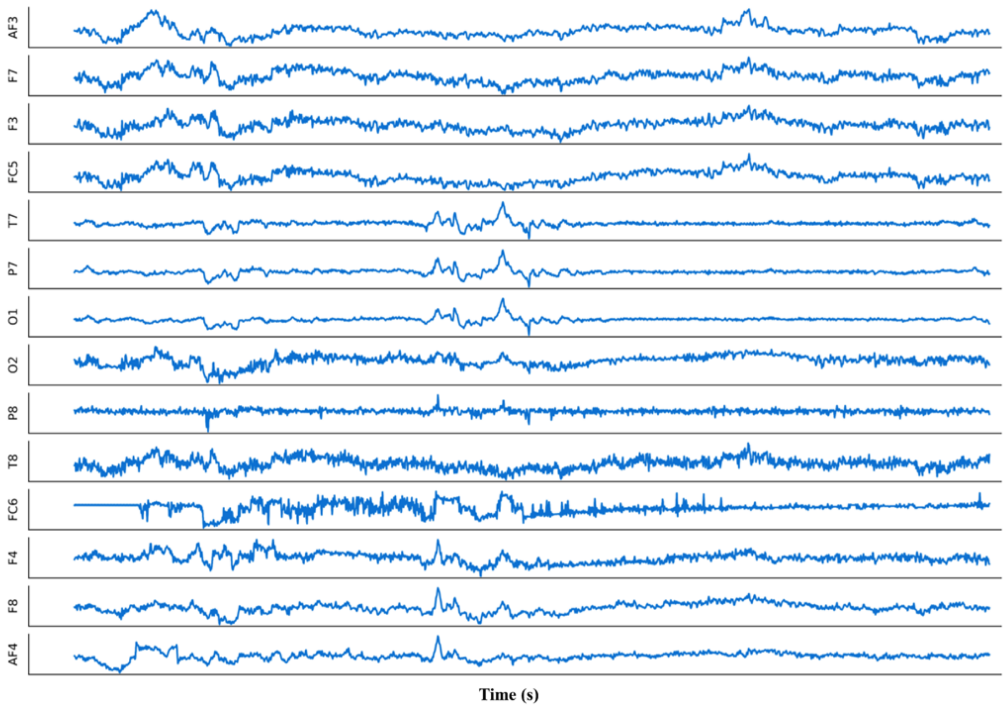

Figure 2.1: Example of multichannel electroencephalography (EEG) signals recorded over time [59]

The EEG signals examined in this thesis present a number of structural challenges in terms of classification and labeling processes. The low amplitude nature of EEG signals, their relatively low signal-to-noise ratio, and high sensitivity to measurement conditions directly affect both manual and automated evaluation processes. These characteristics cause EEG data to require a more complex analysis process compared to other types of biomedical signals [61] .

A significant feature of EEG signals is their non-stationary nature. Significant differences can be observed between EEG signals recorded at different times from the same individual; even within the same recording session, signal characteristics can change [62] . This increases the risk of classification models learning patterns specific to a particular time interval or dataset. One of the main reasons for considering heterogeneous datasets together in this thesis is to observe the effects of this variable nature of EEG signals on different data sources.

Artifacts observed in EEG signals are another critical factor affecting classification performance. Blinking, muscle activity, electrode contact problems, and environmental electrical interference add unwanted components to EEG recordings. These artifacts can lead to inter-rater inconsistencies, especially in manual assessment processes; and in automated classification approaches, they can cause the model to learn incorrect patterns. In this thesis, analyzing the effect of differences in quality and recording conditions between datasets on classification performance is of significant importance in this context.

The frequency components of EEG signals offer a significant representation space for classification problems. However, the variability of these frequency components depending on the task, individual, and recording conditions makes it difficult to define a fixed feature space. This situation limits the generalizability of methods based on hand-built features, while constituting one of the main reasons for preferring deep learning approaches with automated feature learning capabilities in this field. The CNN-based models used in this thesis aim to directly model this complex and variable structure of EEG signals.

EEG datasets reported in the literature can generally be described based on a set of common characteristics that are expected to be explicitly defined in the corresponding studies. These characteristics typically include the experimental setup, the task paradigm (e.g., motor imagery), the number of participants, the number and placement of EEG channels, the sampling frequency, and the overall recording conditions [9, 36]. Clearly defining these elements is essential for ensuring the interpretability, comparability, and reproducibility of EEG-based research.

A typical EEG data acquisition setup consists of several interconnected components: an EEG electrode cap placed on the participant's scalp, a signal amplifier, a data receiver, and a host computer running acquisition software. Electrodes are positioned according to standardized systems, most commonly the

International 10–20 system, which ensures spatial consistency across studies and laboratories [63]. Figure 2.2 illustrates a representative EEG data collection setup used in a motor imagery BCI study, where the participant performs left- and right-hand imagery tasks in response to on-screen cues while multichannel EEG signals are continuously recorded.

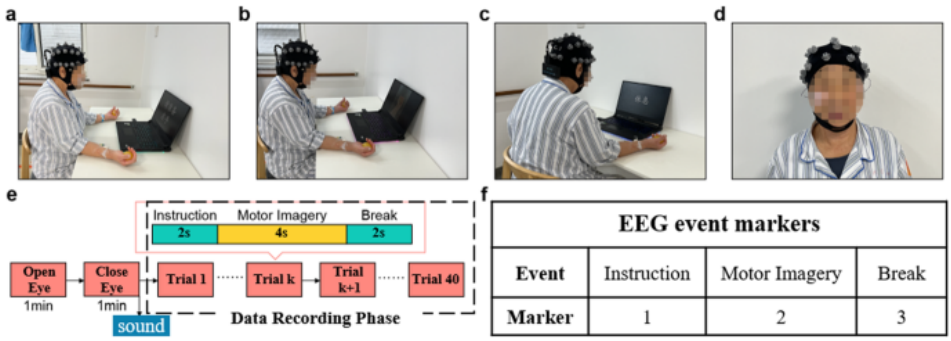

Figure 2.2: A representative EEG data acquisition setup for a motor imagery brain–computer interface study, illustrating the EEG cap, amplifier, data receiver, and host computer, along with the experimental trial structure for left- and right-hand motor imagery tasks. Modified from [64] under CC BY 4.0 license.

Among the most widely studied paradigms in open-source EEG research is motor imagery (MI), in which participants mentally rehearse limb movements—most commonly left- and right-hand movements—without executing them physically [65]. MI-based brain–computer interface (BCI) systems exploit the neural correlates of these imagined movements, particularly event-related desynchronizationerd (ERD) and synchronization (ERS) patterns in the mu (8–12 Hz) and beta (13–30 Hz) frequency bands over sensorimotor cortex [65, 66]. Because these patterns are physiologically well characterized, MI paradigms have become a standard benchmark for evaluating EEG classification algorithms [9, 28].

Several open-source motor imagery EEG datasets have been made publicly available and are widely used in the research community. The BCI Competition IV Dataset 2a [67] provides EEG recordings from nine participants performing four classes of motor imagery (left hand, right hand, feet, and tongue), recorded with 22 channels at 250 Hz. The PhysioNet EEG Motor Movement/Imagery Dataset [42] contains recordings from 109 participants across 14 experimental runs including real and imagined movements of the left and right fists and feet, acquired using 64 channels at 160 Hz. GIGA MI dataset [68] includes 52 subjects performing left and right hand imagery, recorded with 64 channels at 512 Hz. The OpenBMI dataset [69] offers a large-scale benchmark with 54 participants

and three BCI paradigms including MI, recorded at 1000 Hz with 62 channels.

Table 2.1 summarizes the key characteristics of these datasets. It is evident that the datasets differ substantially across all major dimensions: the number of participants ranges from 9 to 109, channel counts vary from 22 to 64, and sampling frequencies span 160 Hz to 1000 Hz. This technical heterogeneity directly affects the representation of EEG signals and, consequently, the performance and generalizability of classification models trained on these data [36, 28].

Table 2.1: Summary of key characteristics of commonly used open-source motor imagery EEG datasets. MI = Motor Imagery; L/R = Left/Right hand.

Dataset	**Subjects**	**Channels**	**Sampling Rate (Hz)**	**MI Tasks**
BCI Comp. IV 2a [67]	9	22	250	L/R hand, feet, tongue
PhysioNet [42]	109	64	160	L/R hand, feet
GIGA [68]	52	64	512	L/R hand
OpenBMI [69]	54	62	1000	L/R hand

The heterogeneity observed across these datasets is not merely a technical inconvenience; it reflects the genuine diversity of experimental contexts in which EEG data are collected. Different laboratories employ different amplifier systems, electrode configurations, and recording protocols, leading to datasets that vary in spatial resolution, temporal resolution, and signal quality [9]. For example, datasets with higher channel counts provide greater spatial resolution and better capture localized cortical activity, while datasets with higher sampling rates preserve fine-grained temporal dynamics relevant to event-related potentials. Conversely, datasets with fewer channels or lower sampling rates may limit the information available to classification algorithms, making direct cross-dataset comparisons unreliable without appropriate harmonization.

Data quality constitutes another critical dimension of variability among open-source EEG datasets. Some datasets provide raw, unprocessed signals, whereas others supply pre filtered or artifact-rejected data [36]. Artifacts arising from muscle activity, eye movements, and electrical interference are well-documented challenges in EEG recording [70]. The absence of a unified preprocessing standard means that the same underlying neural signal may be represented very differently across datasets, depending on the artifact removal methods applied prior to publication. In the context of this thesis, the effects of such quality differences on classification performance are analyzed directly, rather than treating data quality as an exclusion criterion.

Metadata richness represents a further source of variability. Information regarding participant demographics, experimental conditions, and ethical approval procedures is essential for the reusability of a dataset and the interpretation of classification results [47]. Well-documented datasets such as the PhysioNet collection [42] provide detailed participant-level metadata, whereas other datasets offer only aggregate information. In this thesis, available metadata including task descriptions and recording protocols is treated as contextual information that informs both the manual evaluation and the GPT-assisted classification processes described in subsequent chapters.

The structural heterogeneity of open-source EEG datasets constitutes one of the primary motivations for the multi-stage evaluation methodology developed in this thesis. By selecting datasets that differ systematically in participant count, channel configuration, sampling frequency, and preprocessing level, this study examines the conditions under which manual annotation, GPT-based review, and model-based classification converge or diverge. Datasets are therefore treated not merely as sources of training data, but as structured testing environments that reveal the sensitivity of classification approaches to variation in data collection protocols.

2.2 Principles of Systematic Literature Review

Systematic review methodologies provide a structured framework for synthesising existing evidence in a given research domain [71, 72]. In the context of EEG-based research, systematic review approaches have been applied to assess the quality, reproducibility, and comparability of published datasets and classification studies [9, 36]. A key requirement of any systematic review is the explicit definition of inclusion and exclusion criteria, the transparent documentation of the review process, and the consistent application of evaluation criteria across all reviewed items [71].

In this thesis, a human-based systematic review process was applied to a curated set of open-source EEG datasets. The primary objective of this review was to characterise each dataset according to a predefined set of descriptive criteria, including task paradigm, number of participants, recording parameters, preprocessing procedures, and metadata availability. This structured characterisation was intended to provide a consistent and reproducible basis for subsequent analyses, rather than to produce a quantitative performance assessment.

The review was conducted by four individuals with varying levels of domain knowledge and experience in EEG research. This multi-reviewer design was adopted to reduce dependence on the judgement of a single expert and to introduce a degree of variability that reflects the ambiguity inherent in interpreting heterogeneous dataset documentation [73]. Each reviewer examined the available documentation for each dataset including published data descriptor articles, associated files and recorded their observations using a structured data collection form. The collected responses were consolidated in a shared spreadsheet for subsequent comparison and analysis.

It is important to note that the review process described here constitutes a knowledge transfer exercise rather than a formal inter-rater reliability study. The lead investigator communicated the evaluation criteria and the conceptual framework of the study to the participating reviewers prior to the review process. This qualitative phase of knowledge transfer ensured a shared understanding of the review objectives without imposing a rigid coding protocol. As a result, the process captures the naturalistic variability that arises when individuals with different backgrounds interpret the same dataset documentation, which is itself a phenomenon of analytical interest in this study.

One of the principal challenges encountered in systematic review processes of this kind is the consistent definition and application of evaluation criteria across heterogeneous sources [72]. Dataset documentation in the EEG literature varies considerably in depth and format: some datasets are accompanied by detailed data descriptor articles published in peer-reviewed journals [42, 69], whereas others provide only minimal metadata. This variability means that the same evaluation criterion may be straightforwardly applicable to one dataset and ambiguous or unanswerable for another, introducing an inherent asymmetry into the review process.A further methodological consideration concerns the scalability of human-based review. As the number of publicly available EEG datasets continues to grow [47], the time and effort required to conduct thorough manual reviews increases correspondingly. Within the scope of this thesis, the human-based review was applied to a limited and deliberately selected set of datasets. This constraint is acknowledged as a limitation, and it provides one of the motivations for exploring GPT-assisted review as a complementary approach, as discussed in the following section.

Figure 2.3 illustrates the workflow of the human-based review process, from the initial definition of evaluation criteria through the knowledge transfer phase

to the final consolidation of reviewer responses.

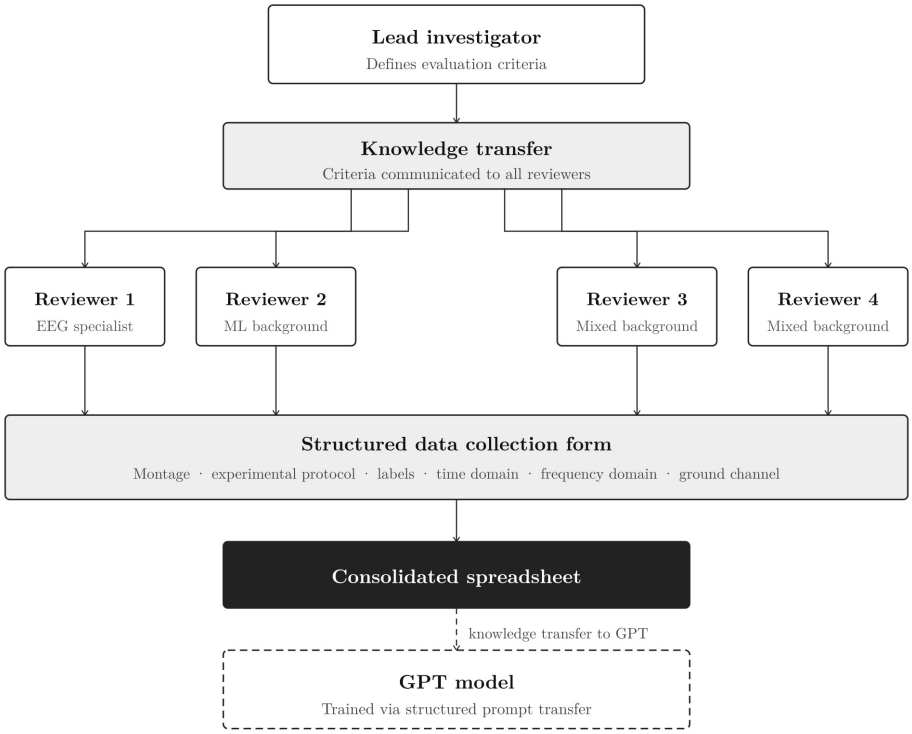

Figure 2.3: Workflow of the human-based systematic review process. The lead investigator defined the evaluation criteria and communicated them to four reviewers through a knowledge transfer process. Each reviewer independently examined dataset documentation and recorded observations using a structured form. Responses were consolidated in a shared spreadsheet for subsequent analysis.

2.3 GPT-Assisted Literature Review

Large language models LLM , and GPT-based models in particular, have attracted growing interest as tools for supporting systematic literature review and information synthesis in scientific research [74, 75]. These models are capable of processing and summarising large volumes of textual information, identifying thematic patterns, and responding to structured queries formulated in natural language [74]. In the context of this thesis, a GPT-based approach was employed as a complementary method for reviewing the EEG dataset literature,

operating in parallel with the human-based review described in the preceding section.

It must be emphasised at the outset that the GPT-based component of this study does not constitute an automated classification system operating on EEG signals. Rather, GPT was used as a literature review instrument: structured prompts were formulated to query the model about characteristics of EEG datasets and their associated publications, and the model's responses were recorded and compared with the findings of the human reviewers. The inputs to the model were entirely textual task descriptions, dataset summaries, and methodological descriptions drawn from published sources and no raw EEG signal data were processed at any stage.

The use of GPT as a literature review tool offers several potential advantages in the context of EEG dataset analysis. First, LLMs can apply a defined set of evaluation criteria consistently across a large number of sources without the fatigue or drift that may affect human reviewers over extended review sessions [75]. Second, because the model's responses are generated from the same prompt structure for each dataset, the review process is in principle reproducible given the same model version and prompt formulation. Third, the systematic application of prompts to a body of literature can surface patterns in dataset documentation that might be difficult to detect through manual review of individual sources.

The practical workflow adopted in this thesis involved the formulation of structured prompts designed to elicit descriptive information about EEG datasets from the published literature. Prompts were constructed to mirror the evaluation criteria used in the human review process, thereby enabling a direct comparison of human and GPT-generated assessments of the same datasets. This parallelism is central to the analytical design of the study: by applying the same conceptual framework to both review methods, it becomes possible to examine the extent to which a language-model based approach can reproduce the evaluative judgements made by human reviewers operating under a shared knowledge framework.

It is acknowledged that GPT-based literature review is subject to a number of important limitations. LLMs can produce responses that are plausible in form but inaccurate in content, a phenomenon commonly referred to as hallucination [61]. This risk is particularly relevant when querying models about specific technical details of individual datasets, where the model's training data may be

incomplete or outdated. Furthermore, the outputs of GPT models are sensitive to the precise formulation of the input prompt; small variations in wording can lead to substantively different responses. These limitations mean that the GPT-assisted review results reported in this thesis should be interpreted as exploratory findings that illuminate the potential and boundaries of LLM-based review tools, rather than as definitive assessments of the datasets in question.

In summary, the GPT-assisted literature review component of this thesis represents a qualitative and exploratory investigation into the applicability of language models as support tools for systematic EEG dataset analysis. The findings contribute to an emerging body of methodological research on the integration of AI tools into scientific review processes [75], and provide a basis for more rigorously controlled investigations in future work.

2.4 Machine Learning and Deep Learning Models

In this thesis, machine learning and deep learning models are employed for the classification of left- and right-hand motor imagery EEG signals across a heterogeneous set of open-source datasets. The comparative use of a classical machine learning method and a deep learning architecture serves a dual purpose: first, to assess the classification performance of each approach on the selected datasets; and second, to examine how structural differences between datasets such as channel configuration, sampling frequency, and preprocessing level affect model behaviour. This framing treats classification models not merely as performance benchmarks, but as analytical instruments for probing dataset characteristics.

Machine learning and deep learning approaches differ fundamentally in how EEG data is represented and processed. Classical machine learning methods typically rely on hand-crafted feature extraction pipelines, in which domain-relevant signal features are computed explicitly before being passed to a classifier [9, 76]. Deep learning models, by contrast, learn feature representations directly from raw or minimally processed data through hierarchical transformations [28, 36]. Evaluating both approaches on the same datasets makes it possible to assess whether performance differences reflect genuine algorithmic advantages or are better explained by dataset-level factors such as signal quality and recording conditions.

SVM are supervised learning algorithms widely used for the classification

of high-dimensional biomedical signals, including EEG [9, 76]. The core principle of SVM is to identify the optimal separating hyperplane that maximises the margin between classes in the feature space [77]. When the data are not linearly separable in the original input space, SVMs employ kernel functions to project the data into a higher-dimensional space where linear separation becomes feasible. Commonly used kernels in EEG classification include the radial basis function Radial Basis Function (RBF), polynomial, and linear kernels [9].

In EEG-based motor imagery research, SVM has been extensively applied to features derived from the EEG signal, most notably band power features computed from the mu (8–12 Hz) and beta (13–30 Hz) frequency bands, and spatial features extracted using methods such as Common Spatial Patterns Common Spatial Pattern (CSP) [76]. SVM is particularly well suited to EEG classification tasks owing to its robustness in high-dimensional feature spaces and its ability to generalise from relatively small training sets [77, 9], both of which are common constraints in EEG research. In the context of this thesis, SVM serves as the classical machine learning baseline, allowing the performance of the deep learning model to be interpreted relative to an established reference method.

Convolutional Neural Networks (CNNs) are deep learning architectures that have demonstrated strong performance in EEG signal classification tasks by learning spatial and temporal feature representations directly from the data [28, 29]. Unlike classical methods that require explicit feature engineering, CNNs apply learned convolutional filters to the input signal, enabling the model to identify task-relevant patterns at multiple scales without prior specification of the features of interest [36].

EEG signals have a natural multi-channel, time-series structure that is well suited to convolutional processing. Temporal convolutions capture the dynamics of neural oscillations over time, while spatial convolutions across electrode channels exploit the topographic organisation of cortical activity [28]. This dual sensitivity is particularly valuable when working with heterogeneous datasets that differ in channel count and sampling rate, as the model must learn representations that are informative despite variation in the input structure.

Among the CNN architectures proposed for EEG classification, EEGNet [29] has emerged as a widely adopted compact model that is applicable across multiple BCI paradigms and dataset configurations. Its design incorporates depthwise and separable convolutions to reduce the number of trainable parameters, which helps to mitigate overfitting in the small-sample settings typical of EEG

research. In this thesis, a CNN-based model is applied to the motor imagery classification task across the selected datasets, and its performance is compared with that of the SVM baseline to evaluate the relative merits under heterogeneous data conditions.

2.5 Related Work

Recent studies have extensively explored the use of electroencephalography (EEG) signals for cognitive and physiological state analysis. For instance, Giulia Cisotto et al. [78] proposed a machine learning-based approach for classifying cognitive workload using in-ear EEG devices. Their work focuses on wearable EEG acquisition and real-time workload estimation. Similarly, Cisotto et al. [79] investigated cognitive performance under sleep deprivation using machine learning techniques applied to EEG signals. Earlier works by the same research group include the classification of grasping tasks based on EEG-EMG coherence and joint compression of EEG and EMG signals for wireless biometric applications [80, 81]. Additionally, Anna V. Guglielmi et al. [62] analyzed frequency-dependent functional connectivity in resting-state brain networks.

Although these studies are closely related to our work in terms of utilizing EEG signals and machine learning techniques, they differ in several key aspects. First, prior studies primarily focus on specific application domains such as cognitive workload detection, sleep deprivation analysis, or motor task classification, whereas our study adopts a broader perspective by systematically analyzing multiple open-source EEG datasets. Second, existing works typically rely on a single dataset or a controlled experimental setup, while our work explicitly investigates cross-dataset variability and its impact on classification performance. Furthermore, previous studies do not address the heterogeneity in dataset reporting or methodological inconsistencies, which constitute a central focus of our research. In contrast, our study introduces a structured framework for evaluating dataset characteristics and their influence on classification outcomes.

Machine learning techniques have also been widely applied in biomedical data analysis, particularly with an increasing focus on model interpretability. For example, Leonardo Badia and collaborators investigated the use of Shapley values to interpret machine learning models in heart disease datasets [82]. Similarly, Borella et al. [83] explored effective sensor selection for human activity recognition using Shapley-based feature importance methods. More recently,

Cisotto et al. [60] applied Shapley value analysis to Random Forest models for sleep apnea detection. While these studies share common ground with our work in employing machine learning techniques for classification tasks, their primary focus lies in model interpretability, feature importance, or adaptive learning strategies. In contrast, our work emphasizes cross-dataset classification performance and the role of dataset characteristics in influencing model outcomes. Moreover, unlike prior works that typically evaluate models on a single dataset, our study compares the performance of both traditional (SVM) and deep learning-based (CNN) approaches across multiple open-access EEG datasets. Another key distinction is that our work integrates a systematic literature review and AI-assisted data extraction process, which is not addressed in the aforementioned studies. Overall, the reviewed literature demonstrates the potential of EEG-based analysis and machine learning techniques in various biomedical applications. However, many existing studies tend to focus on specific EEG-based tasks or on improving model performance and interpretability within a single dataset context.

In this context, the present study aims to introduce a more comprehensive framework that combines systematic dataset analysis, GPT-based methodological information extraction, and cross-dataset classification using both SVM and CNN models.By considering the heterogeneity in EEG dataset reporting and examining how dataset characteristics may influence classification performance, this study is expected to contribute to a more systematic and transparent evaluation of EEG-based machine learning approaches. In this respect, the proposed approach may offer an alternative perspective for improving the comparability and reproducibility of studies in this field.

3

Materials and Methods

This section presents in detail the experimental processes, methods used, and analyses of the results obtained within the scope of this thesis. This section presents a pilot analysis of manual evaluation, GPT based classification, and machine learning-based approaches applied to open-source EEG datasets. The experimental design and comparison strategies are briefly described.

In line with the theoretical and methodological framework presented in previous sections, a multi stage evaluation approach has been adopted in this thesis. EEG datasets were analyzed not only with a single classification method but also through different evaluation layers. The main motivation for this approach is to demonstrate that model performance in EEG based classification studies is strongly dependent not only on the algorithms used but also on the structure of the datasets, the labeling processes, and the evaluation methods. Therefore, the experimental processes presented in the Analysis section are designed to reflect this multi dimensional structure.

The experimental process generally consists of three main part. In the first stage, EEG datasets from Nature Scientific Data were manually evaluated and classified [84]. This evaluation, conducted by individuals with varying levels of knowledge, allowed for the analysis of consistency and differences in human based classification processes. In the second phase, the same datasets were evaluated using a GPT based classification approach and compared with the manual evaluation results. This phase aims to examine the extent to which AI powered systems can reproduce human based evaluations and where they differ [85].

In the third stage, EEG datasets were classified using machine learning models. Specifically, CNN and SVM were employed [86]. The performance of these models was evaluated across different datasets and labeling strategies, and the effects of dataset heterogeneity and evaluation methods on model performance were analyzed [87]. Thus, the study distinguishes between human-based and GPT-supported analysis of EEG-related literature and the use of machine learning models for motor imagery EEG classification , highlighting their complementary roles within the overall framework.

Reproducibility and transparency are considered important aspects of this work. However, the public release of code, GPT based tools, and associated resources is planned as a future perspective, following further development and formal publication of the study.

This section continues by first defining manual evaluation and GPT based classification processes are presented in detail. Subsequently, classification experiments performed with CNN and SVM models are explained, and the evaluation metrics used are defined. The section concludes with a comparative presentation of the results obtained from different evaluation approaches. This structure allows the experimental findings of the thesis to be evaluated within a consistent framework.

The workflow presented in Figure 3.1 illustrates how the three core components of the study are organized within a sequential and integrated structure. The process begins with a systematic manual review of EEG datasets from the literature. In this stage evaluation criteria are refined by a panel of experts from diverse experience levels and disciplines, resolving interpretation discrepancies and establishing a common standardization framework. In the second stage, this framework is transferred to the GPT based model, ensuring that the model adheres to specific rules and automatically extracts information from the datasets. In the final stage, both traditional and deep learning models were applied to the constructed datasets, and their classification performance was evaluated across multiple datasets. This allowed for the examination of model performance under heterogeneous dataset conditions and the impact of dataset variability on classification results.

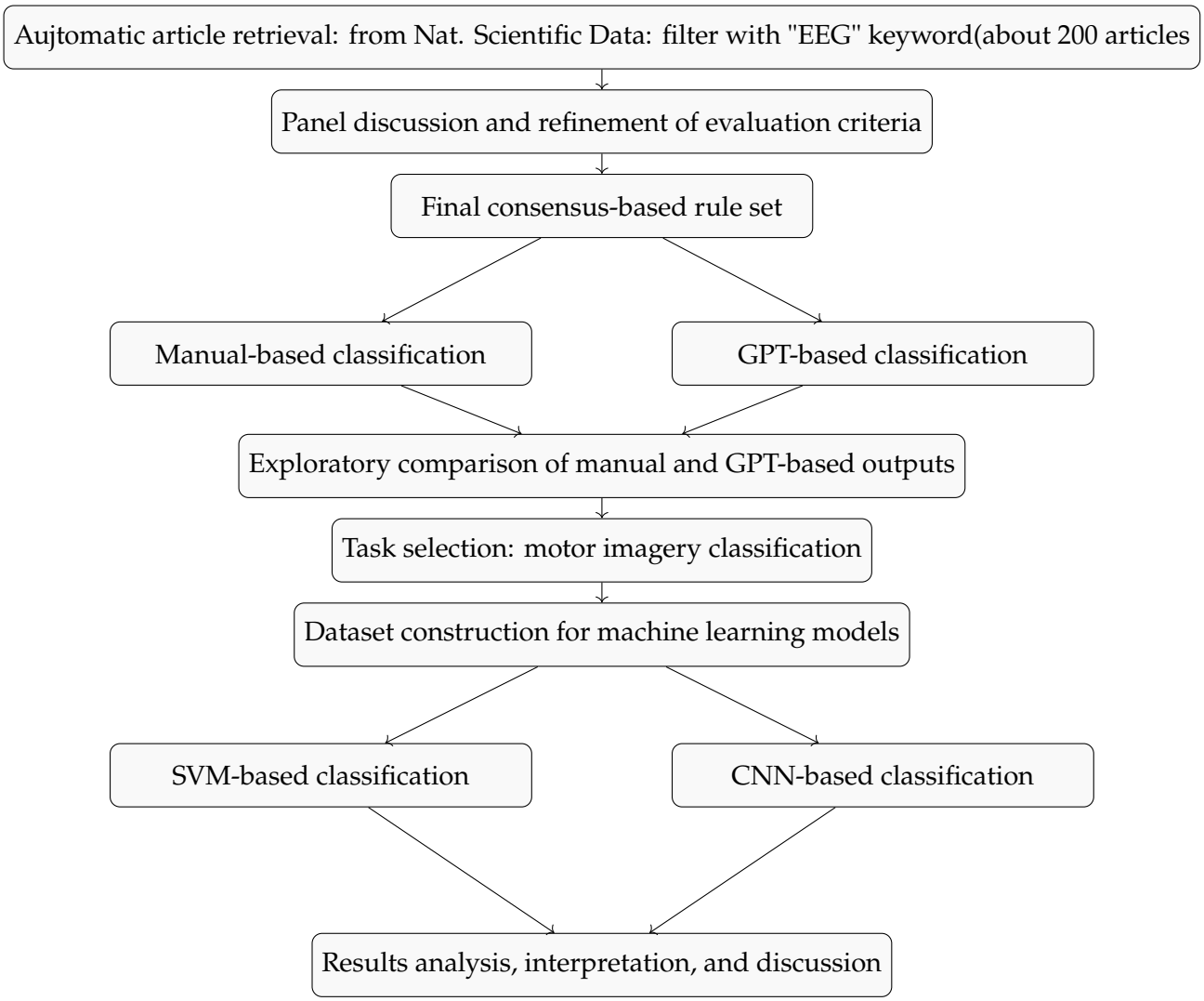

Figure 3.1: Flowchart of the proposed multi stage evaluation and analysis framework

3.1 Scientific Systematic Review

In this thesis, scientific articles containing EEG data from only Nature Science Data were systematically reviewed and classified. The aim of this phase of the study was to evaluate the quality of the EEG datasets used in the iteration and to identify high quality datasets. Manually evaluating all datasets in the initial phase of the study was a critical choice for this study. Because reporting styles in EEG studies are quite heterogeneous in the literature, automated methods can miss critical information.

In this study, the manual evaluation of EEG datasets was not left to the judgment of a single expert. Similar to multi expert decision making approaches commonly used in healthcare, the decision making mechanism was conducted by a team of four people with varying experience levels and disciplines[88, 89]. This team consisted of one researcher with academic expertise in the field (expert evaluator); one researcher who was mentored by the expert throughout this thesis and acquired fundamental EEG knowledge (trained student); one researcher from a discipline related to EEG or neuroscience (related field researcher); and one researcher outside the EEG field with knowledge of data analysis and methodology (different field researcher). The primary purpose of establishing this mechanism is to increase both consistency and standardization by bringing together different perspectives in the evaluation of EEG datasets. Various studies in the literature emphasize that collaborative assessment by multiple experts is preferable to increase the accuracy and reliability of clinical decision making processes; multidisciplinary teams produce more comprehensive, systematic, and reproducible decisions. This multi expert approach is widely used, particularly in healthcare settings where complex data structures are evaluated.

- Expert evaluator: A researcher with academic experience and expertise in EEG signal processing and experimental design.
- Trained (student) evaluator: This individual was trained according to the instructions provided by the expert and served as a mid level evaluator, testing the feasibility of the standardization process.
- Related field researcher evaluator: This individual, working in adjacent fields such as neuroscience or biomedical engineering, contributed to the evaluation of borderline cases through their technical knowledge.
- Different field researcher evaluator: This evaluator, lacking fundamental

EEG knowledge but possessing a scientific understanding of evaluation, helped validate the criteria for understandability and generalizability.

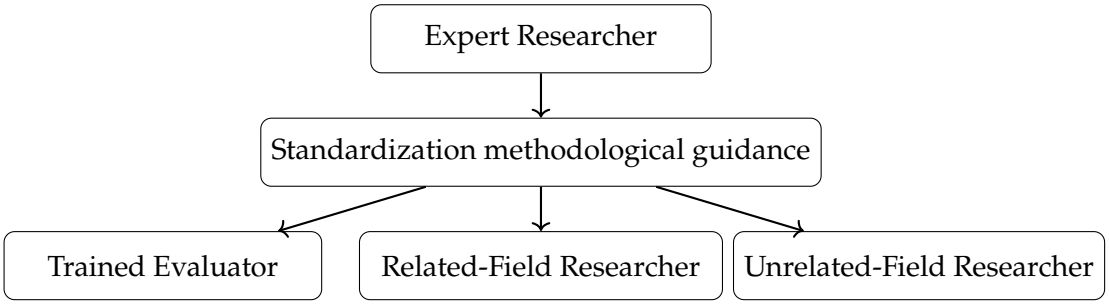

Figure 3.2: Structure of the Multi-Level Evaluation Team

Figure 3.2 illustrates the hierarchical structure and roles of the multi expert panel contributing to the manual evaluation process. The expert researcher defined the methodological framework, establishing the basis for standardization; the trained student served as the framework's implementer. The researcher in the relevant field verified the technical details, while other field researchers contributed to the process by ensuring the clarity, interpretability, and generalizability of the criteria. This structure both improved the evaluation quality and facilitated the creation of a common rule set for training the GPT model.

Several studies have shown that interobserver agreement in EEG interpretation may be limited when assessments are performed by a single expert. However, the use of multiple raters and standardized evaluation criteria has been demonstrated to significantly improve consistency and reliability[90, 91]. Furthermore, systematic reviews have emphasized that clinical decision making processes conducted by multidisciplinary teams provide more comprehensive and reproducible results, and that collaboration among different areas of expertise, particularly in neuroscience and EEG interpretation, increases data quality and consistency [92, 93].

All articles were independently evaluated against a common set of criteria established by the team, and the results were recorded in a single, standardized tabular format. This structure ensured both comparability of datasets and the creation of a common reference framework for subsequent GPT.

All reviewers independently reviewed different datasets and entered scores and explanations into an Excel spreadsheet based on predefined criteria. The classification is structured into three levels.The criteria classified in 3 stages as level 0, 1 and 2 in the study are as follows:

- Level 0: Units, sampling frequency, montage information (number of channels, channel location, reference information), experimental protocol (participants, sessions, blocks, task, rest period, timing), labels.
- Level 1: Time domain, frequency domain.
- Level 2: Output format, download, brand, ground channel.

After the first round of evaluations, datasets that yielded particularly different results were identified, and structured discussions were held within the team. During these discussions, the expert reviewer explained the scientific background of the criteria in detail; the trained student translated these explanations into rules; the relevant field researcher assessed compliance with reporting standards in the literature; and the other field researcher contributed to the clarity, applicability, and openness to interpretation of the criteria.

As a result of this interaction, criteria definitions were clarified, sample lines were created, and strict rules were established regarding the use of the "not found" label. Furthermore, a common decision framework was established for limitedly reported technical elements such as channel placement, electrode information, or protocol timing. This reduced interpretive differences between raters and ensured the creation of consistent evaluation criteria.

The scores and descriptions obtained for each dataset were systematically entered into an Excel spreadsheet, enabling both quantitative and qualitative comparisons. This method provided an objective filter for selecting the datasets to be used in the second phase of the thesis. Furthermore, this approach not only helps with reporting results but also provides a reproducible methodology that allows other researchers to make similar dataset selections in the future. Evaluators (high, medium, low, and naïve experience levels) contributed to this standardization process with varying levels of expertise.High experienced expert evaluators ensured the accuracy of technical parameters, while low-experienced evaluators tested the clear understandability of the criteria. This multilayered structure exemplifies the approach known in the literature as "human in the loop data curation" [94, 95].

The manual evaluation process was designed not only to classify datasets but also as a process of gaining experience to achieve standardization of evaluation criteria. The goal of this process was to develop a consistent, reproducible, and teachable framework for measuring the quality of EEG datasets.

These discrepancies are a frequently highlighted problem in the literature regarding EEG data reporting. For example, [47] and [96] stated that inadequate

reporting standards make data sharing and reuse processes in EEG research difficult. Therefore, the manual evaluation phase in this study was viewed not only as data classification but also as a learning process for standardization training and resolving interpretative differences.

Consequently, this process is a critical step not only for evaluating datasets but also for establishing the conceptual foundation necessary for training the GPT model.In the final stage, the common standard established by this mechanism was transferred to the GPT based model. All rules agreed upon by the panel were clearly reflected in the training instructions of GPT, ensuring that GPT only uses explicitly reported information, marks missing information as "not found," and does not make any interpretations or predictions. Thus, GPT is positioned not merely as a language model but as an implementer of a predefined set of rules reflecting the consensus of multiple experts.

Discussions between evaluators and standardization of criteria provided the necessary empirical knowledge for the GPT to subsequently extract the same information autonomously. Therefore, manual evaluation is not merely a preliminary step in the study; it is the fundamental methodological step that enables the cognitive calibration of the model. This phase not only assessed the content quality of the articles but also created a reference database to test which information the GPT model could correctly extract. Therefore, the manual assessment served as the ground truth for measuring the accuracy of subsequent automated analyses.

This design offers a hybrid assessment architecture that centers human expertise and integrates it with automated systems. Both manual and GPT based assessment processes are based on this multi expert standardization, thus increasing both the reliability and reproducibility of the findings.

3.2 Dataset Overview and Access

In this stage of the study, approximately 200 articles related to EEG were collected using the keyword "EEG" from publicly available sources, particularly the Scientific Data journal (Nature), which provides open access datasets and associated publications. These articles are generally studies created by international research projects that aim to collect and share EEG signals in different types of tasks. All datasets in the study are ethics committee approved, anonymized, reproducible, and provided in BIDS (Brain Imaging Data Structure) compatible

formats [47].

Motor Imagery (MI) refers to the mental simulation of a motor action without any actual physical movement [97]. During motor imagery tasks, specific patterns are observed in EEG signals, particularly over the sensorimotor cortex [25]. These patterns are typically characterized by event-related desynchronization (ERD) and event-related synchronization (ERS) in specific frequency bands, such as the mu (8–13 Hz) and beta (13–30 Hz) rhythms. In particular, imagined movements are associated with a decrease in power (ERD) in the contralateral motor cortex, reflecting the activation of motor-related brain regions [98]. These neurophysiological responses form the basis for distinguishing different motor imagery tasks in EEG-based brain-computer interface systems [99, 100].

Table 3.1: Datasets

Article ID Reference	**Task Type**	**Number of Participants**	**Type**
Dataset-1 [101]	Motor Imagery	60	Healthy subjects
Dataset-2 [64]	Motor Imagery	50	Stroke patients
Dataset-3 [102]	Motor Imagery	62	Healthy subjects

The Table 3.1 summarizes the datasets. This classification step, which was performed before the modeling phase, aims to ensure that comparative analyses in the following sections progress with accurate results.

In this thesis, the selection of EEG datasets aims to maximize data quality and experimental validity. Although all of the datasets used were selected from open access platforms, both technical and contextual criteria were taken into consideration during data selection. The criteria are two basic categories which is inclusion criteria and exclusion criteria.

The selected data sets must adhere to all the specified inclusion criteria. Each inclusion criterion is coded with different numbers. The inclusion criteria determined for this study are listed below.

- IN1: The dataset must include at least 19 EEG electrodes, ensuring sufficient spatial resolution.
- IN2: A minimum of 10 participants must be involved to allow for statistically meaningful analysis.

- IN3: Subject do not perform classification or prediction in experimental paradigms (e.g., classification, prediction, or decoding of cognitive/motor tasks).
- IN4: The experimental protocol must be clearly documented, including task flow, stimulus design, and block structure.
- IN5: The EEG signals must be sampled at a rate of at least 128 Hz to prevent information loss.
- IN7: Only real, non-simulated EEG recordings were considered.
- IN8: Datasets must be recorded from human subjects .

The exclusion criteria are listed below with specific code.

- EX1: Studies based on neurostimulation techniques such as Transcranial Magnetic Stimulation (TMS), Deep Brain Stimulation (DBS), or similar interventions were excluded.
- EX2: Passive EEG recordings that do not include a clear classification objective or task definition were excluded.
- EX3: EEG data acquired from non-human subjects (e.g., rat EEG) were excluded from the analysis.
- EX4: Intra-cranial brain recordings(LPF,EEG,ECoG) were excluded.

Based on the defined inclusion and exclusion criteria, the collected datasets were filtered. While all selected articles (approximately 200) were used in the manual and GPT based evaluations, three representative datasets were specifically chosen for the machine learning experiments. The characteristics of these datasets are described in the following section.

Dataset 1 is one of the open access motor imaging EEG database published [101]. This dataset contains EEG recordings from 60 volunteer participants who participated in motor imaging based brain computer interface experiments. Each participant's experiment session was designed with a total of 6 runs, each containing 40 trials, resulting in 240 trials per participant. Thus, the theoretical total size of the dataset reaches approximately 14,400 trials. However, as stated in the article about the dataset, some participants' recordings were incomplete or interrupted due to technical reasons. Data loss was reported in some runs, particularly for participants A40 and A59. Therefore, these missing recordings were not included in the study, and a total of 13,920 trials were used in the analyses. EEG signals were recorded at a sampling frequency of 512 Hz, and 27 active EEG electrodes were used during the data collection process. However, in this

study, in order to ensure the comparability of the classification process among the datasets, the analysis was performed by selecting 5 channels consisting of C3, Cz, C4, P3 and Pz electrodes, which are common to all datasets.

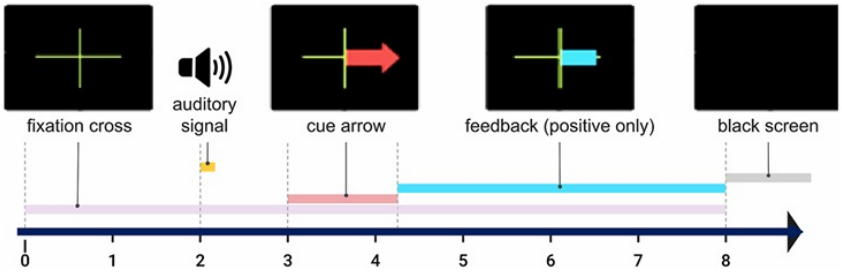

Figure 3.3: Experimental Setup of Dataset 1 [101]

Dataset 2 is an open access dataset containing EEG recordings from acute stroke patients for use in brain computer interface research based on motor imagery. The dataset includes EEG signals collected from a total of 50 acute stroke patients and aims to analyze brain activity during motor imagery. During the experiments, participants were asked to mentally visualize left and right hand movements (motor imagery), thus creating a two class classification problem.In the experimental protocol, motor imagery experiments were performed for each participant according to a specific task structure. Each trial consists of three main phases: a preparation phase where task instructions are given, a motor imagery phase where the participant mentally visualizes the relevant movement, and a short rest period. This structure allows for the analysis of brain activity during motor imagery in a specific time segment. A total of 40 trials were performed for each participant, and there are approximately 2000 trials in the dataset overall.EEG signals were recorded using a multi channel EEG head positioned according to the international 1020 electrode placement system, and a total of 32 channels (30 EEG channels and 2 EOG channels) were used in the data acquisition process. Recordings were obtained at a sampling frequency of 500 Hz. Furthermore, to improve signal quality, the reference electrode was placed at position Cz and the grounding electrode at position Fpz. The raw EEG signals obtained during data acquisition were then subjected to various preprocessing steps, and the time intervals corresponding to the motor imagery task were separated for analysis.A significant feature of this dataset is that the EEG recordings were obtained from a clinical population, namely patients with acute stroke. Neurophysiological changes occurring in the motor cortex after stroke

can cause the EEG patterns observed during motor imagery to be more variable and weaker compared to healthy individuals. Therefore, this dataset constitutes an important resource for evaluating the performance of motor imagery based brain computer interface systems in clinical rehabilitation applications.

Dataset 3 is a high quality motor imagery EEG dataset containing multiple session and multi day recordings. This dataset includes EEG recordings from 62 healthy participants and aims to analyze brain activity during motor imagery tasks. The dataset includes two different task configurations: two class (2C) and three class (3C) motor imagery paradigms. In this study, only the 2C dataset was used to limit the classification problem to a binary structure. In the 2C dataset, participants performed left hand grasping and right hand grasping motor imagery tasks.For each participant, the experiments consisted of three separate recording sessions conducted on different days.

In the dataset 3, each session contained five motor imagery blocks, and each block contained 40 trials. As a result of this structure, a total of 200 trials were obtained in each recording session, with an equal number of samples for both classes. EEG signals were recorded using a 64-channel EEG head positioned according to the international 10–20 electrode placement system and obtained at a sampling frequency of 1000 Hz.

Table 3.2: Comparison of the main characteristics of the EEG motor imagery datasets used in this study

Feature	**Dataset-1**	**Dataset-2**	**Dataset-3**
Subjects	60	50	62
Population	Healthy	Stroke patients	Healthy
MI Tasks / Classes	Left vs Right	Left vs Right	Left vs Right
Sampling Frequency	512 Hz	500 Hz	1000 Hz
EEG Channels	27	30	59
Reference Electrode	Left earlobe	CPz	Pz

3.3 Preprocessing and Feature Extraction

Preprocessing constitutes a critical stage in EEG based motor imagery classification, as the quality and consistency of the input signal directly influence the discriminative capacity of the subsequent classification models. Given the

inherent characteristics of EEG data including low signal to noise ratio, susceptibility to muscular and ocular artefacts, and high inter subject variability a carefully designed preprocessing pipeline is essential to ensure that the models receive reliable and informative representations of neural activity. In this study, each dataset was subjected to a tailored preprocessing and feature extraction pipeline reflecting both the specific recording conditions of that dataset and the distinct input requirements of the two classification models employed, namely a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM). While the CNN operates directly on normalized time series representations of the EEG signal and learns discriminative features through successive convolutional operations, the SVM requires handcrafted feature vectors derived from the preprocessed epochs. Accordingly, the preprocessing pipeline diverges at the feature extraction stage depending on the classifier, while the upstream signal conditioning steps including channel selection, filtering, and segmentation remain consistent across both models within each dataset.

The raw EEG signals in Dataset 1 were preprocessed through a standardized pipeline prior to model training. Five electroencephalographic channels were selected from the original 27 channel montage, namely C3, Cz, C4, P3, and Pz, which overlie the sensorimotor cortex and are well established as the most informative sites for motor imagery decoding. The continuous signals were first bandpass filtered between 4 and 38 Hz using a zero phase filter to preserve motor related frequency components, including the mu (8–12 Hz) and beta (13–30 Hz) rhythms, while attenuating slow drifts and high frequency noise. A notch filter was subsequently applied at 50 Hz and its harmonic at 100 Hz to suppress power line interference. The filtered signals were segmented into epochs of 1.0 second duration with a cue offset of 0.5 seconds, yielding fixed length trials of 256 samples per channel at a sampling frequency of 256 Hz. Each epoch was assigned a binary label corresponding to left hand or right hand motor imagery, and the resulting dataset comprised 41,760 trials each represented as a matrix of shape 256×5.

To ensure generalizability and prevent data leakage, a subject wise train–validation split was employed using `GroupShuffleSplit`, whereby all trials belonging to a given subject were assigned exclusively to either the training or the validation set, with approximately 80% of subjects allocated to training and 20% to validation. Class balance in the validation set was enforced by iteratively sampling splits until the positive class ratio fell within the range [0.40, 0.60].

Following the split, per subject z-score standardization was applied independently to each subject's training trials, computing the channel wise mean and standard deviation across that subject's data and normalizing accordingly. The same subject specific statistics were applied to normalize the corresponding validation trials, ensuring that no information from the validation set influenced the normalization parameters. For the CNN model, Gaussian noise with a standard deviation of 0.01 was additionally applied to the input during training as a data augmentation strategy to improve generalization. For the SVM classifier, spectral features were extracted from the preprocessed epochs using Welch's method to estimate the Power Spectral Density (PSD) of each channel, and the mean power within the 8–25 Hz frequency band was computed per channel, yielding a five dimensional feature vector per trial. These feature vectors were subsequently standardized using z-score normalization computed from the training set statistics.

The dataset 2 was obtained from a publicly available EEG motor imagery dataset recorded from acute stroke patients. The dataset comprises 2,000 trials recorded from 50 subjects across 33 EEG channels at a sampling frequency of 500 Hz. Five channels corresponding to C3, Cz, C4, P3, and Pz were selected from the original 33 channel montage, as these electrode positions overlie the sensorimotor cortex and are most relevant for motor imagery decoding. A Common Average Reference (CAR) was applied to each trial by subtracting the mean signal across the selected channels at each time point, thereby reducing spatially diffuse noise and common mode interference. A 4 second segment was extracted from each trial, and a fourth order zero phase Butterworth bandpass filter with cutoff frequencies of 8 and 30 Hz was subsequently applied along the time axis to isolate the mu and beta frequency bands associated with motor imagery related sensorimotor rhythms.

For the CNN model, the filtered signals were transposed to the format $C \times T$ (channels $\times$ time), yielding trial representations of shape $5 \times 2{,}000$. A subject wise train–validation split was performed using `GroupShuffleSplit` with 80% of subjects assigned to training and 20% to validation, ensuring that no subject's data appeared in both sets. Per subject z-score normalization was applied independently to each subject's trials, and Gaussian noise with a standard deviation of 0.01 was added to the input during training as a regularization strategy. For the SVM classifier, trial-wise z-score normalization was applied by subtracting the temporal mean and dividing by the standard deviation computed across the

time dimension of each individual trial. Then, common spatial patterns (CSP) with three components were applied to maximize the variance ratio between the two motor imagery classes, followed by log-variance computation to produce a compact, discriminative feature vector per trial. The extracted features were standardized using `StandardScaler` prior to classification, and model evaluation was conducted through a 5-fold stratified cross-validation scheme in subject wise applied independently within each subject's trials, with performance averaged across all folds and subjects.

The dataset 3 was sourced from the publicly available WBCIC_SHU Motor Imagery dataset, which provides preprocessed EEG recordings from multiple subjects performing left and right hand motor imagery tasks. The data were recorded at a sampling frequency of 1,000 Hz across 58 EEG channels, with each trial comprising 1,000 samples corresponding to a 1-second epoch. As the dataset was distributed in a preprocessed format, bandpass filtering and artefact removal had been applied by the data providers prior to release; no additional filtering was therefore performed in this study. Five channels were selected from the original 58 channel montage, namely C3, Cz, C4, P3, and Pz, corresponding to electrode positions overlying the sensorimotor cortex and well established as the most informative sites for motor imagery decoding.

For the CNN model, global z-score normalization was applied across all trials by computing the channel wise mean and standard deviation over the entire dataset and normalizing each channel accordingly, ensuring a consistent amplitude scale across subjects and sessions. Model evaluation was performed using a 10-fold stratified cross-validation scheme, in which the dataset was partitioned into ten folds with preserved class proportions, with nine folds used for training and one held out for evaluation in each iteration. For the SVM classifier, the preprocessed channel signals were directly concatenated into a fixed length feature vector per trial by flattening the $T \times C$ representation, yielding a feature vector of dimensionality $1{,}000 \times 5 = 5{,}000$. Within each cross-validation fold, the training feature matrix was standardized using `StandardScaler`, and the resulting statistics were applied to normalize the test fold, ensuring that no information from the test set influenced the normalization parameters. Classification was performed using a support vector machine with an RBF kernel, and model evaluation followed the same 10-fold stratified cross-validation scheme as employed for the CNN, with performance metrics averaged across all folds.

Across all three datasets, the preprocessing pipelines were designed with

two overarching principles in mind: the prevention of data leakage between training and evaluation sets, and the maximization of signal quality through physiologically motivated filtering and channel selection. The consistent selection of the C3, Cz, C4, P3, and Pz electrode subset across all experiments reflects the well established role of sensorimotor cortical regions in mediating motor imagery related oscillatory activity, and ensures a controlled and comparable basis for cross-dataset evaluation. The divergence in feature extraction strategies between the CNN and SVM models raw normalized time series versus handcrafted spectral or spatial features is a deliberate methodological choice intended to exploit the complementary strengths of deep learning and traditional machine learning approaches. The outputs of these pipelines serve as the direct inputs to the classification models described in the following section.

3.4 GPT Dataset Curator

In this thesis, the multi expert evaluation mechanism developed during manual review is designed not only for the classification of datasets but also to establish the conceptual foundation necessary for the implementation of an automated system. As a result of the collaborative evaluation processes conducted by the expert, the trained student, and researchers from two different disciplines during the manual review process, a comprehensive decision framework for classifying EEG datasets was created. This framework relies not only on individual expert opinions but also on collective standardization achieved across different experience levels and disciplinary perspectives. Therefore, the second phase, in which GPT is used, focuses not on replacing human expertise but, on the contrary, on transforming this expertise into a systematic set of rules and automating it.

Because EEG datasets are reported in highly heterogeneous formats in the literature, systematically extracting specific technical parameters and methodological information is both time consuming and error prone. Especially in large scale literature reviews, the reproducibility of human based classification processes may be reduced, and inter rater inconsistencies can arise due to subjective interpretation and variability among evaluators [103]. In this context, automated approaches based on large language models, such as GPT, have been proposed to improve consistency and scalability in text based classification tasks [74]. It should be noted that this model does not act as an independent decision

maker; rather, it functions as an automation tool that applies predefined rules or criteria derived from human expert input with high consistency and efficiency.

At the heart of GPT's work is the shared knowledge base established during the manual evaluation process. Training instructions on how to evaluate parameters such as criteria definitions, sampling rate, the terms of use of the "not found" label, channel placement, and protocol information, determined during panel discussions, were transferred step by step to a prompt.

In this process, the expert evaluator in the study ensured the scientific accuracy of the technical concepts. A trained student translated the expert explanations into clear and applicable rules. A researcher in the relevant field verified compliance with the reporting standards in the EEG literature. A researcher in the other field tested the understandability and generalizability of the criteria.

The combination of these four perspectives ensured that the instructions transmitted to GPT achieved scientific and practical standardization. Strict rules, such as relying solely on explicitly reported information, allowing no interpretation, and marking all missing information as "not found," were established by consensus among the team and formed the basis for GPT's work.

To ensure the model can produce consistent and reproducible results, the command prompt is designed in a multi layered structure.To ensure consistent and structured outputs from the GPT based classification process, a three layer prompting framework was designed. In the first layer, referred to as the instruction layer, the overall task definition and classification objectives were clearly specified. In the second layer, the schema layer, the structure of the 13 column classification table was explicitly defined, and the meaning as well as the expected format of each column were described in detail. In the final layer, the sample output layer, example rows with correctly formatted outputs were provided to guide the model in producing consistent and properly structured results. These examples illustrated both the expected content and formatting requirements. The promt created with some rules that the model must follow are clearly stated.These rules can be listed as follows:

- Only information directly provided in the article can be used.
- Estimation, interpretation, or inference is prohibited.
- Missing information for each column must be clearly written as "not found."
- The sampling frequency should be reported only numerically.

- Terms such as channel number, brand, and reference electrode information should be taken as they appear in the article.

Through this multi layered design, the decision making process defined by the human expert panel was systematically translated into a reproducible and structured algorithmic workflow.

GPT processes each article individually during the review process; it reads the text only once and populates the table after the first reading. This approach closely mimics the guiding principle of manual review. For each article, the model output was compared with the manual review. This process has two main objectives. The first is to measure GPT's level of compliance with the panel standard. The second is to assess the extent to which the panel's decisions are amenable to automation. Instances of non compliance typically arise from information embedded in figures or supplementary materials within the article, or from the text's failure to provide clear information.

The model does not function as an expert or an independent decision maker; rather, it operates as a consistent system for applying the rules defined by the research team. In this context, GPT can be considered as an automated mechanism for implementing standardized criteria within the evaluation framework. Due to the absence of human like intuition and domain specific reasoning, its performance may be limited in tasks requiring contextual interpretation. However, it demonstrates high reliability in applying clearly defined rules and extracting well specified technical parameters.

This work aligns with the approach known in the literature as human in the loop data curation, where human expertise is integrated into automated systems to guide, validate, and constrain model outputs [104, 105]. In this framework, GPT is incorporated as part of the loop but does not assume the final decision making role. Instead, it operates as a reliable automation tool that applies the boundaries and criteria defined by the human panel, ensuring consistency while maintaining human oversight.

This approach is compatible with the increasingly widespread use of human AI hybrid models in the literature. Numerous studies have shown that guiding AI models with rules defined by human experts significantly increases accuracy, particularly in automated structured knowledge extraction from biomedical texts [106].Jensen et al. [107] reported that explicitly defined rules significantly reduce the error rate of large language models in clinical text mining. Similarly, Moradi and Samwald [108] demonstrated that using human generated

templates in GPT based systems for extracting data from biomedical literature improves consistency. Furthermore, a comprehensive review by Chapman et al. [109] emphasizes that multi expert standardized criteria are the gold standard in NLP based knowledge extraction processes and directly impact model performance. Therefore, the method used in this study not only offers a practical solution but also provides a methodologically consistent framework with hybrid human AI knowledge extraction strategies proposed in the literature.

3.5 Classification Models

This study addresses the binary classification of left hand versus right hand motor imagery from EEG signals, a well established paradigm in brain–computer interface research [23]. Two fundamentally different classification strategies were evaluated across all three datasets: a traditional machine learning approach based on Support Vector Machines (SVM) and a deep learning approach based on Convolutional Neural Networks (CNN). Applying both approaches to the same datasets under controlled conditions enables a systematic and rigorous comparison of classical and modern classification paradigms for EEG based motor imagery decoding. A key contribution of this work lies in the evaluation of these pipelines across three heterogeneous datasets collected under different experimental protocols, with different devices, different subject populations.

Support Vector Machine

Support Vector Machines are a well established supervised learning method widely applied in EEG based brain–computer interface research owing to their strong generalization performance in high dimensional feature spaces and their robustness to overfitting when the number of training samples is limited relative to the feature dimensionality [77, 110]. The SVM algorithm seeks to identify an optimal separating hyperplane in the feature space by maximizing the margin between classes, with the decision boundary defined solely by the support vectors the subset of training samples lying closest to the boundary. In this study, a Radial Basis Function (RBF) kernel was employed for all SVM classifiers, as commonly adopted in the literature [111]. The RBF kernel was selected because EEG derived feature spaces are inherently nonlinear, and the RBF kernel implicitly maps the input features into a high dimensional space in which linear separation

becomes feasible, effectively modelling complex nonlinear decision boundaries without requiring explicit feature engineering [77]. Here, C consistently denotes the number of EEG channels.The kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{3.1}$$

where γ controls the width of the kernel and was set to `scale` (i.e., $\gamma = 1/(n_{\text{features}} \cdot \text{Var}(\mathbf{X}))$), and the regularization parameter was set to $C = 1.0$ for Datasets 1 and 3, and $C = 10.0$ for Dataset 2. Class imbalance was addressed in Dataset 1 by setting `class_weight = balanced`, which adjusts the cost of misclassification inversely proportional to class frequencies. For Dataset 2, the SVM was integrated into a pipeline combining Common Spatial Pattern (CSP) feature extraction [76] with `StandardScaler` normalization prior to classification.

Convolutional Neural Network

In the deep learning approach, automatic feature extraction and classification were performed jointly using a Convolutional Neural Network architecture based on EEGNet, a compact and computationally efficient model specifically designed for EEGbased brain–computer interface applications [29]. Unlike traditional machine learning approaches that rely on handcrafted features, CNNs learn hierarchical representations directly from the raw or minimally processed EEG signal, capturing both temporal and spatial patterns through successive convolutional operations [112].

The EEGNet architecture, illustrated in Figure 3.4 and detailed in Table 3.3, processes the EEG input through three distinct convolutional stages. In the first stage, a standard `Conv2D` layer with 8 filters and a kernel size of (1×128) is applied along the temporal dimension to learn frequency specific features from the EEG time series, operating analogously to a set of bandpass filters [29]. This is followed by Batch Normalization to stabilize the activations. In the second stage, a `DepthwiseConv2D` layer with a kernel of $(C \times 1)$ – where C denotes the number of EEG channels – is applied to learn spatial filters across electrodes, a procedure functionally equivalent to traditional spatial filtering methods such as Common Spatial Patterns [76]. This depthwise operation uses a depth multiplier of 2, yielding 16 feature maps. Batch Normalization, ELU activation, Average Pooling (1×4), and Dropout $(p = 0.3)$ follow to introduce nonlinearity, reduce temporal resolution, and regularize the representations. In the third stage, a

`SeparableConv2D` layer with 16 filters and kernel size (1×32) is applied to capture temporal dependencies at a reduced scale while minimizing the number of trainable parameters [29]. A second Average Pooling (1×8) and Dropout $(p = 0.3)$ further compress the representation. The resulting feature map is then flattened and passed through a fully connected Dense layer with 64 units and ELU activation, followed by a single output neuron with sigmoid activation for binary classification. L2 regularization (weight decay $\lambda = 10^{-4}$) was applied to all convolutional and dense layers throughout the network to penalize large weight values and reduce overfitting. Additionally, a `GaussianNoise` layer with standard deviation $\sigma = 0.01$ was prepended to the input as a data augmentation strategy during training.

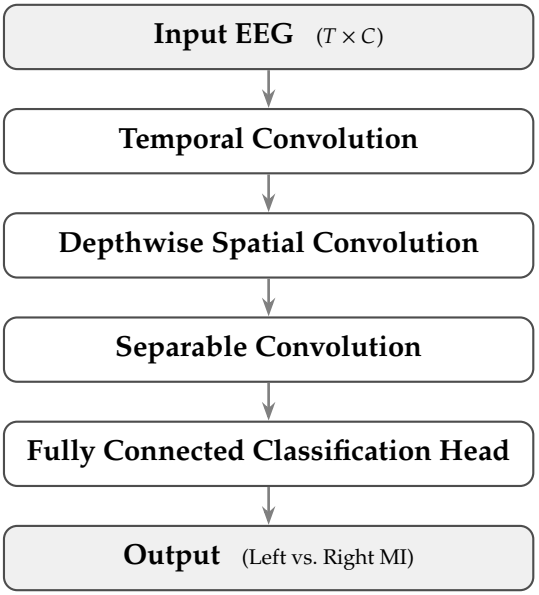

Figure 3.4: General pipeline of the CNN architecture employed in this study. Adapted from [29].

Training Procedure and Hyperparameter Configuration

All CNN models were trained using the Adam optimizer [113] with an initial learning rate of $\eta = 3 \times 10^{-4}$ and binary cross entropy as the loss function, given the binary nature of the left versus right motor imagery classification task. Training was conducted for a maximum of 120 epochs with a mini batch size of 64 for Dataset 1 and Dataset 3, and 32 for Dataset 2. To prevent overfitting and avoid unnecessarily long training processes, two adaptive training strategies were applied. First, Early Stopping monitored the validation loss and terminated

Table 3.3: Layer-wise architectural details of the CNN model. T denotes the number of time points and C the number of EEG channels in the input.

Layer	Type	Parameters	Output shape
1	Input	–	(T, C)
2	GaussianNoise	$\sigma = 0.01$	(T, C)
3	Conv2D	8 filters, (1×128), same	$(C, T, 8)$
4	BatchNormalization	–	$(C, T, 8)$
5	DepthwiseConv2D	$(C \times 1)$, depth×2	$(1, T, 16)$
6	BatchNormalization	–	$(1, T, 16)$
7	ELU Activation	–	$(1, T, 16)$
8	AveragePooling2D	(1×4)	$(1, T/4, 16)$
9	Dropout	$p = 0.3$	$(1, T/4, 16)$
10	SeparableConv2D	16 filters, (1×32), same	$(1, T/4, 16)$
11	BatchNormalization	–	$(1, T/4, 16)$
12	ELU Activation	–	$(1, T/4, 16)$
13	AveragePooling2D	(1×8)	$(1, T/32, 16)$
14	Dropout	$p = 0.3$	$(1, T/32, 16)$
15	Flatten	–	$(T/32 \times 16)$
16	Dense	64 units, ELU	(64)
17	Dense (output)	1 unit, sigmoid	(1)

training if no improvement was observed for 15 consecutive epochs (`patience` = 15), restoring the model weights corresponding to the best validation loss at termination. Second, a Learning Rate Reduction on Plateau strategy halved the learning rate (factor = 0.5) whenever the validation loss failed to improve for 6 consecutive epochs (`patience` = 6), with a minimum learning rate floor of $\eta_{\min} = 10^{-6}$. Additionally, a `ModelCheckpoint` callback was used to save the best performing model weights throughout training based on the validation loss, ensuring that the final evaluation was always performed using the weights associated with the lowest validation loss rather than the last training epoch. The complete set of hyperparameters used for training across all datasets is summarized in Table 3.4.

Evaluation Metrics

Model performance was assessed using a comprehensive set of classification metrics to enable a multi-perspective evaluation of each model's discriminative

Table 3.4: Hyperparameter configuration for CNN training across all datasets.

Hyperparameter	**Value**	**Notes**
Optimizer	Adam [113]	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Initial learning rate	3×10^{-4}	–
Loss function	Binary cross-entropy	Binary classification
Max epochs	120	–
Batch size	64 (DS1, DS3), 32 (DS2)	–
Early stopping patience	15 epochs	Monitor: val_loss
LR reduction factor	0.5	Patience: 6 epochs
Min learning rate	10^{-6}	–
L2 regularization	$\lambda = 10^{-4}$	All conv + dense layers
Dropout rate	$p = 0.3$	After each pooling
Gaussian noise	$\sigma = 0.01$	Input augmentation
Random seed	42	Reproducibility

capacity. The primary metric was classification accuracy, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Class-level performance was further characterized using precision, recall, and F1-score:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

The area under the Receiver Operating Characteristic curve (AUC-ROC) was computed to evaluate discriminative performance independently of the classification threshold [114]. The Average Precision (AP) derived from the Precision–Recall curve was also reported, which is particularly informative under class imbalance [115]. Confusion matrices were generated for all experimental conditions to provide a complete breakdown of classification outcomes across both classes.

Cross-Validation and Data Separation Strategy

Given the well documented risk of overly optimistic performance estimates when data from the same subject appear in both training and evaluation sets [116], a subject based data separation strategy was employed for Datasets 1 and 2.

Specifically, all trials for a given subject were assigned exclusively to the training set or the validation set, with no subject appearing in both partitions. This subject independent evaluation scheme implemented via `GroupShuffleSplit` with an 80/20 subject ratio provides a conservative and realistic estimate of the model's generalizability to previously unseen individuals.

For Dataset 3, which does not provide subject level metadata in a format compatible with subject wise splitting, a 10-fold stratified cross validation scheme was applied instead. In this approach, the data set was partitioned into ten folds with preserved class proportions; in each iteration, nine folds were used for training and one was held out for evaluation, with performance metrics averaged across all folds. The same 10-fold cross-validation scheme was applied to both the CNN and SVM classifiers on Dataset 3 to ensure comparability. For SVM of Dataset 2, a subject wise 5-fold stratified cross-validation was applied within each subject's trials independently, and the results were averaged across subjects, providing a generalization estimate within subjects that accounts for individual variability.

4

Results and Discussion

This section offers a systematic presentation of the results, accompanied by an analytical interpretation of the findings obtained during the experimental phases. The analyzes encompass experimental results based on manual evaluation, automated GPT based classification, and machine learning models performed on open source EEG datasets. The results are presented by comparing the outcomes across different datasets and methods.

The presented results are intended not only to quantitatively report classification performances, but also to examine the structural characteristics of the datasets used and the effects of evaluation approaches on these performances. In this context, three key dimensions were systematically investigated: evaluator based systematic review processes, the effectiveness and reliability of GPT based labeling, and the comparative performance of machine learning models across heterogeneous EEG datasets.

The results are presented in a way that highlights the performance differences, strengths, and limitations of each method across different datasets. The results of manual evaluation provide findings that aim to understand the consistency but also subjectivity of human based decision making processes; The GPT based evaluation results demonstrate the extent to which automated approaches agree with human evaluations. Findings from machine learning experiments allow for the evaluation of data set heterogeneity and the effects of labeling strategies on model performance.

The experimental findings presented in this section provide the foundation for subsequent comparative analyses of machine learning model performance

and GPT based labeling effectiveness across different EEG datasets.Thus, the results are not limited to numerical performance metrics but also provide methodological and practical insights.

4.1 Systematic Review Results

In the first phase of this study, EEG datasets were evaluated independently by participants representing four different levels of expertise. The results indicate that the proposed classification criteria can be applied with a certain level of consistency, while also revealing variations across evaluators. This observation is consistent with previous findings on inter rater reliability and EEG evaluation processes, which highlight the influence of evaluator expertise on assessment outcomes [117].

Figure 4.1 presents only a few examples of the articles reviewed in this study. This figure shows a single review of the article performed by evaluators at different levels of expertise, including an Expert Reviewer (Giulia Cisotto), an Instructed Senior Reviewer (Daniela DAuria), an Instructed Junior Reviewer (Betül Sena Petek), and a Naive Reviewer (Davide Chicco). In the Level 0 criteria (e.g., sampling frequency, number of channels, number of participants, and protocol structure), the evaluations showed only minor differences between raters. This is likely due to the fact that this information is typically reported explicitly and consistently in the original studies. Consequently, the assessments were generally consistent for these criteria. In later stages of the evaluation process, a more consistent interpretation of the criteria was observed. In general, similar evaluation patterns were observed in most cases, although these observations should be interpreted with caution.

In contrast, greater variability was observed in the interpretation of the Level 1 and Level 2 criteria between evaluators. This variability was mainly related to how specific technical aspects were identified and interpreted, such as frequency-domain analyzes, montage details, ground electrode information, and device specifications. In several cases, relevant information was not explicitly reported in the text and had to be inferred from figures or supplementary materials, requiring additional interpretation.

Annotator	name	Dataset	Level 0					Level 1		Level 2			
			Units (opt)	sampling frequency [Hz]	Montage	Experimental protocol (planned, a priori, intentions; timing is important)	Labels (given or annotated afterwards, interpretation; timing is important)	time-domain	frequency-domain	output format	download	brand	ground channel
Naive Reviewer	***template*	DOI URL	µV or quantization levels?	number	Number of EEG channels: XXX Channels location: XXX Reference: XXX Details: XXX. Other non-EEG data: XXX	Sessions: XXX Blocks: XXX Task: XXX Rest periods: XXX Timing: XXX Details: see Section "XXX"	Protocol-based or annotations?	visualizations: XXX data quality: XXX	to fill	to fill	platform and URL	to fill	to fill
Expert Reviewer	BCI Database	https://www.nature.com/articles/s41597-023-02445-z	??	512 Hz	27 active scalp electrodes (10–20 system), referenced to the left earlobe.	Eight different experimenters conducted the experiments, Eighty-seven (87) participants completed one of the two BCI experiments based on the same protocol.	triggers (a.k.a events) associated to each cue and phase of the BCI trials and runs (see section "Usage NOts")	time-freq plots (fig.10)	time-freq plots (fig.10)	format that can be read with EEGLAB (Matlab-based) or MNE (Python-based)	directly from Zenodo	g.USBAmp amplifiers, recorded without any hardware filters	Fpz
Instructed Junior Reviewer	exoskeleton	https://www.nature.com/articles/s41597-023-02243-7	µV	200	Number of EEG channels: 32 Channels location: 10-10 system Reference: A1(left ear lobe) Details: Figure 1 Other non-EEG data: EOG	Participants: 50 Sessions: 16 Blocks: 3 Task: 15 trials while the subject wears the exoskeleton and performs diferent mental tasks Rest periods: 4 minutes Timing: Figure 3 Details: see Section "Protocols"	Protocol based	visualizations: Figure 4 and Figure 10 data quality: Artifact Subspace Reconstruction (ASR) or Independent Component Analysis (ICA)	Gamma band activity analysis	.csv, .mat, json,.txt	Figshare https://github.com/bmislab/DECODED.	Brain Products actiCAP and actiCHamp amplifier	A2 (right earlobe)
Naive Reviewer	Inner Speech Dataset	https://doi.org/10.1038/s41597-022-01147-z	µV (Figure 5)	1024	Number of EEG channels: 128. Channels location: standard. Reference: in the lef and right lobe of each ear. Details: Figure 9. Other non-EEG data: none.	Participants: 10. Sessions: 3. Blocks: 7 (Figure 2). Task: four mental tasks in three diferent conditions: inner speech, pronounced speech and visualized condition. Rest periods: . A rest interval, with a variable duration of between 1.5 seconds and 2 seconds, was given between trials. Timing: Figure 3. Details: see Section "Exerimental procedures".	protocol based	time-course (Figures 5, 6), time-frequency (Fig.7), time-frequency representation section	spectra (Figure 8), time-frequency (Figure 7), time-frequency representation section	.bdf EEGLAB Matlab	yes, on OpenNeuro.org	BioSemi ActiveTwo	not found
Instructed Senior Reviewer	Tinnitus	https://www.nature.com/articles/s41597-022-01622-w	not found	256	Number of EEG channels: 16 Channels location: 10-20 International Standard System Reference: Cz Other non-EEG data: THI and HADS questionnaires Details: Figure 3a	Participants: 71 Sessions: 4 Task: resting-state (few minutes x2), auditory therapy (1), passive mode(4), active mode(2) Timing: Figure 4 Details: see Section "Experimental procedure"	Protocol-based	visualizations: not found data quality: not found	not found	.mat, .xls, gdf	Mendeley (free download)	gTech *g.USBAmp*	A1

Figure 4.1: Example of structured dataset annotation sheet used in the manual review process

These challenges highlighted the need for clearer and more standardized reporting practices in EEG studies. Throughout the evaluation process, interactions between evaluators enabled the transfer of domain knowledge from more experienced experts to less experienced reviewers, as well as the GPT based labeling process. This knowledge transfer supported a more informed and consistent interpretation of complex or ambiguously reported criteria.

As a result, shared guidelines were progressively developed for handling unclear or missing information, including the use of the "not found" label and the treatment of borderline cases. This process contributed to a more consistent application of the evaluation criteria, without implying a formal measurement of agreement.

The input from the expert rater appeared to facilitate a clearer definition of the evaluation criteria. Over time, these clarifications were incorporated by all evaluators, supporting a more consistent application of the criteria throughout the evaluation process.

The analysis conducted in this study highlighted heterogeneity in the reporting of EEG datasets. While some criteria were consistently reported across studies (e.g., sampling frequency and number of channels), other elements, such as protocol details and specific technical configurations, were often incomplete or not explicitly reported, requiring additional interpretation during the evaluation.

This finding is consistent with the long standing issue of reporting standards in the EEG literature. Pernet et al. [47] demonstrated that a significant portion of EEG studies report incomplete or inconsistent descriptions of electrode placement, referencing, and data processing steps. Similarly, Robbins et al. [96] reported that EEG datasets are heterogeneous in terms of shared formats and that metadata elements, in particular, are far from standardized in most studies. Previous work has highlighted that inconsistencies in EEG preprocessing pipelines can affect the interpretability and reproducibility of results [118].Gramfort et al. [119] emphasized that complete reporting of experimental protocols, channel configurations, and preprocessing steps is critical for the reliable reuse of open EEG databases. The deficiencies observed in the panel assessment are in line with findings reported in the literature, suggesting that reporting issues in EEG studies may be relatively widespread.

Basic parameters such as sampling frequency and number of channels are generally clearly stated; however, channel locations, reference, and ground elec-

trode information are missing or only presented in the figures in most studies. Although the number of participants is included in almost all articles, the structure of the session, block, rest periods, and task times are often unclear.Although time-domain signals are visualized in figures by most studies, frequency-domain analyzes, advanced processing methods (waveform, spectral power analysis, etc.), and artifact processing strategies are only reported in detail in a limited number of articles.Significant differences were also observed between the data formats. Some studies used standard formats (EDF, BDF, MAT), while others presented proprietary data structures or restricted access files. Information about the brand and model of the device is often reported.These findings highlight the lack of standardized reporting guidelines for EEG datasets in the literature. The openness of some criteria to interpretation during the evaluation process created inter rater variability, confirming the context sensitive nature of manual evaluation. The successful management of this process by the human panel provided a very strong reference base for training the GPT based model.

The findings of this phase suggest that manual review conducted by a multi expert human panel may contribute positively to the accuracy of EEG dataset assessments and may support the establishment of methodological consistency. The involvement of reviewers with varying levels of experience appears to benefit from the diversity among researchers reported in the literature, potentially enabling the interpretation of incomplete, ambiguous, or implicitly presented information through expert informed judgment. The resolution of inconsistencies through intra team discussions may have facilitated the clarification of evaluation criteria and contributed to a more structured and reproducible review process. Overall, the manual systematic review can be considered not only as an initial step in the study, but also as an important methodological component that may support data quality and inform the development of automated analysis systems. These observations may provide a basis for further exploration of humanmachine interaction, criteria standardization, and hybrid analysis approaches, which are discussed in the following sections.

4.2 GPT Performance

In the second phase of this study, the common rules and standardization principles established during the manual evaluation process were transferred to the GPT based model and the extent to which the model could accurately, con-

sistently and systematically evaluate the same datasets was examined. The GPT evaluation was designed as an automated implementation of the framework developed by the human panel; therefore, the model was required to extract only the information explicitly stated in the text, to avoid interpreting ambiguous statements, and to explicitly label missing information as "not found." The results demonstrated that GPT produced remarkably systematic, fast and format consistent output, but it had limitations in areas requiring contextual inference.

Table 4.1: Error Analysis of GPT on EEG Dataset Articles

Dataset Name	**Total Parameters**	**Missing**	**Incorrect**	**Error Rate (%)**
ChineseEEG	11	1	0	9.1
EEG-ECG Dataset	11	7	0	63.6
LPPHK	11	1	0	9.1
BMI-HDEEG	11	0	7	63.6
PEARL	11	1	1	18.2
Sustained-Attention Driving	11	7	0	63.6
Voice-User	11	1	1	18.2
Mind-Body-Brain	11	2	1	27.3
Exoskeleton	11	0	2	18.2
DS000117	11	8	0	72.7

The dataset level error analysis as Table 1 reveals variability in the models performance across different EEG dataset articles. While some datasets such as ChineseEEG and LPPHK exhibit relatively low error rates, others, including DS000117 and the EEG-ECG Dataset, show substantially higher error rates. This variation suggests that the models performance is influenced by the structure and clarity of the source articles. In particular, datasets with more complex or less standardized reporting tend to result in higher numbers of missing or incorrect fields. Overall, the findings indicate that the model is capable of extracting information with reasonable accuracy in some cases, but its performance is not consistent across all datasets.

The model achieves high accuracy for certain parameters, such as labels, indicating that clearly defined and standardized information is more reliably extracted. In contrast, lower accuracy is observed for parameters such as output format and experimental protocol, where information is often less structured or more context dependent. Intermediate performance is seen in parameters such as montage, time domain, and frequency domain, suggesting partial success in capturing more technical details. These results highlight that the models effectiveness varies depending on the nature and presentation of the parameter

within the source articles.

The tabular outputs generated by the model were compared to generated by the manual panel, and performance was evaluated in three main dimensions. The first of these was explicitly providing numerical and conceptual information in the text. GPT demonstrated high accuracy in this class. GPT produced highly consistent and reproducible results for the following parameters:

- Sampling frequency (Hz)
- Brand and model
- Labeling method (protocol-based / task-based)

The commonality of these parameters is that they are often reported in the literature as single and unambiguous statements. GPT extracted information with high accuracy from such statements without requiring any interpretation; the results remained identical when the same prompt was repeated.

The technical parameters, such as the sampling frequency, were detected almost accurately by the model. For example, even indirect expressions like "sampled at 500 Hz" and "resampled to 500 Hz" were accurately parsed as the value "500." This demonstrates GPT's strong discrimination capacity for numbers and technical terms.

Another class of information required contextual interpretation. GPT provided limited inference for these parameters. The model exhibited significant limitations in information that was not directly included in the text of the articles or presented only in figures. These include:

- Output format
- Experimental Protocol
- Montage
- Time Domain
- Frequency Domain

Because this information was generally presented only in figures or embedded in the text, GPT's reach was limited, and unlike the multi expert panel, the model did not infer context. In particular, the model's performance was significantly lower than that of the human panel in technical aspects such as channel layout information.

The final class is ambiguous or incomplete information.The most incomplete information related experimental protocol such as participants, sessions, blocks,

task, rest period and timing information. GPT demonstrated high consistency across these parameters. The model consistently applied the "not found" label to any information not explicitly found in the text.This helped eliminate the tendency to "complete by interpretation" sometimes seen in manual evaluation and produced a more objective database.

All these data suggest that GPT is strong in explicit text based information but limited in areas requiring contextual inference.The contribution of GPT to the second stage is not limited to the precision of the information extraction. All tables filled by the model:

- followed the same column order,
- used the same verbal expressions,
- uniformly marked all missing information,
- contained no variation.

This level of structural consistency is a highly valuable feature in large scale literature reviews.

Although manually evaluating a document takes minutes on average, GPT completed the same task in 35 seconds.This speed difference has allowed larger databases for EEG research to be examined with large volumes of literature.In addition, comprehensive quality screenings to be performed in a much shorter time.Thanks to GPT,radically reducing the workload of the human panel.

Although the human panel could partially estimate missing information through contextual interpretation in some cases, GPT strictly retrieved only the explicit information in the text. This approach increased the objectivity of reporting and made critical omissions more visible. GPT contributed to the widespread heterogeneity in data quality.

Overall, the results suggest that the GPT based analysis system can provide consistent and structured outputs in the analysis of EEG datasets, particularly for parameters that are explicitly defined and text based. The model appears to facilitate the data extraction process from the literature, reducing the time required for manual review, while following predefined rules that may help limit certain types of extraction inconsistencies. However, its performance was comparatively lower in cases requiring interpretative judgment, such as information presented in visual materials, complex experimental protocols, or ambiguously reported technical details. These observations indicate that the effectiveness of automated extraction is influenced by how information is presented within

the source articles. Taken together, the findings suggest that GPT may function more effectively as a supportive tool that enhances efficiency, rather than as a standalone assessment mechanism. In this context, integrating rule based structures derived from human review into the GPT framework may contribute to improving efficiency in large scale literature analysis, while potentially supporting methodological consistency. More broadly, these results indicate that positioning automated systems as complementary tools alongside human expertise may be a practical approach in EEG dataset evaluations, particularly in contexts where both efficiency and interpretative flexibility are required.

4.3 Cross-dataset EEG Statistical Analysis

In this study, the amplitude distributions of three different EEG datasets were analyzed descriptively, and the fundamental differences between the datasets were comparatively evaluated. The analysis involved examining the empirical distributions of EEG signal samples obtained from each dataset through histogram representations. In addition, a reference normal distribution curve was overlaid on each histogram to facilitate a visual comparison with an idealized distribution pattern. Basic descriptive statistical measures, namely the mean and standard deviation values, were also calculated to summarize the central tendency and variability of the data.

Figure 4.2 shows the amplitude distribution of EEG signals from Dataset 1 . Examining the figure, it is observed that the signal values are largely concentrated around zero microvolts (0 μV) and the distribution generally exhibits a symmetrical structure. An inspection of the resulting histogram indicates that the amplitude values exhibit an approximately bell shaped distribution. However, this is only a visual assessment and does not represent any claim to statistical modeling. In this context, the normal distribution curve shown in the graph is used only for reference purposes and provides a visual comparison of the signal distribution with an idealized structure.

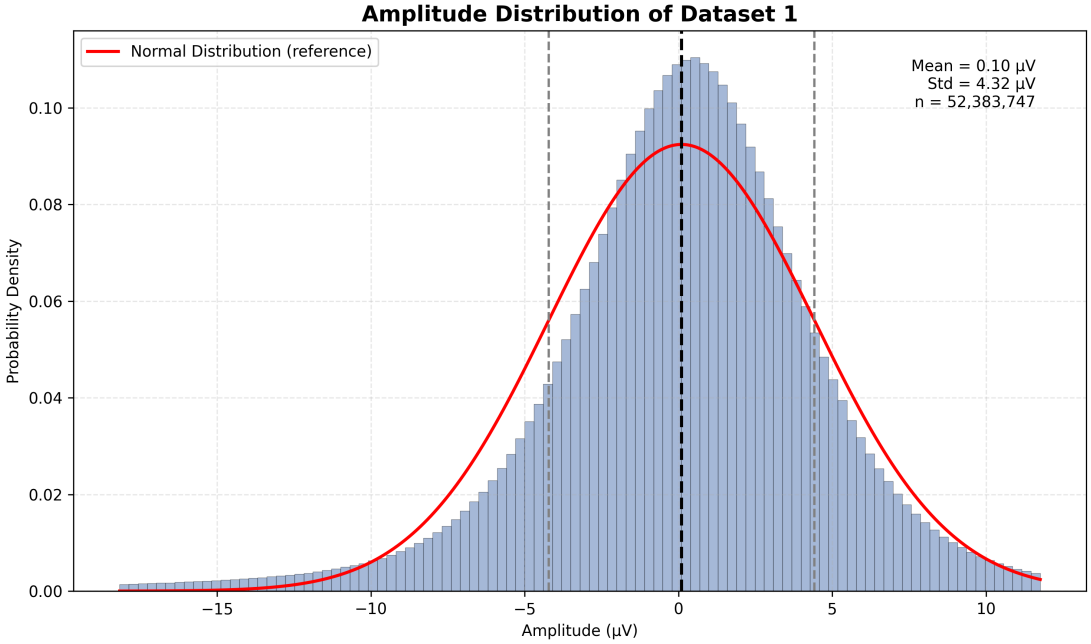

Figure 4.2: Amplitude Distribution of Dataset 1

The fact that the calculated average value is approximately close to zero (Mean 0.10 μV) indicates that the signal is balanced in terms of the DC component and does not contain a significant offset. The relatively low standard deviation (Std 4.32 μV) reveals that the signal variance is limited and the amplitude values are concentrated in a narrow range. This indicates that the dataset has a low noise level and high signal quality.

The low density values observed at the extremes of the distribution may be related to rare amplitude deviations or artifact related components frequently encountered in EEG signals. However, thanks to the percentage-based filtering approach applied to reduce the impact of extreme values in data visualization, the basic characteristics of the distribution have been revealed more reliably.

In conclusion, the amplitude distribution of Dataset 1 exhibits a balanced, low variance, and approximately symmetrical structure, indicating that the dataset contains good quality and stable EEG recordings. The findings suggest that this dataset provides a suitable basis for further analysis and modeling; however, these inferences are limited to observational distribution analysis and may require further statistical validation.

Figure 4.3 shows the statistical analysis of the amplitude distribution of EEG signals from Dataset 2. When the amplitude distribution of the EEG signals from Dataset 2 is examined, it is observed that the histogram generally exhibits

a bell shaped form, but the distribution is not symmetrical around zero. The calculated average amplitude value of approximately +8.4 μV indicates that the distribution is not centered on zero and that the amplitude values are shifted in the positive direction.

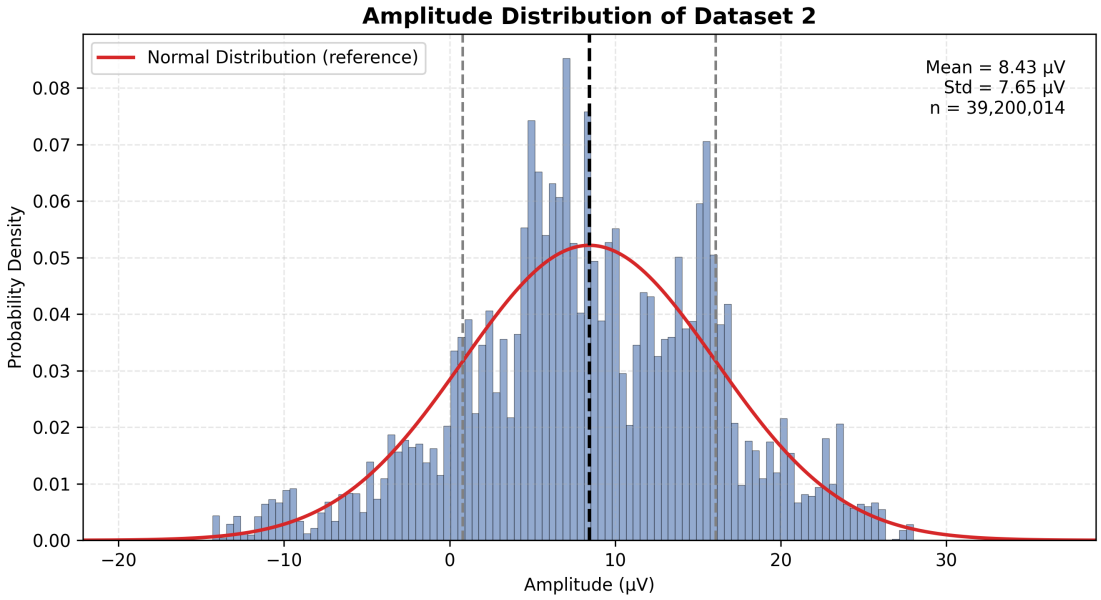

Figure 4.3: Amplitude Distribution of Dataset 2

When the spread of the distribution is examined, it is observed that the standard deviation value (7.65 μV) is higher compared to Dataset 1. This indicates that the amplitude values are distributed over a wider range in the dataset and that the variation is greater. The histogram also reveals that the distribution deviates to a certain extent from the ideal normal distribution and exhibits a longer tail structure, especially in the positive amplitude range.

The analysis performed in this section focuses on examining the empirical distribution of the amplitude values of EEG signals. Accordingly, the basic characteristics of the distribution were evaluated through histogram representation and measures of central tendency and dispersion. The findings provide a descriptive framework regarding the general structure of the distribution.

Figure 4.4 shows the statistical analysis of the amplitude distribution of EEG signals from Dataset 3. When the amplitude distribution of the EEG signals from Dataset 3 is examined, it is observed that the histogram exhibits a distinctly bell shaped and symmetrical structure. The calculated average amplitude value of approximately 0.06 μV indicates that the distribution is centered around zero and

that the signal amplitudes are evenly distributed in both positive and negative directions. This reveals that the dataset does not contain a significant centroid in terms of amplitude distribution.

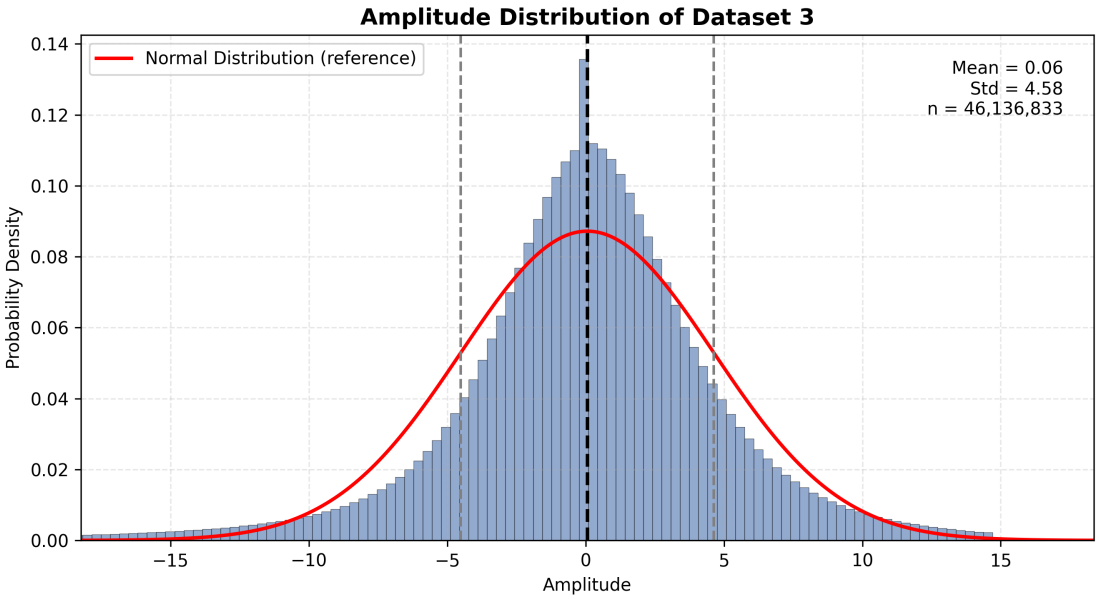

Figure 4.4: Amplitude Distribution of Dataset 3

When the spread of the distribution is evaluated, the standard deviation is found to be approximately 4.58 μV. This value indicates that the amplitude variations are concentrated within a certain range and that the dataset exhibits a relatively low variance structure. The unimodal structure and highly symmetrical appearance of the histogram reveal that the distribution exhibits characteristics closer to an ideal normal distribution compared to other datasets. The agreement between the reference normal distribution curve and the histogram supports this observation.

Furthermore, when the tail regions of the distribution are examined, it is seen that the outliers are limited and the amplitude values are largely concentrated around the center. This indicates that the dataset presents a more regular and balanced structure in terms of amplitude distribution. The findings reveal that the amplitude distribution characteristics of Dataset 3 exhibit a more stable character compared to datasets with wider variation and significant asymmetry.

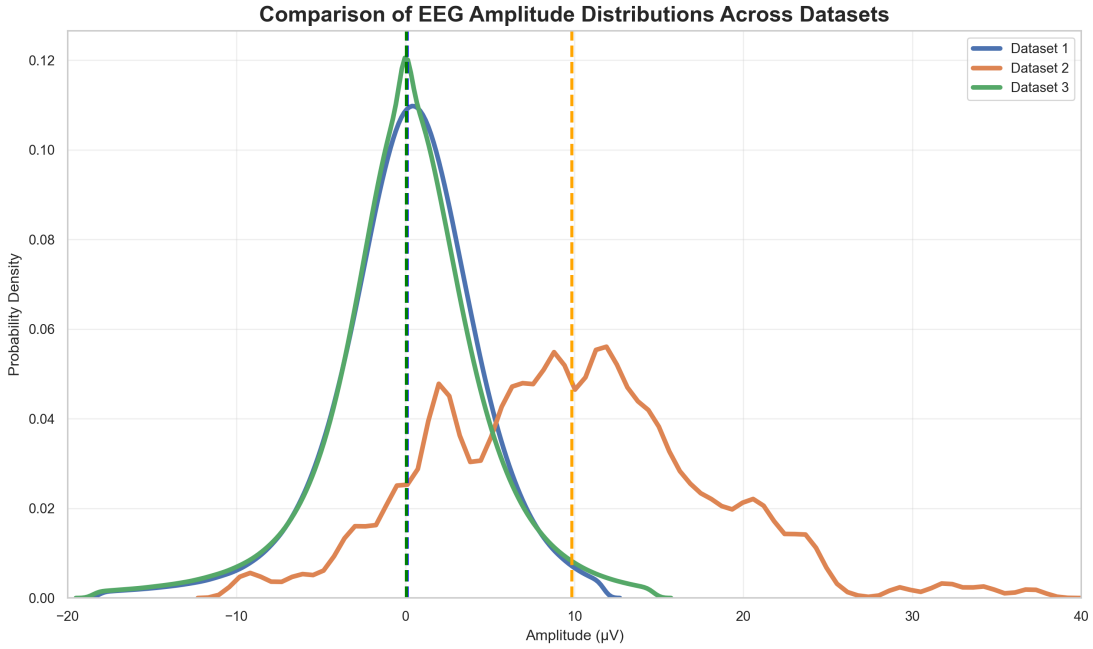

Figure 4.5: Amplitude Distribution Graph of All Dataset

When the results are compared, it is observed that data sets 1 and 3 exhibit similar statistical characteristics in their amplitude distributions. In both datasets, the mean amplitude is very close to zero and the amplitude distribution shows approximately normal characteristics. Furthermore, the very close standard deviation values indicate that the signal amplitudes in these datasets are distributed within similar variation ranges. This shows that both datasets reflect the typical amplitude characteristics of physiological EEG signals and that the signals are distributed in a balanced manner relative to the reference level. However, data set 2 shows significantly different statistical distribution characteristics compared to data set 1 and 2. The positive shift in the mean amplitude and the higher standard deviation compared to the other datasets indicate that the signal amplitudes have a wider variation range. When the histogram distribution is examined, it is observed that the amplitude values are distributed over a wider range, and the distribution exhibits a slightly right skewed character.A significant reason for the differing statistical characteristics of Dataset 2 is that it was obtained from patients with acute stroke. Structural and functional changes in brain tissue after a stroke can affect the regulation of cortical activity and neuronal synchronization. Such neurophysiological changes can lead to statistically significant differences in the amplitude distribution of EEG signals, such as higher variance, wider amplitude range, and shifts in the mean ampli-

tude level. Furthermore, the activation of alternative neural networks or the occurrence of compensatory cortical activations during motor imagery tasks in stroke patients can also affect the distribution characteristics of the EEG signal amplitudes.

These differences between datasets highlight the importance of the characteristics of the data sets in EEG data analysis and brain computer interface (BCI) applications. EEG data from different populations can differ not only in terms of experimental conditions but also in terms of the statistical characteristics of the signal. This demonstrates the need to consider the characteristic features of datasets when developing EEG based machine learning models. These findings indicate that the statistical properties of EEG signals can be influenced not only by technical recording conditions but also by the neurophysiological characteristics of the participant population to which the data set belongs. Therefore, considering the population characteristics of the dataset in EEG analyzes performed on different datasets is considered an important factor for the correct interpretation of the results obtained.

4.4 Intra and Cross-dataset Classification

This section presents a comparative performance analysis of the classification models used in this study on different EEG datasets.The performance of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) based models, used to classify motor imaging EEG signals, was analyzed using accuracy, confusion matrix, ROC curves, and other evaluation metrics. The results obtained were examined separately for each dataset, and visual and numerical evaluation outputs were presented together to provide a more detailed understanding of the model behavior in different data structures.

4.4.1 Results on Dataset 1

In this part, the results of classification experiments conducted on Dataset 1 were represented. Both SVM nd CNN based models were trained on this dataset, which contains EEG motor imagery data from healthy participants, and the resulting performance was comparatively evaluated. To analyze model performance more comprehensively, accuracy values, confusion matrix results, and

ROC and Precision-Recall curves were examined, thus evaluating the classification success of the models through different metrics.

SVM Results of Dataset 1

The SVM model implemented on Dataset 1 demonstrated moderate performance in the motor image classification task. The general accuracy of the model was calculated to be approximately 55.7%. This result indicates that the model has a certain discrimination capacity compared to random classification, but the distinction between classes is quite limited.

Examining the matrix given in Figure 4.6, it is seen that SVM classifier trained on Dataset 1 yielded an overall accuracy of approximately 55.7%. With 2204 true left and 2478 true right predictions correct, the model shows a slight tendency toward predicting the right class more accurately. However, the high misclassification counts 1996 left samples predicted as right and 1722 right samples predicted as left.

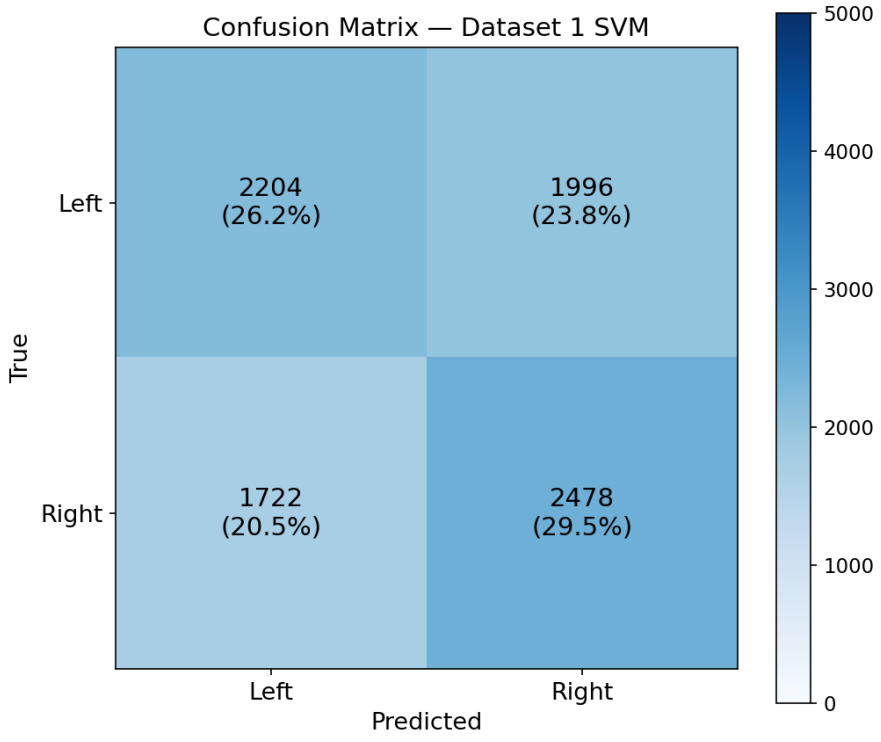

Figure 4.6: Confusion Matrix results for Dataset 1 for SVM

This indicates that the model is slightly more successful in detecting right hand motor imagery samples compared to the left hand class.

Table 4.2: Classification Report of the SVM Model for Dataset 1

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.5614	0.5248	0.5425	4200
Right	0.5539	0.5900	0.5714	4200
Accuracy			0.5574	8400
Macro Avg	0.5576	0.5574	0.5569	8400
Weighted Avg	0.5576	0.5574	0.5569	8400

ROC curve analysis provides additional information on the discrimination capacity of the model . The area under the ROC curve (AUC) value was calculated as approximately 0.58 in the Figure 4.8. This value shows that the model's ability to distinguish classes is limited, but it performs better than a purely random prediction. Similarly, the area under the PrecisionRecall curve (Average Precision) value was also obtained as approximately 0.58. The Precision-Recall curve was observed to show higher precision values at lower recall levels, but the precision value gradually decreases as the recall increases in the Figure 4.7.

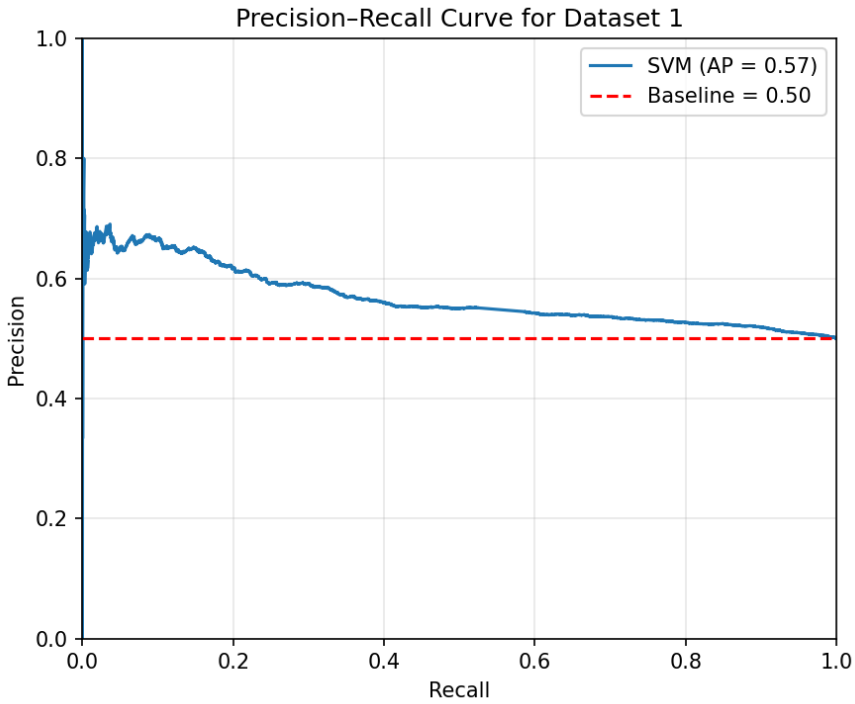

Figure 4.7: Precision Recall Curve for Dataset 1 for SVM

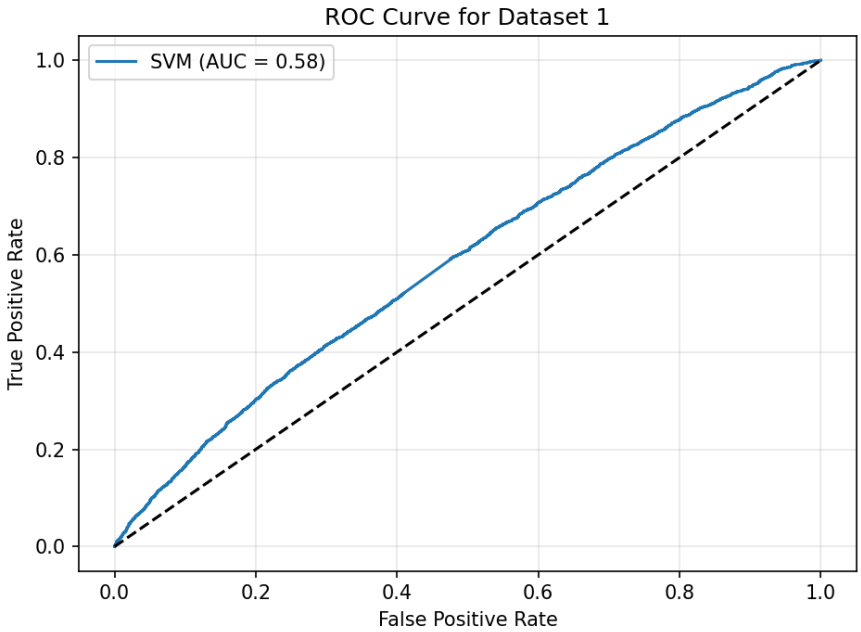

Figure 4.8: ROC for Dataset 1 for SVM

The SVM model applied to Dataset 1 achieves moderate classification performance in the motor imagery task. This result may be influenced by the characteristics of the dataset, the selected feature extraction method, and the variability of EEG signals.

When the SVM based classification results performed on Dataset 1 are examined, it is seen that the model achieves an accuracy level of approximately 55%–56%. Support vector machines have been one of the widely used methods for classifying motor imaging EEG signals for many years [26].

Similarly, a comprehensive review by Lotte et al [9] indicated that the performance of traditional machine learning approaches used in motor imaging BCI systems can vary significantly depending on the characteristics of the dataset and participant differences . This study specifically notes that the CSP + SVM combination is considered a strong fundamental method in many datasets, but performance generally remains in the 80%–90% accuracy range [120]. In this context, the SVM results obtained on Dataset 1 appear to be consistent with typical performance ranges reported in the literature.

CNN Results of Dataset 1

An analysis of the results obtained from the CNN model trained on Dataset 1 indicates that the model exhibits comparable performance across both classes. When the classification report and confusion matrix are evaluated jointly, it is evident that the model does not demonstrate a significant bias toward either class and is capable of distinguishing between them with similar levels of accuracy. However, the overall accuracy remaining at approximately 62% suggests that the models discriminative capability on this dataset is limited.

Examining the complexity matrix presented in Figure 4.9, it is seen that 2531 of the 4200 samples belonging to the left hand motor image class were correctly classified, while 1669 were incorrectly classified. Similarly, in the right hand class, 2606 correct and 1594 incorrect classifications occurred. These results show that the model exhibited balanced performance for both classes, and was slightly more successful in detecting the right hand motor image class.

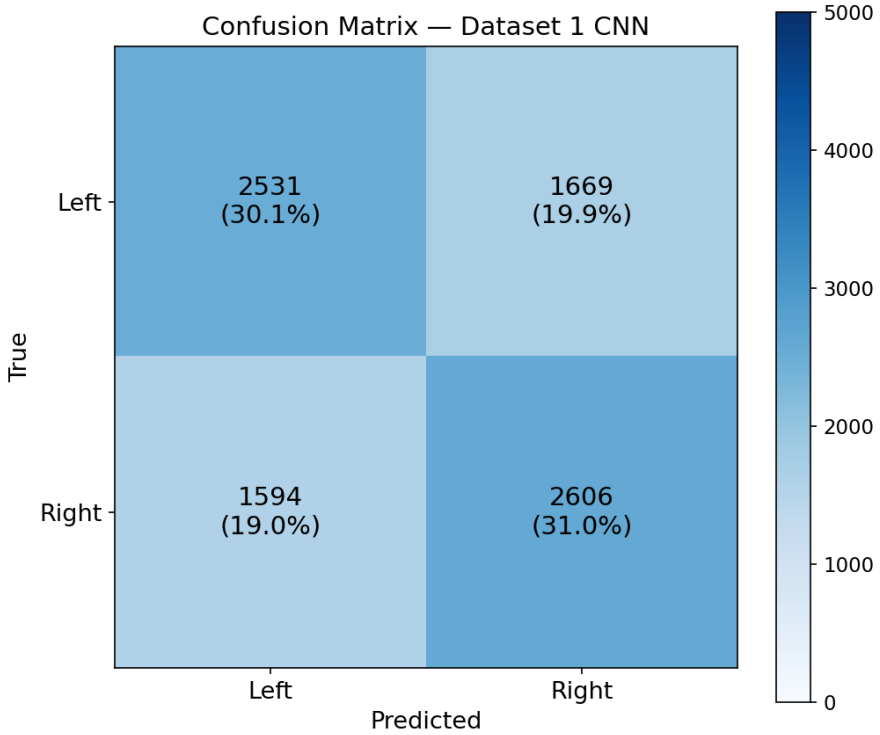

Figure 4.9: Confusion Matrix results for Dataset 1 for CNN

The CNN model applied to Dataset 1 demonstrated a balanced but limited classification performance, achieving an overall accuracy of approximately 61%. The model correctly classified 2531 left-hand and 2606 right-hand samples, while

misclassifying 1669 left samples as right and 1594 right samples as left

Table 4.3: Classification Report of the CNN Model for Dataset 1

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.6136	0.6026	0.6080	4200
Right	0.6096	0.6205	0.6150	4200
Accuracy			0.6115	8400
Macro Avg	0.6116	0.6115	0.6115	8400
Weighted Avg	0.6116	0.6115	0.6115	8400

An examination of the training and validation accuracy curves reveals that the model undergoes a rapid learning phase during the initial epochs, followed by a gradual stabilization in performance in later epochs. The observation that training accuracy is slightly higher than validation accuracy is expected and indicates that the model achieves a better fit on the training data. The relatively small gap between the training and validation curves suggests that the model does not exhibit a strong tendency toward overfitting.

A similar trend is observed in the loss curves, where the training loss decreases steadily, while the validation loss stabilizes after a certain point. This behavior indicates that the learning process progresses in a stable manner.

Throughout the training process, it was observed that validation performance did not show significant improvement beyond approximately 5060 epochs. The absence of notable performance gains after this range suggests that the model reaches a stable performance level within this interval. Therefore, extending the training beyond this point does not appear to provide a substantial benefit.Furthermore, the similarity in trends between the training and validation curves after the normalization process suggests that the data scaling step may have contributed positively to the stability of the models learning process.

The CNN based EEGNet model implemented in Dataset 1 exhibited higher performance compared to the SVM approach, reaching an accuracy of approximately 61%. The ability of deep learning based methods to perform automatic feature extraction from EEG signals has provided a significant advantage in BCI research in recent years. In particular, the EEGNet architecture proposed by Lawhern et al. [29] is designed as a lightweight deep learning architecture capable of learning both temporal and spatial features of EEG signals within the same model, and has achieved more competitive results compared to traditional methods on many motor imaging datasets.

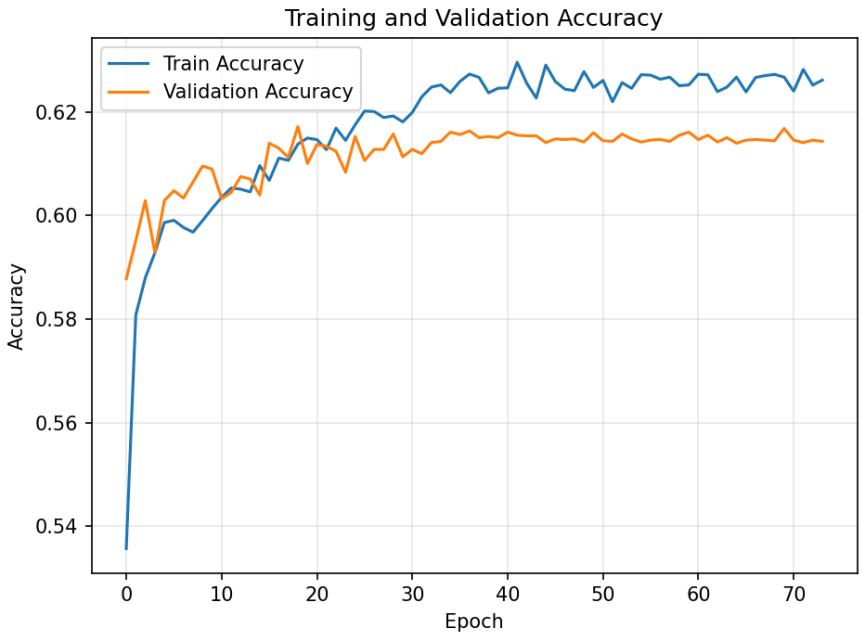

Figure 4.10: Training and validation accuracy curves for Dataset 1

Figure 4.11: Training and validation loss curves for Dataset 1

The literature reports that the EEGNet architecture generally performs in the 60 - 70% accuracy range on different motor imaging datasets (Lawhern et al., 2018). Similarly, the study by Schirrmeister et al. [28] demonstrated that the ability of convolutional neural networks to learn complex spatio temporal patterns in EEG signals provides higher classification performance compared to traditional feature extraction methods. These studies demonstrate that deep learning based approaches can offer significant advantages in motor image classification problems, particularly due to their ability to perform direct learning from raw EEG signals.In Cisotto et al. [121], cross subject classification performance on EEG based motor imagery data was investigated, and CNN based approaches, including the proposed DynamicNet architecture, were shown to achieve competitive performance. The study also highlights that classification performance varies significantly depending on the dataset.

4.4.2 Results on Dataset 2

This part presents the results of the classification experiments performed on Dataset 2. This dataset differs from others in that it contains EEG recordings obtained from a clinical population. Therefore, it is predicted that the classification performance may vary depending not only on the model architecture but also on the participant profile and the recording conditions of the dataset. In this section, the performance of SVM and CNN based models in Dataset 2 is analyzed through accuracy values, confusion matrix results, and ROC and Precision-Recall curves. The obtained results are evaluated in detail to better understand the behavior of the models on clinical data structures.

SVM Results of Dataset 2

When the results of the SVM based classification experiments performed on Dataset 2 are examined, it is seen that the model's performance in distinguishing between the two classes (Left and Right) of motor imagery tasks is limited. The total accuracy value obtained on the test dataset was calculated as 52.7%.

An examination of the confusion matrix reveals that the model correctly classified 520 examples belonging to the Left class, while incorrectly predicting 480 examples as the Right class. Similarly, 534 examples were correctly classified as Right, while 466 were incorrectly assigned to the Left class in Figure 4.12.

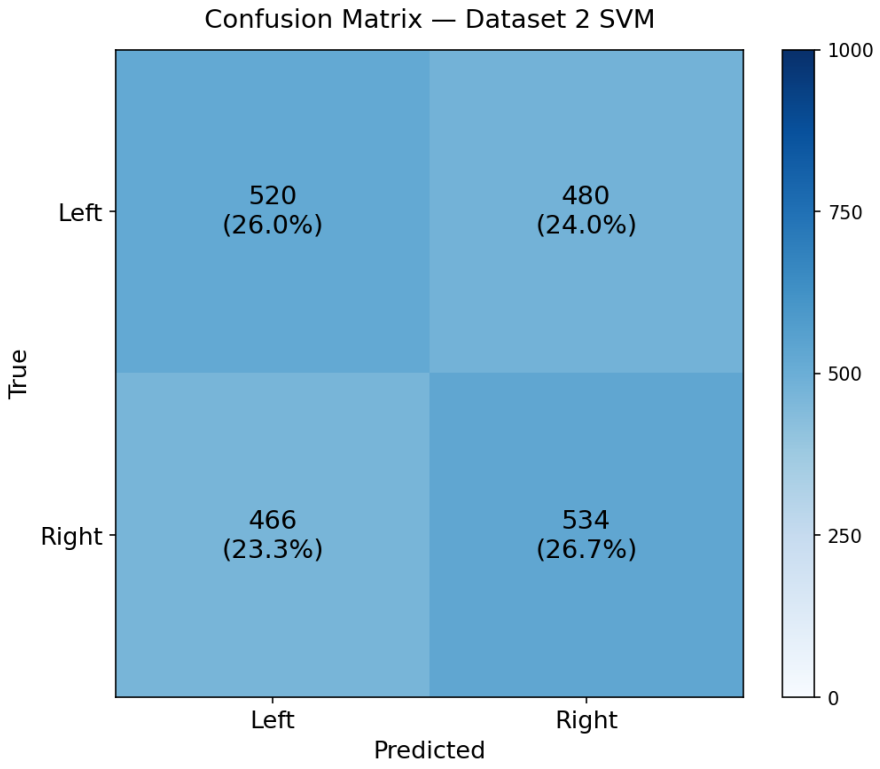

Figure 4.12: Confusion Matrix results for Dataset 2 for SVM

Table 4.4: Classification Report of the SVM Model for Dataset 2

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.5274	0.5200	0.5237	1000
Right	0.5266	0.5340	0.5303	1000
Accuracy			0.5270	2000
Macro Avg	0.5270	0.5270	0.5270	2000
Weighted Avg	0.5270	0.5270	0.5270	2000

To evaluate the model's discrimination capacity, we also examined the ROC curve and Precision-Recall curves. The area under the ROC curve (AUC) was calculated to be approximately 0.53. This value indicates that the model provides only a limited performance improvement compared to random classification. Similarly, the average precision value (AP) under the Precision-Recall curve was obtained as approximately 0.52. These results show that the SVM model struggles to distinguish between classes on Dataset 2 and that the structural characteristics of the dataset may limit the model's performance.

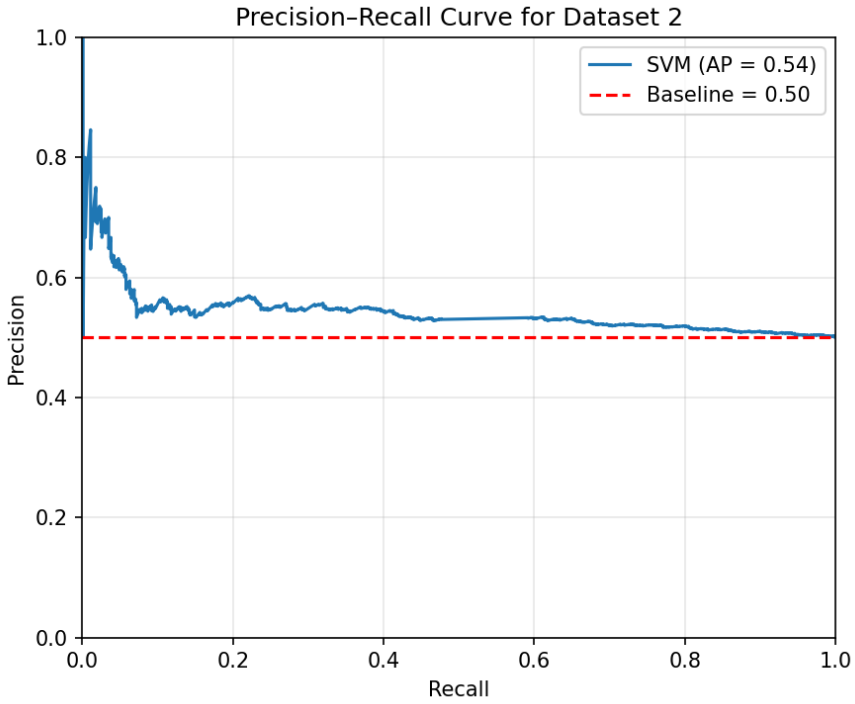

Figure 4.13: Precision Recall Curve for Dataset 2 for SVM

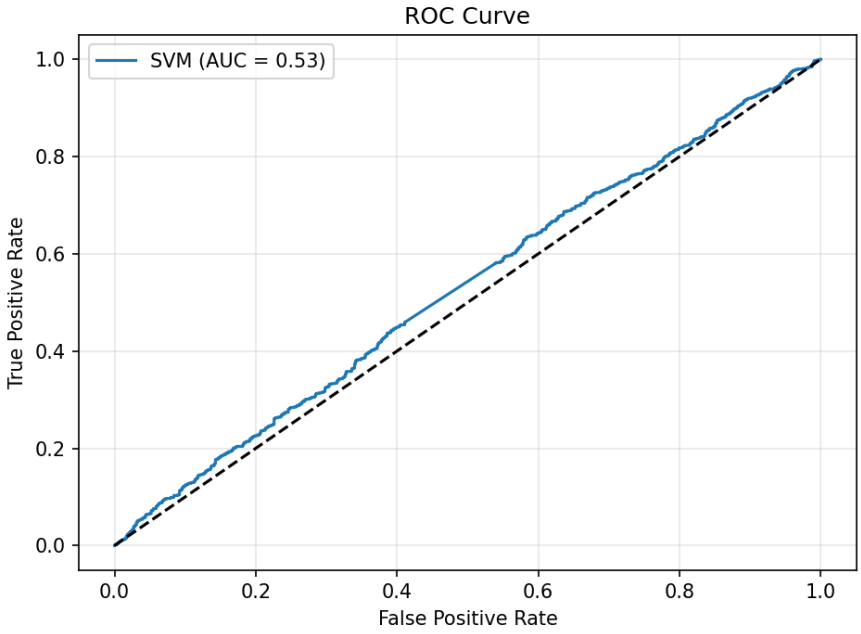

Figure 4.14: ROC for Dataset 2 for SVM

Overall, the results obtained on Dataset 2 indicate that the SVM model offers

moderate classification performance for this data set. The observed performance of the SVM model may be influenced by factors such as inter individual variability in motor imagery EEG signals, feature overlap between classes, and limited feature representation. Since these factors have not been explicitly investigated in this study, they should be interpreted as potential contributors rather than definitive causes. Therefore, evaluating CNN based models on the same dataset is essential to assess whether improved feature learning can enhance classification performance.

In SVM based classification experiments performed on Dataset 2, the model's accuracy was approximately 52.7%. This performance value is lower compared to Dataset 1. This can be largely attributed to the structural characteristics of the data set. Dataset 2 contains motor imagery EEG signals obtained from patients with acute stroke, unlike EEG recordings obtained from healthy individuals. Neurophysiological changes in the motor cortex after a stroke can cause the EEG patterns that appear during motor imagery to be weaker and more irregular [122]. In addition, the limited number of trials per participant in this dataset and the relatively low total amount of data can make it difficult for classical machine learning methods to learn distinguishing features.

SVM classifiers, especially those used in conjunction with CSP or spectral features, are known to be a strong foundational method in numerous studies [24]. However, it has been frequently reported in the literature that classification performance in EEG data obtained from clinical populations may be lower compared to datasets obtained from healthy individuals. For example, studies by Ang et al. showed that classifying motor imagery EEG signals from stroke patients was more difficult and that accuracy values varied significantly between individuals [123] . Similarly, Blankertz et al. emphasized that signal quality and user dependent variability have a significant impact on classification performance in EEG based BCI systems [26] . In this context, it can be said that the SVM performance obtained on Dataset 2 is consistent with the values reported in the literature for clinical populations.

CNN Results of Dataset 2

The CNN based EEGNet model implemented on Dataset 2 demonstrated higher classification performance compared to the SVM model. Figure 4.15 shows the confusion matrix for the CNN model evaluated on Dataset 2. The

model correctly classified 107 left and 130 right instances, while 93 left samples were misclassified as right and 70 right samples were misclassified as left, yielding an overall accuracy of 59.3%.

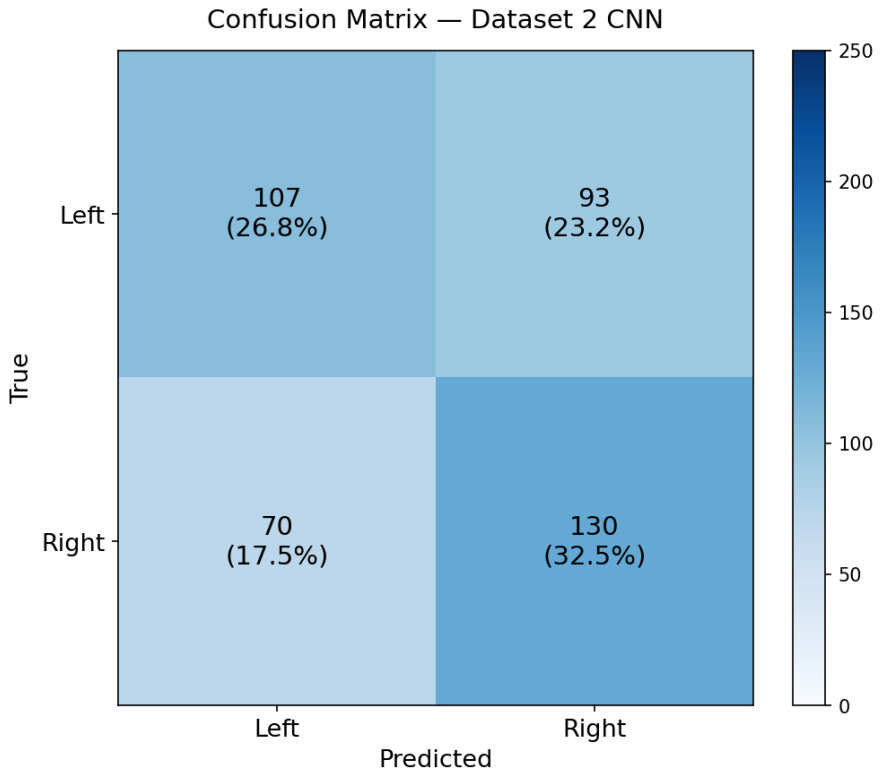

Figure 4.15: Confusion Matrix results for Dataset 2 for CNN

Table 4.4 shows that, the precision value for the Left class was obtained as 0.604, the recall value as 0.535, and the F1-score as 0.567. For the Right class, the precision was calculated as 0.583, the recall as 0.650, and the F1-score as 0.614. These values indicate that the model has a higher sensitivity, especially in identifying the Right class.

Table 4.5: Classification Report of the CNN Model for Dataset 2

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.6045	0.5350	0.5676	200
Right	0.5830	0.6500	0.6147	200
Accuracy			0.5925	400
Macro Avg	0.5937	0.5925	0.5911	400
Weighted Avg	0.5937	0.5925	0.5911	400

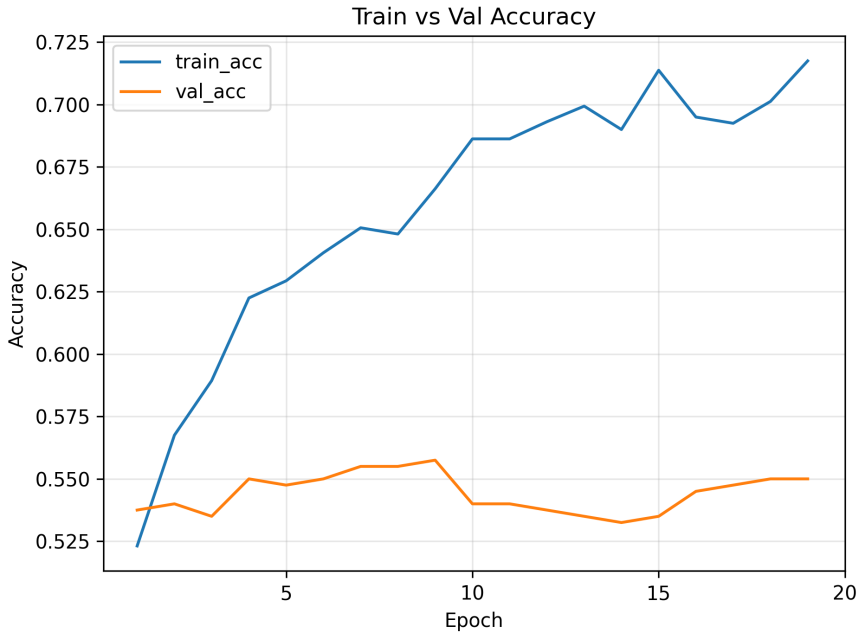

Figure 4.16: Training and validation accuracy curves for Dataset 2

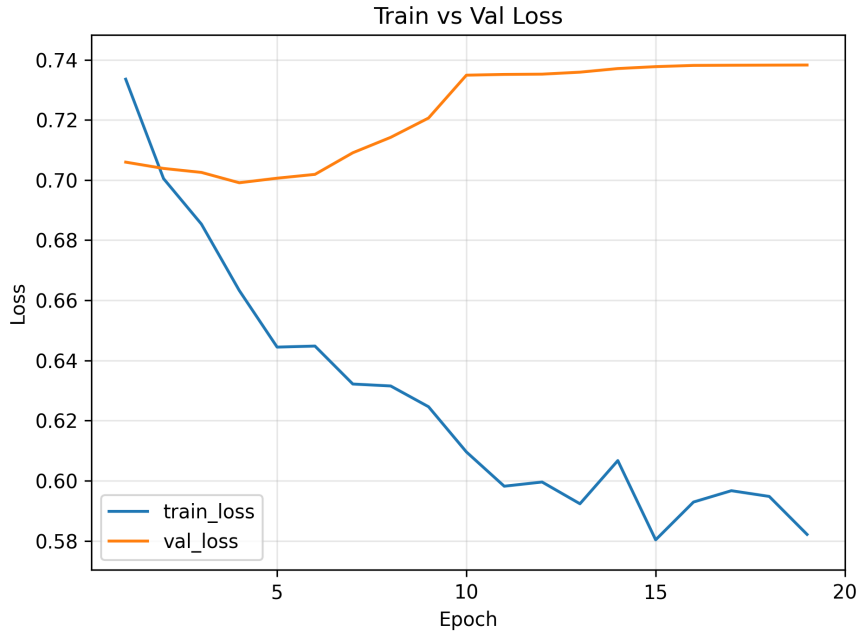

Figure 4.17: Training and validation loss curves for Dataset 2

Examining the accuracy curves of the model's training process reveals that the training accuracy continuously increases as the epoch progresses, reaching approximately 71%. In contrast, validation accuracy remains stable at around

53–55%. This indicates that the model has a high capacity to learn from the training data but cannot maintain the same performance on the validation data. Similarly, while the training loss continuously decreases, the validation loss tends to increase after a certain point, suggesting a partial overfitting tendency in the model.

The CNN based EEGNet model implemented in Dataset 2 achieved approximately 59.25% accuracy, demonstrating higher performance compared to the SVM model. The ability of CNN models to automatically extract features from raw or minimally processed EEG signals allows for more efficient learning of spatial and temporal patterns related to motor imagery tasks [124, 125]. However, an examination of the training and validation accuracy curves shows that while the model's accuracy increased on the training data, the validation accuracy remained within a more limited range. The fact that the EEG recordings in Dataset 2 were obtained from a clinical population may result in lower signal quality compared to datasets obtained from healthy individuals. Therefore, although the CNN model performed better than the SVM, the structural characteristics of the dataset appear to have limited the model's generalization performance to some extent.

In recent years, the use of deep learning methods in EEG based brain computer interface studies has increased significantly. In particular, the ability of convolutional neural networks to automatically learn complex spatial and temporal patterns in EEG signals provides significant advantages compared to traditional feature extraction methods [126]. The EEGNet architecture proposed by Lawhern et al. [29] is widely used in the literature as a compact CNN architecture that achieves competitive performance on different BCI datasets. Experimental results show that EEGNet achieves high performance across multiple paradigms, with AUC values exceeding 0.8 in several tasks, while maintaining a substantially lower number of parameters compared to conventional CNN models. However, it is known that the performance of deep learning models largely depends on the amount of data and the homogeneity of the dataset. study by Schirrmeister et al. [**schirrmeister**] indicated indicated that CNN based models achieve higher performance, especially in large and balanced EEG datasets, but that model generalization can be difficult when the amount of data is limited. While the CNN performance obtained on Dataset 2 is generally consistent with these studies and offers better results compared to the SVM model, it is considered that a certain performance limitation may have been encountered due to

the clinical characteristics of the dataset and the limited amount of data.

4.4.3 Results on Dataset 3

This section presents the results of the classification experiments performed in Dataset 3. This data set, which contains EEG signals from different participants and multiple recording sessions, has a broader structure in terms of data diversity and heterogeneity compared to other datasets used in the study. Therefore, it provides an important testing environment to evaluate the generalization performance of classification models. In this section, the performance of SVM and CNN based models in Dataset 3 is examined using accuracy rates, confusion matrix analyzes, and ROC and Precision-Recall curves. The findings are then comparatively analyzed to evaluate the performance of the models in more heterogeneous data distributions.

SVM Results of Dataset 3

In SVM based classification experiments conducted on Dataset 3, the overall accuracy of the model was obtained as 65.63%. Figure 4.18 shows the confusion matrix for the SVM model evaluated on Dataset 3. The model correctly classified 10437 left and 9639 right instances, while 4858 left samples were misclassified as right and 5657 right samples were misclassified as left, resulting in an overall accuracy of 65.6%. This indicates a certain level of overlap between the two motor imagery classes. The fact that EEG signals in motor imagery tasks produce activity patterns in similar frequency bands, particularly in sensorimotor cortex regions, can make it difficult for classification algorithms to distinguish between these two classes.

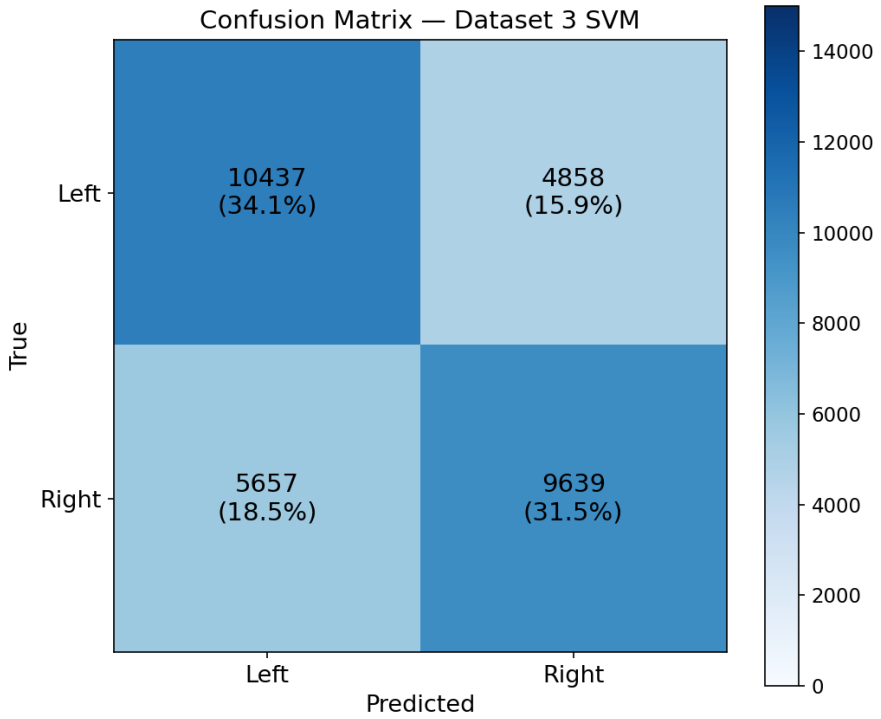

Figure 4.18: Confusion Matrix results for Dataset 3 for SVM

The precision value for the left hand motor imaging class was calculated as 0.6485, the recall value was 0.6824, and the precision value for the right hand motor imaging class was 0.6649 and the recall value 0.6302, as shown in Table 4.7.

Table 4.6: Classification Report of the SVM Model for Dataset 3

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.6485	0.6824	0.6650	15295
Right	0.6649	0.6302	0.6471	15296
Accuracy			0.6563	30591
Macro Avg	0.6567	0.6563	0.6560	30591
Weighted Avg	0.6567	0.6563	0.6560	30591

These results show that the model can distinguish both classes at similar levels, but the recall value in the right hand class is slightly lower compared to the left hand class.

The model's performance was also evaluated using ROC and Precision-Recall analyses. The area under the ROC curve (AUC) was approximately 0.70, indicating that the model exhibits significant discrimination compared to random estimation. The average precision (AP) value obtained from the Precision-Recall

curve was calculated as 0.67. When these metrics are evaluated together, it is seen that the SVM model exhibits a moderate level of classification performance on Dataset 3.

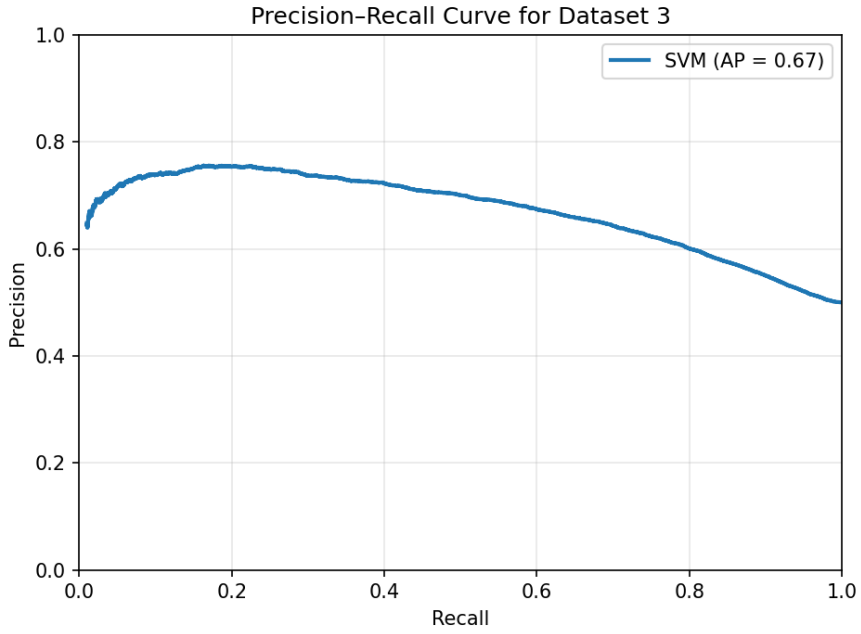

Figure 4.19: Precision Recall Curve for Dataset 3 for SVM

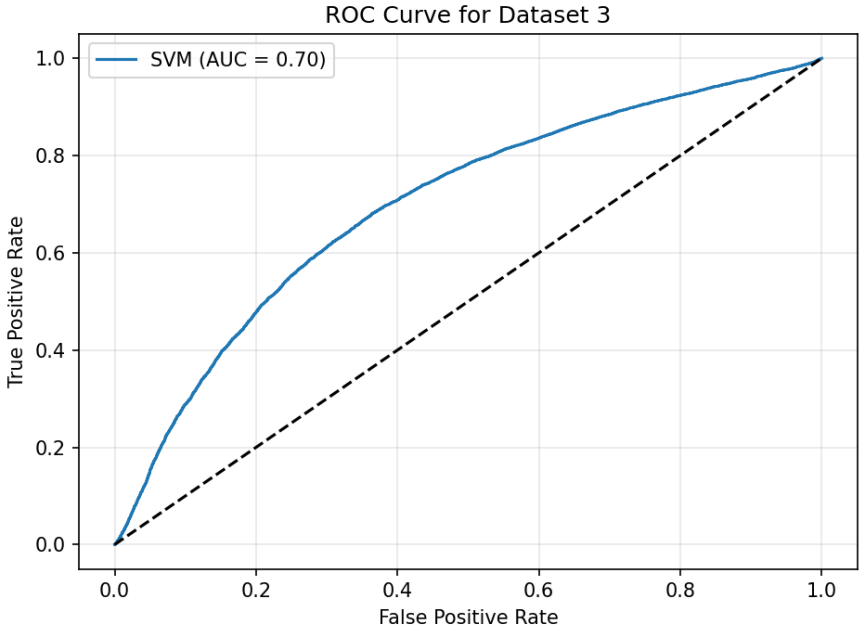

Figure 4.20: ROC for Dataset 3 for SVM

The fact that Dataset 3 has a multi participant and multi session structure is one of the important factors affecting this performance level. The data set contains EEG signals obtained from 62 different participants and three separate recording sessions. The variation between subjects and between sessions that arises from physiological differences between participants and recordings made on different days can lead to significant changes in the statistical properties of the EEG signals. This situation makes it difficult for classical machine learning methods to generalize all variation in the dataset, thus limiting the accuracy of the classification.

In the literature, the Support Vector Machine (SVM) algorithm has long been considered one of the most widely used traditional machine learning methods for classifying motor imagery EEG signals. Especially when used in conjunction with spatial filtering methods such as the Common Spatial Pattern (CSP), the SVM algorithm can achieve powerful and stable results in motor imagery classification problems. This approach has been used as a fundamental reference method in many early BCI studies [127, 128].

On the other hand, the literature also highlights some limitations of SVM based methods. The high dimensional, noisy, and inter individual variations in the nature of EEG signals can limit the generalization ability of traditional machine learning methods. It has been reported that the performance of classical algorithms like SVM can decrease, particularly in datasets obtained from different participants and data structures containing multiple recording sessions. This is due to the fact that motor imagery EEG signals have unique characteristics and that the statistical properties of the signals can vary significantly between individuals.

Therefore, many studies in recent years have compared traditional machine learning methods with deep learning based approaches. The literature generally shows that CNN based models achieve higher performance, especially in large and complex EEG datasets. The main reason for this is that deep learning models can learn the spatial, temporal, and frequency components of EEG signals together and create more abstract feature representations. In this context, the SVM results obtained on Dataset 3 show a trend consistent with the performance levels reported for traditional methods in the motor imaging EEG classification literature.

CNN Results of Dataset 3

Figure 4.21 shows the confusion matrix for the CNN model evaluated on Dataset 3. The model correctly classified 14117 left and 8593 right instances, while 1178 left samples were misclassified as right and 6703 right samples were misclassified as left, yielding an overall accuracy of 74.2%.

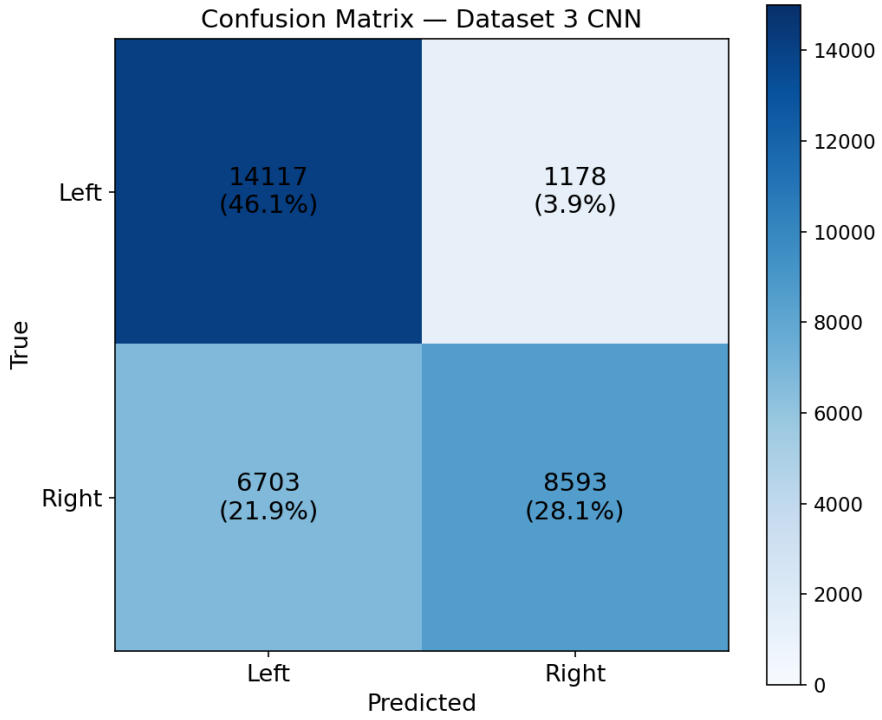

Figure 4.21: Confusion Matrix results for Dataset 3 for CNN

Table 4.7: Classification Report of the CNN Model for Dataset 3

Class	**Precision**	**Recall**	**F1-score**	**Support**
Left	0.6780	0.9230	0.7818	15295
Right	0.8794	0.5618	0.6856	15296
Accuracy			0.7424	30591
Macro Avg	0.7787	0.7424	0.7337	30591
Weighted Avg	0.7787	0.7424	0.7337	30591

When class based performance metrics were examined, the precision value for the left hand motor imaging class was calculated as 0.678, the recall value as 0.923, and the F1-score as 0.7818. For the right hand class, the precision value was obtained as 0.8794, the recall value as 0.5618, and the F1-score as 0.6856.

These results show that the model can capture left hand motor imaging samples with high accuracy, but the recall performance is lower in the right hand class.

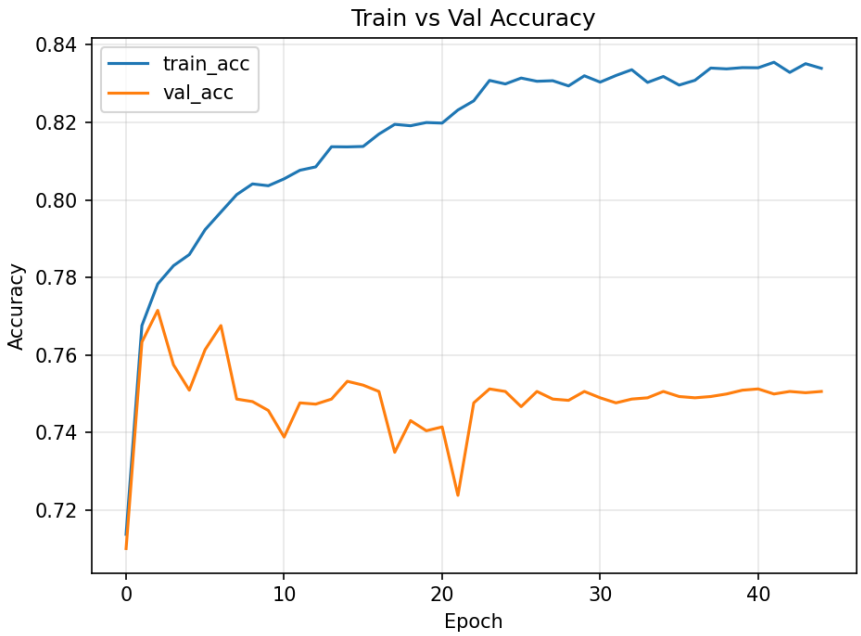

Figure 4.22: Training and Validation Accuracy of Dataset 3

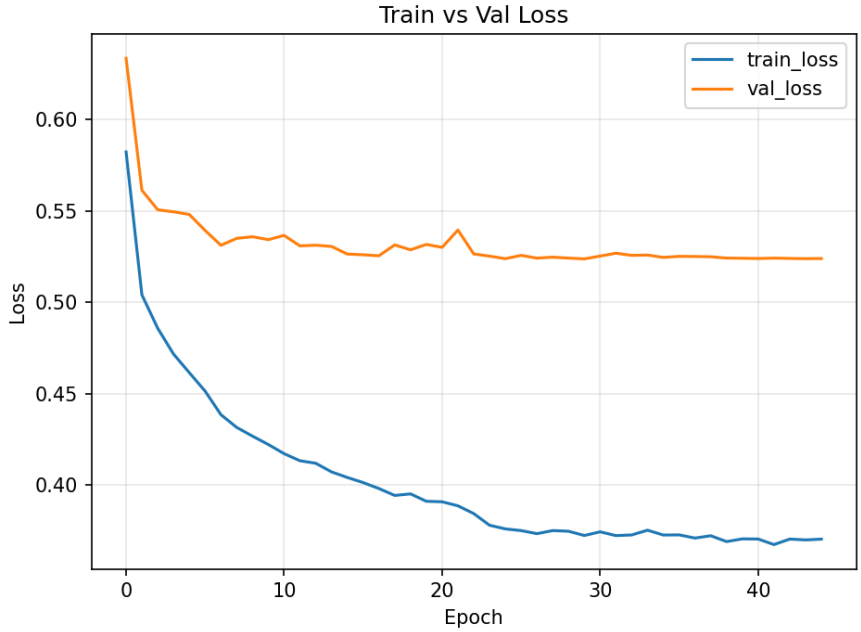

Figure 4.23: Train vs Validation Loss Graph of Dataset 3

When the model's training process was examined, it was observed that the training accuracy increased to 83% as the epoch progressed, while the validation accuracy stabilized at approximately 75%. Similarly, while the training loss decreased steadily, the validation loss showed a more stable trend after a certain point. This indicates that the model can effectively learn patterns in the dataset, but the between participant and between session variation in the dataset may limit the model's generalization performance.

The structural characteristics of Dataset 3 are also among the important factors that affect the performance results obtained. The data set has a complex EEG data structure that involves multiple participants, multiple sessions, and different motor imaging tasks. This situation can particularly make it difficult for deep learning models to generalize the signal patterns of different participants. Nevertheless, the higher accuracy achieved by the CNN model compared to the SVM model is due to the capacity of deep learning based architectures to learn both spatial and temporal characteristics of EEG signals simultaneously.

The obtained results are consistent with prior work demonstrating that convolutional neural networks can provide improved classification performance over traditional machine learning methods in motor imagery EEG analysis, mainly due to their ability to learn discriminative features directly from raw EEG signals [129]. CNN architectures can directly extract features from raw EEG signals, thus reducing dependence on manual feature extraction processes required in traditional methods. Therefore, CNN based models are increasingly used in motor imagery based brain computer interface (BCI) studies. Indeed, a comprehensive review by Craik et al. [10] indicated that deep learning based approaches are increasingly preferred in EEG classification problems, and that convolutional neural networks, in particular, exhibit strong performance in motor imagery tasks .

Numerous studies in the literature report that CNN architectures, particularly through deep learning models such as EEGNet and DeepConvNet, achieve strong classification performance in motor imaging EEG signals. For example, the EEGNet architecture proposed by Lawhern et al. can effectively learn frequency filters and spatial patterns in EEG signals through convolution layers applied in time and channel dimensions, thus providing successful classification results in different datasets [29] . Studies show that CNN based models have a stronger representation learning capability, especially in multi channel and high dimensional EEG data, compared to traditional methods.

However, the literature also emphasizes that motor imagery EEG classification remains a challenging problem. The low signal to noise ratio of EEG signals, inter individual neurophysiological differences, and inter session variations can directly affect classification performance. For example, the study by Roy et al. [36] emphasized that due to the complex and variable nature of EEG signals, different architectures and training strategies should be used to improve the generalization performance of deep learning models. These findings show that the results obtained on Dataset 3 are consistent with the performance ranges reported in the motor imagery EEG classification literature.

To complement the dataset specific evaluations presented above, an additional set of experiments was conducted in which all three datasets were pooled into a single combined dataset and used to train and evaluate both the SVM and CNN classifiers under a unified cross dataset framework. This analysis was motivated by the observation that real world brain computer interface applications may require models capable of generalizing across subjects, recording conditions, and device configurations simultaneously. Given the inherent heterogeneity of the three datasets which differ in terms of sampling frequency, recording protocol, subject population, and preprocessing level the combined setting introduces substantially greater distributional variability than any individual dataset, and may therefore provide a more conservative and ecologically valid assessment of each model's generalization capacity. Prior to pooling, all datasets were resampled or temporally cropped to a uniform representation of 256 time points at 256 Hz, and the same five channel selection (C3, Cz, C4, P3, Pz) was maintained across all sources. A subject-wise train–validation split was applied to the combined pool to ensure that no subject's data appeared in both sets.

The SVM classifier was evaluated on the combined dataset following the same preprocessing pipeline applied to the individual datasets. Raw time series signals were subject to per trial z-score normalization followed by flattening into fixed length feature vectors of dimensionality $256 \times 5 = 1{,}280$, after which StandardScaler normalization was applied. Due to the computational constraints imposed by the size of the combined dataset ($n > 70{,}000$ trials), a stratified random subsample of 10,000 balanced training trials and 4,000 balanced validation trials was drawn from the combined pool, maintaining equal class proportions in both sets. The combined SVM model achieved a validation accuracy of 60.20% and an AUC of 0.659, with an Average Precision of 0.661. Precision and F1-score

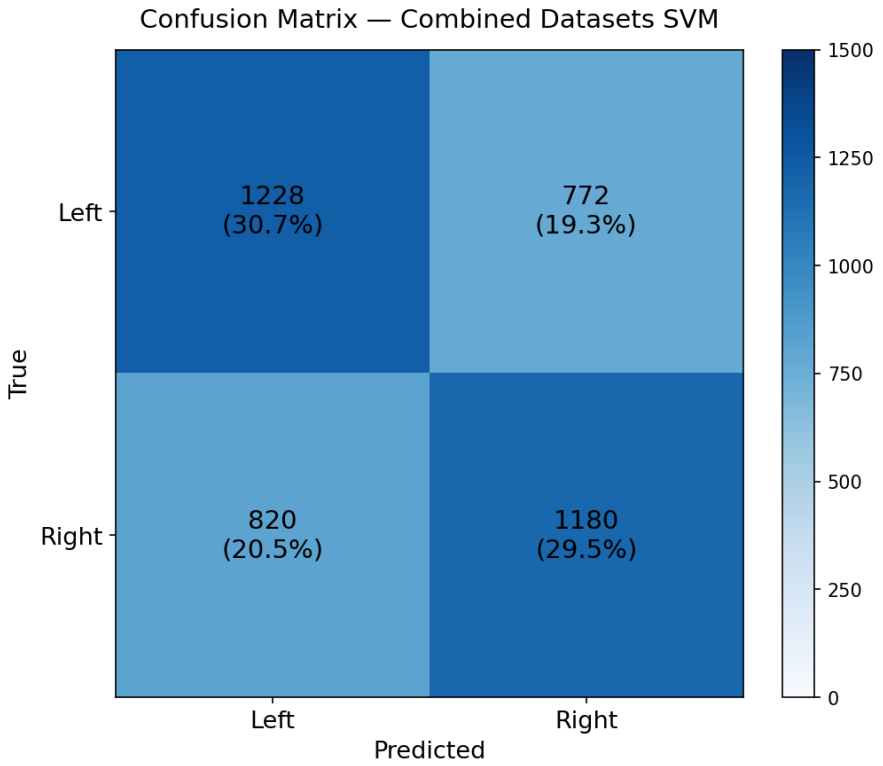

Figure 4.24: Confuion Matrix for Combined Dataset for SVM

values were 0.5996 and 0.6067 for the left-hand class, and 0.6045 and 0.5972 for the right hand class, respectively, suggesting a broadly symmetric classification behaviour across both motor imagery classes. These results may indicate that raw temporal feature representations retain a degree of cross dataset discriminability when combined with per trial normalization, though the observed performance relative to individually trained SVM models could be interpreted as evidence that fixed feature extraction strategies may face increased difficulty in adapting to the distributional heterogeneity introduced by pooling data from sources with different recording devices, sampling frequencies, and experimental protocols.

The CNN model trained on the combined dataset achieved a validation accuracy of 67.21% and an AUC of 0.755. The precision and F1 score values for the left hand class were 0.6835 and 0.6617, respectively, while the right hand class yielded a precision of 0.6621 and an F1-score of 0.6819, suggesting a modest but consistent discriminative capacity across both classes.

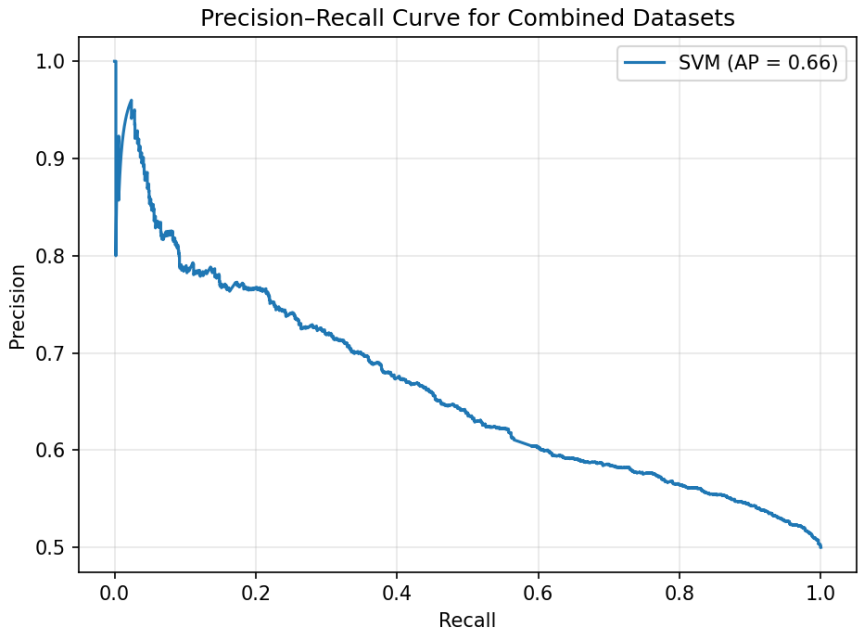

Figure 4.25: Precision Recall Curve for Combined Dataset for SVM

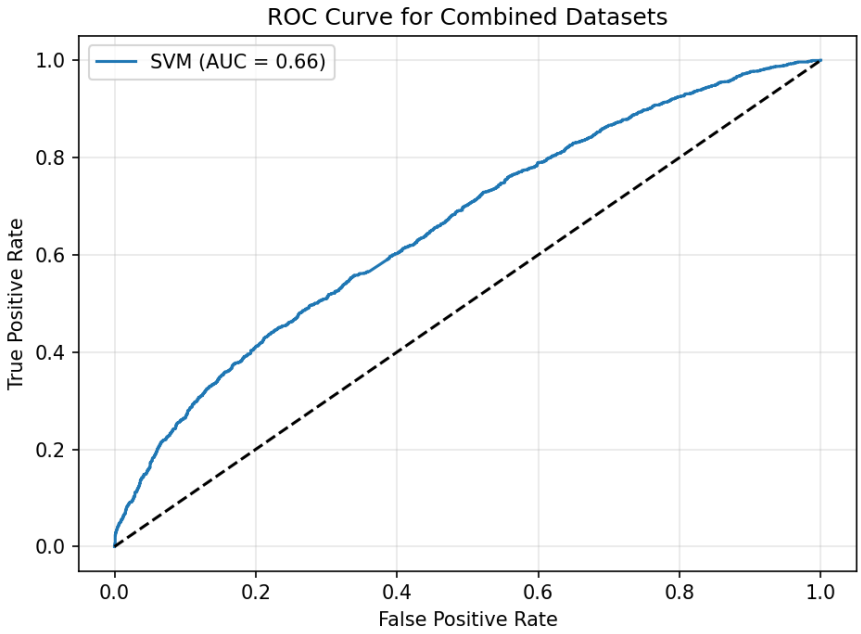

Figure 4.26: ROC for Combined Dataset for SVM

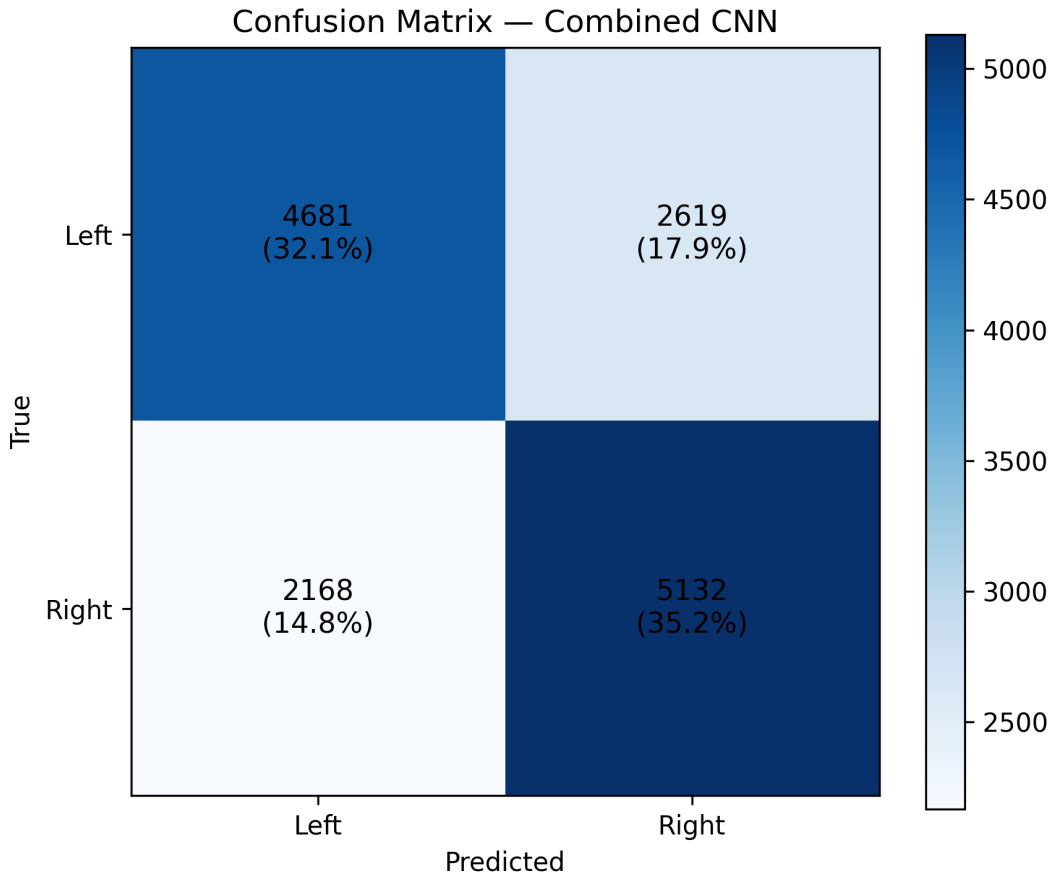

Figure 4.27: Confuion Matrix for Combined Dataset for CNN

The training and validation accuracy curves followed closely parallel trajectories throughout the training process, with no substantial divergence observed between the two, which may suggest that the model did not exhibit severe overfitting despite the considerable heterogeneity of the combined training data. The validation loss similarly decreased in a stable manner across epochs, which could be interpreted as an indication that the model was able to extract at least partially generalizable representations from signals recorded under different experimental conditions, sampling frequencies, and subject populations. It should be noted that the observed accuracy represents a reduction compared to the performance obtained on certain individual datasets, a pattern that may be expected given the increased variability introduced by combining data from sources with inherently different recording protocols and preprocessing levels.

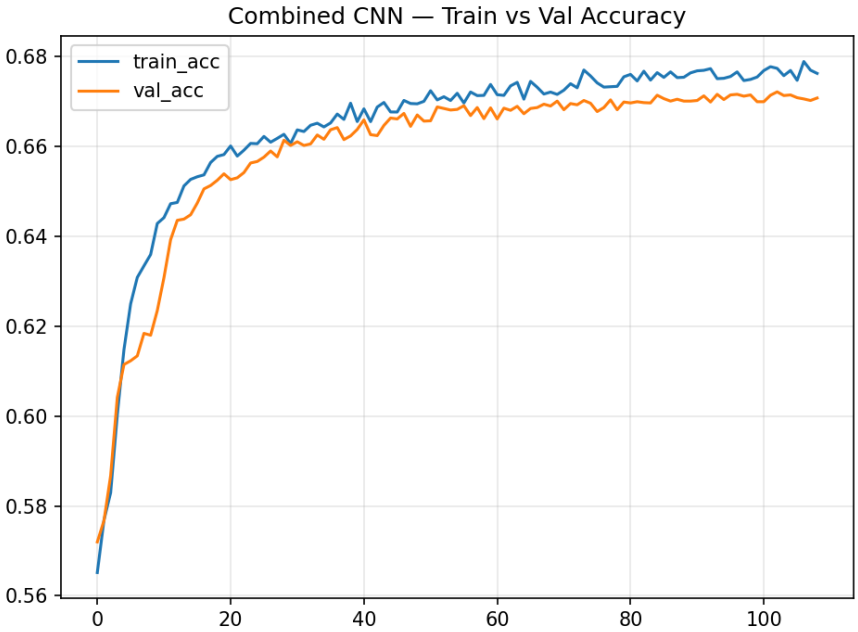

Figure 4.28: Training and Validation Accuracy of Combined Dataset

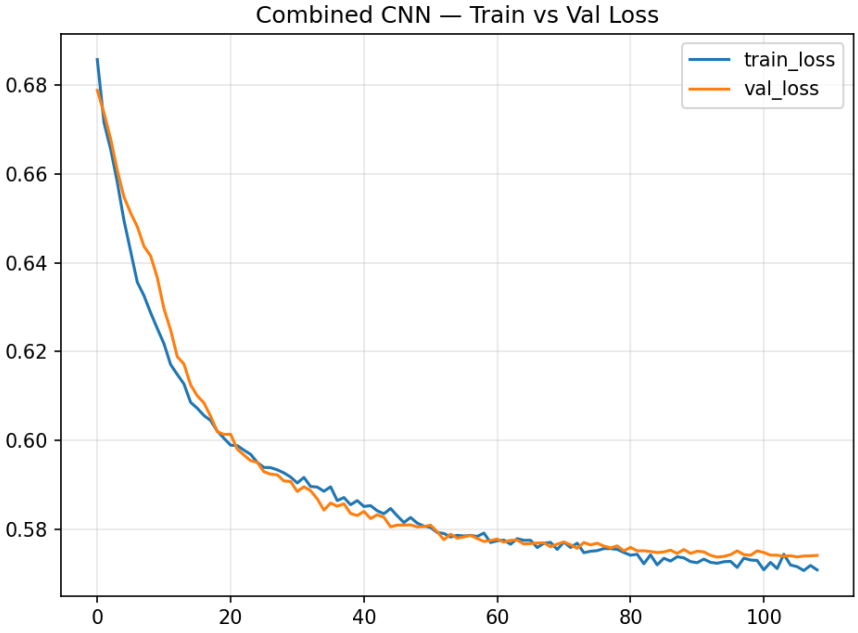

Figure 4.29: Training and Validation Loss of Combined Dataset

Taken together, the results obtained from the combined dataset experiments may suggest several preliminary observations regarding the cross dataset generalizability of the two classification approaches.

The CNN model, which relies on data driven feature learning through successive convolutional operations, appeared to retain a higher level of discriminative performance under the combined heterogeneous setting compared to the SVM classifier, which depends on fixed feature representations. This pattern could be interpreted as tentative evidence that deep learning architectures may be better positioned to adapt to the distributional variability inherent in multi source EEG data, though the observed differences in performance should be interpreted with caution given the methodological differences in feature extraction and the subsampling applied to the SVM pipeline. Whether these findings would generalize to larger or more diverse collections of EEG datasets, or whether domain adaptation and transfer learning strategies might further improve cross dataset classification performance, remains an open question that may warrant further investigation. Nevertheless, the fact that both models achieved performance substantially above chance level on a combined dataset drawn from three heterogeneous sources could be considered as preliminary support for the feasibility of developing EEG based motor imagery classifiers that are not strictly tied to a single recording context.

4.5 Discussion

This study's evaluation process not only relied on performance analysis of classification models but also included a two stage analysis process examining the reporting structure of EEG datasets and methodological heterogeneity in the literature. In the first stage, a systematic literature review was conducted by a panel of four individuals representing different levels of expertise, and methodological criteria for EEG datasets were independently evaluated. During this process, differences in evaluation among team members were analyzed, and common decision making mechanisms were developed to standardize the evaluation criteria. In the second stage, this standard framework created by the human panel was transferred to a GPT based model, and the extent to which the model could analyze the same literature data accurately, consistently, and systematically was examined. This two stage evaluation process formed the methodological basis of the study, demonstrating how both human expertise and AI based automated analysis tools can play complementary roles in data extraction processes in the EEG literature.

After these two stage, the Support Vector Machine (SVM) as a traditional

machine learning method and the Convolutional Neural Network (CNN) as a deep learning based architecture were comparatively evaluated on three different open access EEG datasets for the classification of motor imagery based EEG signals. Overall, the CNN based model achieved higher classification accuracy than the SVM model across all datasets. This can be explained by the ability of deep learning based approaches to directly learn the complex spatial and temporal structure of EEG signals. Since EEG signals are high dimensional, noisy, and time varying biological signals, traditional machine learning methods often rely on predefined feature extraction methods to obtain distinguishing features from these signals. In contrast, convolution based architectures can create more robust feature representations by modeling both the temporal dynamics of signals and the spatial relationships between electrodes. It has been shown in the literature that the EEGNet architecture proposed by Lawhern et al (2018) can provide higher classification performance compared to traditional methods by learning the temporal and spatial features of EEG signals together [29].

An examination across the three datasets indicates that classification performance may be related not only to the model architecture but also to the structural characteristics of the datasets. Both SVM and CNN models were observed to achieve relatively high accuracy values on Dataset 1.This dataset contains EEG recordings obtained from motor imagery tasks performed under controlled experimental conditions on healthy participants. Conducting motor imagery tasks within a structured protocol and controlled environment may increase the prominence of sensorimotor rhythms, particularly the μ (8–13 Hz) and β (13–30 Hz) frequency bands, in sensorimotor cortex regions. This, in turn, may facilitate the ability of classification algorithms to learn EEG patterns associated with motor imagery. In the literature, classification accuracies are often reported to be relatively higher in motor imagery EEG datasets collected from healthy individuals. [26, 9].

The results obtained on Dataset 2 indicate lower classification performance compared to other datasets. This is closely related to the participant population and recording structure of the dataset. The EEG recordings used in Dataset 2 were obtained not from healthy individuals, but from stroke patients experiencing motor function loss. Neurophysiological changes in sensorimotor cortex regions after stroke can cause EEG patterns during motor imagery to be weaker and more variable [130]. Therefore, the classification of EEG data obtained from stroke patients is generally considered a more difficult problem in motor im-

agery based brain computer interface systems. The literature reports that the classification performance of motor imagery EEG signals obtained from stroke patients may be lower compared to healthy individuals [123, 131]. In addition, the fact that Dataset 2 has a multisession structure is one of the important factors affecting the classification performance. Since EEG signals can vary over time, recordings made on different days can lead to significant differences in signal distributions. This phenomenon, referred to in the literature as crosssession variability, is considered one of the fundamental problems that hinders the generalization performance of BCI systems [132].

Dataset 3 results show a performance distribution reflecting the heterogeneous nature of the dataset, which includes a larger participant population and multiple recording sessions. Since this dataset contains EEG recordings from different participants, interindividual neurophysiological differences can directly affect the classification process. This is consistent with previous studies showing that EEG signals vary significantly across subjects, resulting in distribution shifts that hinder the generalization performance of BCI systems [9, 133]. In contrast, CNN based models, thanks to their more flexible feature learning capacity, can model signal patterns from different participants more effectively. This is consistent with literature findings showing that deep learning approaches can provide more successful results, especially in large and heterogeneous EEG datasets [28].

One of the important aspects of this study is that the evaluations were not limited to a single dataset, but were performed on multiple open access EEG datasets with different structural characteristics. A common problem in EEG based brain computer interface studies is that the developed methods are evaluated only on a specific dataset, and therefore the generalizability of the results is limited. The datasets used in this study exhibit different characteristics in terms of participant population, experimental protocol, recording sessions, and data size. Thanks to this heterogeneous structure, it was possible to observe how the classification models used behave under different data conditions, rather than only fitting to a specific data distribution.

The results showed that GPT based systems can extract technical parameters directly reported in the plaintext with high accuracy and consistency, but human expertise still plays a significant role in situations requiring contextual interpretation. These results indicate that combining human and GPT based evaluation improves consistency and reduces subjectivity in EEG literature classification.

The experimental findings obtained in this study reveal the following key conclusions:

1. CNN based models consistently outperformed the SVM model across all datasets, demonstrating the advantage of deep learning approaches, especially in heterogeneous and multi participant EEG scenarios.

2. The structural characteristics of EEG datasets (participant population, recording sessions, and experimental protocol) directly affect classification accuracy.

3. Evaluations performed on different datasets allow for a more reliable analysis of the generalization performance of the models.

Conclusions and Future Works

This thesis presents a multi stage research approach for the analysis and classification of motor imagery based EEG datasets. The study first involves a systematic evaluation process to examine the methodological characteristics of datasets used in the EEG literature. Following this, the applicability of this evaluation framework to automated analysis tools, with a particular focus on Generative AI (GenAI) systems such as the ChatGPT model, is investigated. Finally, the EEG classification performance of different machine learning methods is comparatively evaluated.

In the first phase of the study, were systematically examined by a human panel of evaluators representing four different levels of expertise. These evaluations revealed significant heterogeneity in the reporting methods of datasets in the EEG literature. While fundamental technical parameters such as sampling frequency and number of channels were explicitly reported in most studies, critical methodological details such as channel configuration, reference electrode information, session structure, and task protocol were found to be incomplete or indirectly presented in some studies. The decision making mechanisms developed for dataset selection and classification, along with the standardized evaluation criteria for EEG datasets, may contribute to a more systematic analysis of datasets in the literature.

In the second phase of the study, this standard evaluation framework created by the human panel was transferred to a GPT based model, and the model's capacity to automatically analyze information from the same datasets was evaluated. The results showed that the GPT based analysis system could extract technical parameters explicitly reported in the text with high accuracy and consistency. Parameters that are clearly and explicitly reported in the articles, including sampling frequency, number of channels, data format, and number of participants, were reliably extracted by the model. However, certain limitations

of the model may become apparent in cases that require contextual interpretation. For example, information related to channel configuration, often described indirectly through references to standard electrode systems (e.g., 10–10 or 10–20 systems), may not always be explicitly identified by the model. Similarly, details of the experimental protocol, such as session structure, block organization, and trial timing, can be more challenging to extract, as they are sometimes distributed across different sections of the text or embedded within figures rather than being clearly stated. In such cases, the models performance may be affected by the implicit and fragmented nature of the information. Nevertheless, the GPT based approach may still provide advantages in terms of speed and consistency, particularly for extracting clearly defined and explicitly reported parameters.

In the final phase of the study, the performance of CNN and SVM based classification models was comparatively evaluated on the selected EEG datasets. The results indicate that both models were capable of performing the motor imagery EEG classification task; however, CNN based architectures tended to achieve higher classification accuracy across all datasets. Specifically, for Dataset 1, CNN achieved an accuracy of 61.15%, while SVM achieved 55.74%. For Dataset 2, CNN reached 59.25%, compared to 52.69% for SVM. Similarly, in Dataset 3, CNN achieved 74.24%, whereas SVM reached 65.63%.

These results suggest that the observed performance differences may not be solely attributed to the model architecture, but may also be influenced by dataset-specific characteristics. In particular, factors such as the participant population (e.g., healthy subjects vs. clinical populations), variability across subjects, the structure of the experimental protocol (e.g., session design and trial organization), and recording conditions (e.g., number of channels and signal quality) may contribute to variations in classification performance. For example, datasets collected from clinical populations or with higher inter subject variability may present additional challenges for classification, potentially leading to lower accuracy.

Therefore, these findings highlight the importance of evaluating model performance in EEG based brain–computer interface research in conjunction with dataset characteristics, rather than attributing performance differences solely to the choice of model.

One of the significant contributions of this study is demonstrating that human expertise and AI based analysis tools can be used together in the evaluation of EEG datasets. While human panel evaluations offer significant advantages

in situations requiring contextual interpretation and methodological inference, GPT based systems stand out as an effective tool for the rapid and consistent extraction of plaintext based information. Therefore, in the future, hybrid approaches combining human expertise with automated analysis systems in EEG literature analysis can improve both the accuracy and scalability of data extraction processes.

In conclusion, this thesis presents a research framework that combines systematic analysis of EEG datasets, automated data extraction from the literature, and a comparative evaluation of motor imagery EEG classification models. This approach contributes to a more systematic analysis of datasets in the EEG literature, while also allowing for more reliable and comparable evaluation of classification studies performed on different datasets.

Although this study evaluates systematic analysis of EEG datasets, GPT-based automated data extraction, and machine learning-based classification approaches together, future studies may further expand this framework. Rather than focusing solely on improving classification performance, future work may investigate the generalization capacity of models across different EEG datasets acquired under varying experimental setups and protocols. In particular, differences in subject populations, recording conditions, and experimental designs may significantly affect model performance. Therefore, developing models that are robust to such variations and capable of generalizing cross data sets remains an important research direction. Additionally, advanced deep learning approaches, such as Transformer-based models or hybrid architectures, may be explored in this context to improve generalization performance rather than only dataset specific accuracy.

Furthermore, the application of transfer learning and domain adaptation methods across different datasets may represent an important research direction for improving model generalization on heterogeneous EEG datasets. In addition, the development of LLM-based analysis systems capable of processing multimodal data sources (e.g., text, figures, and tables) may enable a more comprehensive and automated data extraction from the EEG literature [74]. Finally, future studies may investigate the potential adaptation of such approaches for real-time brain–computer interface systems and explore their applicability in clinical settings, which could contribute to enhancing the practical usability of EEG based BCI technologies.

References

[1] Noor Kamal Al-Qazzaz et al. "A Review of Brain Activity and EEG-Based Brain–Computer Interfaces for Rehabilitation Application". In: *Bioengineering* 9.12 (2022), pp. 1–28.

[2] Xiu-Yun Liu et al. "Recent Applications of EEG-Based Brain-Computer-Interface in the Medical Field". In: *Military Medical Research* 12.14 (2025), pp. 1–35.

[3] Essam H. Houssein et al. "Deep Learning Approaches for EEG-Based Healthcare Applications: A Comprehensive Review". In: *Frontiers in Human Neuroscience* 19.1 (2025), pp. 1–40.

[4] Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. "Removal of Artifacts from EEG Signals: A Review". In: *Sensors* 19.5 (2019), pp. 1–27.

[5] Mamunur Rashid et al. "EEG Datasets for Seizure Detection and Prediction: A Review". In: *Epilepsia Open* 5.3 (2020), pp. 354–373.

[6] Valerie Van Brabant et al. "Reproducible Machine Learning Research in Mental Workload Classification Using EEG". In: *Frontiers in Neuroergonomics* 5.1 (2024), pp. 1–22.

[7] Qusai Khraisha et al. "The Emergence of Large Language Models as Tools in Literature Reviews: A Large Language Model-Assisted Systematic Review". In: *JMIR AI* 4.1 (2025), pp. 1–18.

[8] Alice Ferretti et al. "Evaluation of a Large Language Model (ChatGPT) versus Human Researchers in Assessing Risk-of-Bias and Community Engagement Levels: A Systematic Review Use-Case Analysis". In: *European Journal of Public Health* 35.3 (2025), pp. 1–8.

[9] Fabien Lotte et al. "A review of classification algorithms for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 15.3 (2018), p. 031005.

[10] Alexander Craik, Yongtian He, and José L. ContrerasVidal. "Deep learning for electroencephalogram (EEG) classification tasks: a review". In: *Journal of Neural Engineering* 16.3 (2019), p. 031001. DOI: 10.1088/1741-2552/ab0ab5.

[11] Yi Zhang et al. "Efficient Feature Extraction for EEG-Based Classification: A Comparative Review of Deep Learning Models". In: *AI* 7.2 (2026), pp. 1–30.

[12] Gaurav Dhiman et al. "A systematic review of machine learning and deep learning techniques for EEG signal analysis". In: *Artificial Intelligence Review* 56.2 (2023), pp. 1531–1578.

[13] Md. Shamim Hossain et al. "Deep learning approaches for EEG-based classification: A comprehensive review". In: *Expert Systems with Applications* 213 (2023), p. 118955.

[14] Giulia Cisotto et al. "hvEEGNet: A Novel Deep Learning Model for High-Fidelity EEG Reconstruction". In: *Frontiers in Neuroinformatics* 18 (2024), p. 1459970. DOI: 10.3389/fninf.2024.1459970.

[15] Revati Shriram, Mahalingam Sundararajan, and Nivedita Daimiwal. "EEG Based Cognitive Workload Assessment for Maximum Efficiency". In: *International Journal of Advanced Research in Computer Science and Software Engineering* 2.8 (Aug. 2012), pp. –.

[16] Kevin Murungi et al. "Machine learning-based EEG signal classification: Methods and applications". In: *Biomedical Signal Processing and Control* 85 (2023), p. 104834.

[17] A. R. Hassan et al. "EEG-based brain state classification using traditional machine learning approaches". In: *Computers in Biology and Medicine* 156 (2023), p. 106628.

[18] Cédric Rommel et al. "Deep learning for EEG decoding: From convolutional neural networks to transformers". In: *IEEE Signal Processing Magazine* 39.2 (2022), pp. 24–36.

[19] Alexey Dosovitskiy et al. "Transformer-based deep learning models for EEG classification". In: *Neurocomputing* 500 (2022), pp. 123–135.

[20] Soumya Samal et al. "EEG-based cognitive and affective state classification: A review". In: *Information Fusion* 103 (2024), p. 102084.

[21] Marta Bilucaglia et al. “EEG-based emotion and cognitive workload assessment”. In: *Frontiers in Neuroscience* 15 (2021), p. 646528.

[22] Yifan Zhao et al. “Multimodal EEG-based classification using deep learning”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 32 (2024), pp. 455–466.

[23] Gert Pfurtscheller and Christa Neuper. “Motor imagery and EEG-based control”. In: *Clinical Neurophysiology* 112.3 (2001), pp. 416–431.

[24] Luis F. Nicolas-Alonso and Jaime Gomez-Gil. “Brain–computer interfaces, a review”. In: *Sensors* 12.2 (2012), pp. 1211–1279.

[25] Gert Pfurtscheller et al. “Event-related desynchronization and synchronization in EEG”. In: *Progress in Brain Research* 159 (2006), pp. 65–76.

[26] Benjamin Blankertz et al. “Single-trial analysis and classification of EEG”. In: *Clinical Neurophysiology* 121.11 (2010), pp. 1738–1747.

[27] Vinay Jayaram and Alexandre Barachant. “MOABB: Trustworthy algorithm benchmarking for BCIs”. In: *Journal of Neural Engineering* 15.6 (2018).

[28] Robin Tibor Schirrmeister et al. “Deep learning with convolutional neural networks for EEG decoding”. In: *Human Brain Mapping* 38.11 (2017), pp. 5391–5420.

[29] Vernon J. Lawhern et al. “EEGNet: A compact convolutional neural network for EEG-based BCIs”. In: *Journal of Neural Engineering* 15.5 (2018).

[30] Wojciech Samek et al. “Explainable deep learning for EEG-based brain–computer interfaces”. In: *arXiv preprint arXiv:1708.02666* (2017).

[31] Hyun Cho et al. “Generalization issues of CNN-based EEG classification”. In: *Sensors* 20.15 (2020).

[32] Wolfgang Klimesch. “EEG alpha and theta oscillations and cognitive processes”. In: *International Journal of Psychophysiology* 33.2 (1999), pp. 169–195.

[33] Gianluca Borghini et al. “EEG-based mental workload assessment”. In: *Neuroscience & Biobehavioral Reviews* 44 (2014), pp. 1–16.

[34] Alan Gevins and Michael E. Smith. “Neurophysiological measures of cognitive workload”. In: *Theoretical Issues in Ergonomics Science* 4.1-2 (2003), pp. 113–131.

[35] Anne-Marie Brouwer et al. “Estimating workload using EEG”. In: *International Journal of Psychophysiology* 83.1 (2012), pp. 109–120.

[36] Yannick Roy et al. “Deep learning-based EEG classification: A review”. In: *Journal of Neural Engineering* 16.5 (2019).

[37] Jian Zhang et al. “Multi-task learning for EEG-based classification”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), pp. 120–131.

[38] Russell A. Poldrack et al. “Scanning the horizon: Towards transparent and reproducible neuroimaging research”. In: *Nature Reviews Neuroscience* 18 (2017), pp. 115–126.

[39] Krzysztof J. Gorgolewski et al. “The Brain Imaging Data Structure (BIDS)”. In: *Scientific Data* 3 (2016), p. 160044.

[40] Katherine S. Button et al. “Power failure: Why small sample size undermines the reliability of neuroscience”. In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376.

[41] Ary L. Goldberger et al. “PhysioNet: Components of a new research resource for complex physiological signals”. In: *Circulation* 101.23 (2000), e215–e220.

[42] Gerwin Schalk et al. “BCI2000: A general-purpose brain–computer interface system”. In: *IEEE Transactions on Biomedical Engineering* 51.6 (2004), pp. 1034–1043.

[43] Arman Babayan et al. “A mind–brain–body dataset of MRI, EEG, cognition, emotion”. In: *Scientific Data* 6 (2019), p. 180308.

[44] Arnaud Delorme et al. “EEGLAB: An open source toolbox for EEG analysis”. In: *Journal of Neuroscience Methods* 134.1 (2012), pp. 9–21.

[45] Mehrdad Fatourechi et al. “EMG and EOG artifact removal from EEG”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15.2 (2007), pp. 151–157.

[46] Jason Onton and Scott Makeig. “Information-based modeling of EEG dynamics”. In: *Progress in Brain Research* 159 (2006), pp. 99–120.

[47] C. R. Pernet et al. “EEG-BIDS, an extension to the brain imaging data structure for electroencephalography”. In: *Scientific Data* 6.1 (2019), p. 103. DOI: 10.1038/s41597-019-0104-8.

[48] Emily S. Kappenman and Steven J. Luck. "The effects of electrode impedance on EEG data quality". In: *Psychophysiology* 47.5 (2010), pp. 888–904.

[49] Steven J. Luck. *An introduction to the event-related potential technique*. MIT Press, 2014.

[50] Ernst Niedermeyer and Fernando Lopes da Silva. *Electroencephalography: Basic principles, clinical applications*. Lippincott Williams & Wilkins, 2005.

[51] Patrick E. Shrout and Joseph L. Fleiss. "Intraclass correlations: Uses in assessing rater reliability". In: *Psychological Bulletin* 86.2 (1979), pp. 420–428.

[52] Andrew Webb et al. "Inter-rater variability in EEG interpretation". In: *Clinical Neurophysiology* 126.4 (2015), pp. 814–821.

[53] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[54] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[55] Yifan Zhang et al. "Large language models for data annotation". In: *Nature Machine Intelligence* 5 (2023), pp. 223–235.

[56] Fabrizio Gilardi et al. "ChatGPT outperforms crowd workers for text annotation". In: *Proceedings of the National Academy of Sciences* 120.30 (2023).

[57] Emily M. Bender et al. "On the dangers of stochastic parrots". In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (2021).

[58] Mehves Koksa. "Developing an NLP-based information retrieval tool for analysing the impact of AI on EEG studies". Master's Thesis in ICT for Internet and Multimedia. University of Padova, 2025.

[59] Giulia Cisotto and Davide Chicco. "Ten quick tips for clinical electroencephalographic (EEG) data acquisition and signal processing". In: *PeerJ Computer Science* 7 (2021), e506. DOI: 10.7717/peerj-cs.506.

[60] Giulia Cisotto et al. "Feature Importance via Shapley Values in Random Forests for Sleep Apnea and Hypopnea Detection". In: *Proc. ITAL-IA*. 2025.

[61] Zijiao Ji et al. "Survey of hallucination in natural language generation". In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[62] Anna V Guglielmi et al. “Frequency-dependent functional connectivity of brain networks at resting-state”. In: *Proc. Biomed. Engin. Int. Conf. (BMEiCON)*. 2022, pp. 1–5.

[63] G. H. Klem et al. “The ten-twenty electrode system of the International Federation”. In: *Electroencephalography and Clinical Neurophysiology* 52 (1999), pp. 3–6.

[64] Haijie Liu et al. “An EEG motor imagery dataset for brain computer interface in acute stroke patients”. In: *Scientific Data* 10.1 (2023), pp. 1–13. DOI: 10.1038/s41597-023-02787-8.

[65] Gert Pfurtscheller and Fernando H. Lopes da Silva. “Event-related EEG/MEG synchronization and desynchronization: Basic principles”. In: *Clinical Neurophysiology* 110 (1999), pp. 1842–1857.

[66] K. K. Ang et al. “A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke”. In: *Clinical EEG and Neuroscience* 46 (2015), pp. 310–320.

[67] Clemens Brunner et al. *BCI Competition 2008 – Graz Data Set A*. Graz University of Technology. 2008.

[68] Hyeon Cho et al. “EEG datasets for motor imagery brain-computer interface”. In: *GigaScience* 6.7 (2017), pp. 1–8.

[69] Min-Ho Lee et al. “EEG dataset and OpenBMI toolbox for three BCI paradigms”. In: *GigaScience* 8.5 (2019), giz002.

[70] Jon Ander Urigüen and Begonya Garcia-Zapirain. “EEG artifact removal: State-of-the-art and guidelines”. In: *Journal of Neural Engineering* 12.3 (2015), p. 031001.

[71] Alessandro Liberati et al. “The PRISMA statement for reporting systematic reviews and meta-analyses”. In: *PLOS Medicine* 6.7 (2009), e1000097.

[72] Julian P. T. Higgins et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd. Wiley-Blackwell, 2019.

[73] J. Richard Landis and Gary G. Koch. “The measurement of observer agreement for categorical data”. In: *Biometrics* 33.1 (1977), pp. 159–174.

[74] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems* (2020).

[75] Shuang Wang et al. “Scientific discovery in the age of artificial intelligence”. In: *Nature* 620 (2023), pp. 47–60.

[76] Benjamin Blankertz et al. “Optimizing spatial filters for robust EEG single-trial analysis”. In: *IEEE Signal Processing Magazine* 25.1 (2008), pp. 41–56.

[77] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (1995), pp. 273–297.

[78] Giulia Cisotto et al. “Machine Learning-Based Classification of Cognitive Workload via In-Ear EEG”. In: *2025 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2025, pp. 1–6.

[79] Giulia Cisotto et al. “Machine Learning Based Assessment of Cognitive Performance Under Sleep Deprivation”. In: *2025 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2025, pp. 1–6.

[80] Giulia Cisotto et al. “Classification of grasping tasks based on EEG-EMG coherence”. In: *Proc. IEEE Int. Conf. e-Health Netw. Appl. Serv. (Healthcom)*. 2018, pp. 1–6.

[81] Giulia Cisotto et al. “Joint compression of EEG and EMG signals for wireless biometrics”. In: *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. 2018, pp. 1–6.

[82] Daniele Scapin et al. “Shapley value as an aid to biomedical machine learning: a heart disease dataset analysis”. In: *Proc. IEEE Int. Symp. Cluster Cloud Internet Comput. (CCGrid)*. 2022, pp. 933–939.

[83] Elisa Borella et al. “Effective sensor selection for human activity recognition via Shapley value”. In: *Proc. IEEE Int. Workshop Metrology Living Environ. (MetroLivEnv)*. 2024, pp. 22–27.

[84] Angelique C. Paulk et al. “Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals”. In: *Scientific Data* 7 (2020), p. 180. DOI: 10.1038/s41597-020-0532-5.

[85] Patryk Jurczak, Tomasz Wozniak, and Malgorzata Jedrzejewska-Szczerska. “Improved Manual Annotation of EEG Signals through Convolutional Neural Network Guidance”. In: *eNeuro* 9.5 (2022), ENEURO.0160–22.2022. DOI: 10.1523/ENEURO.0160-22.2022.

[86] Muhammad Sudipto Siam Dip et al. "Optimized Feature Selection and Neural Network-Based Classification of Motor Imagery Using EEG Signals". In: *arXiv preprint arXiv:2504.03984* (2025).

[87] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, et al. "EEG-TCNet: An Accurate Temporal Convolutional Network for Embedded Motor-Imagery Brain-Machine Interfaces". In: *arXiv preprint arXiv:2006.00622* (2020).

[88] B. Kea et al. "Consensus development for healthcare professionals". In: *Journal of Emergency Medicine* 47.3 (2014), pp. 364–370.

[89] M. H. J. van de Pol et al. "Expert and patient consensus on a dynamic model for shared decision making". In: *Patient Education and Counseling* 99.6 (2016), pp. 1067–1074.

[90] Selim R. Benbadis et al. "Interobserver reliability in EEG interpretation". In: *Epilepsy Behavior* 15.3 (2009), pp. 361–365. DOI: 10.1016/j.yebeh.2009.05.014.

[91] W. O. Tatum et al. "Clinical EEG consensus and standardization". In: *Clinical Neurophysiology* 132.1 (2021), pp. 3–9. DOI: 10.1016/j.clinph.2020.10.003.

[92] Sara Rosenthal et al. "Multidisciplinary decision-making in healthcare". In: *Health Services Research* 52.5 (2017), pp. 1990–2008. DOI: 10.1111/1475-6773.12555.

[93] Christoph M. Michel and Denis Brunet. "EEG source imaging and methodological consensus". In: *NeuroImage* 199 (2019), pp. 133–146. DOI: 10.1016/j.neuroimage.2019.05.038.

[94] David A. Cohn et al. "Human-in-the-loop: Interactive and interpretable machine learning". In: *Communications of the ACM* 64.8 (2021), pp. 62–71. DOI: 10.1145/3454122.

[95] Mark P. Sendak et al. "A path for translation of machine learning products into healthcare delivery". In: *NPJ Digital Medicine* 3.1 (2020), pp. 1–4. DOI: 10.1038/s41746-020-0226-0.

[96] K. A. Robbins et al. "How reproducible are EEG data analyses? A comparison of workflows". In: *Frontiers in Neuroscience* 14 (2020), p. 113. DOI: 10.3389/fnins.2020.00113.

[97] Marc Jeannerod. "Neural simulation of action: a unifying mechanism for motor cognition". In: *NeuroImage* 14.1 (2001), S103–S109. DOI: 10.1006/nimg.2001.0832.

[98] Dennis J. McFarland et al. "Mu and beta rhythm topographies during motor imagery and actual movements". In: *Brain Topography* 12.3 (2000), pp. 177–186. DOI: 10.1023/A:1023437823106.

[99] Jonathan R. Wolpaw et al. "Brain-computer interfaces for communication and control". In: *Clinical Neurophysiology* 113.6 (2002), pp. 767–791. DOI: 10.1016/S1388-2457(02)00057-3.

[100] Christa Neuper, Michael Wörtz, and Gert Pfurtscheller. "ERD/ERS patterns reflecting sensorimotor activation and deactivation". In: *Progress in Brain Research* 159 (2006), pp. 211–222. DOI: 10.1016/S0079-6123(06)59014-4.

[101] Pauline Dreyer et al. "A large EEG database with users profile information for motor imagery brain-computer interface research". In: *Scientific Data* 10.1 (2023), pp. 1–12. DOI: 10.1038/s41597-023-02445-z.

[102] Banghua Yang et al. "A multi-day and high-quality EEG dataset for motor imagery brain-computer interface". In: *Scientific Data* 12.1 (2025), pp. 1–15. DOI: 10.1038/s41597-025-04826-y.

[103] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596.

[104] Tongshuang Wu, Michael Terry, and Carrie J. Cai. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts". In: *CHI Conference on Human Factors in Computing Systems* (2022).

[105] Saleema Amershi et al. "Power to the People: The Role of Humans in Interactive Machine Learning". In: *AI Magazine* 35.4 (2014), pp. 105–120.

[106] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer learning in biomedical natural language processing". In: *Briefings in Bioinformatics* 20.2 (2019), pp. 806–820.

[107] P. B. Jensen, L. J. Jensen, and S. Brunak. "Mining electronic health records: towards better research applications and clinical care". In: *Nature Reviews Genetics* (2022). DOI: 10.1038/s41576-022-00475-4.

[108] Mohammad Moradi and Matthias Samwald. "Evaluating GPT models for structured information extraction in biomedical research: The role of prompt engineering and rule-based constraints". In: *Journal of Biomedical Informatics* 141 (2023), p. 104410. DOI: 10.1016/j.jbi.2023.104410.

[109] Wendy W. Chapman et al. "Overcoming barriers to NLP for clinical text: the role of shared tasks and expert-annotated corpora". In: *Journal of the American Medical Informatics Association* 27.8 (2020), pp. 1239–1244. DOI: 10.1093/jamia/ocaa078.

[110] Fabien Lotte et al. "A review of classification algorithms for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 4.2 (2007), R1–R13.

[111] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011), p. 27. DOI: 10.1145/1961189.1961199.

[112] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[113] Diederik P Kingma and Jimmy Ba. "Adam: a method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[114] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874.

[115] Jesse Davis and Mark Goadrich. "The relationship between precision-recall and ROC curves". In: (2006), pp. 233–240.

[116] Steven Lemm et al. "Introduction to machine learning for brain imaging". In: *NeuroImage* 56.2 (2011), pp. 387–399.

[117] Kevin A. Hallgren. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". In: *Tutorials in Quantitative Methods for Psychology* 8.1 (2012), pp. 23–34.

[118] A. Zorzetto. "Explainable AI for EEG Signals Analysis". Masters Thesis, supervised by G. Cisotto. MA thesis. University of Padova, Department of Information Engineering, 2022.

[119] A. Gramfort et al. "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroscience* 15 (2021), p. 720. DOI: 10.3389/fnins.2021.720.

[120] Yijun Wang, Xiaorong Gao, and Shangkai Gao. "A study on EEG classification using common spatial patterns and support vector machines". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14.3 (2006), pp. 369–372. DOI: 10.1109/TNSRE.2006.875576.

[121] Giulia Cisotto et al. "CNN-based Approaches For Cross-Subject Classification in Motor Imagery: From the State-of-the-Art to DynamicNet". In: *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2021, pp. 1–8. DOI: 10.1109/CIBCB49929.2021.9562821.

[122] Fabrizio De Vico Fallani et al. "Multiscale Topological Properties of Functional Brain Networks during Motor Imagery after Stroke". In: *PLoS ONE* 8.2 (2013), e56915. DOI: 10.1371/journal.pone.0056915.

[123] Kai Keng Ang et al. "A Clinical Study of Motor Imagery-Based Brain–Computer Interface for Upper Limb Robotic Rehabilitation". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23.2 (2015), pp. 362–372. DOI: 10.1109/TNSRE.2014.2329155.

[124] X. Liu et al. "A Hybrid Convolutional Neural Network with Attention-Based Feature Extraction for Motor Imagery EEG Classification". In: *Neuroscience* 530 (2025), pp. 120–132. DOI: 10.1016/j.neuroscience.2025.01.015.

[125] Y. Li et al. "Automatic Feature Extraction and Fusion Recognition of Motor Imagery EEG Using Multilevel Multiscale Convolutional Neural Network". In: *Biomedical Signal Processing and Control* 68 (2021), p. 102643. DOI: 10.1016/j.bspc.2021.102643.

[126] Hubert Cecotti and Axel Gräser. "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. DOI: 10.1109/TPAMI.2010.125.

[127] Christoph Guger, Herbert Ramoser, and Gert Pfurtscheller. "Real-time EEG analysis with subject-specific spatial patterns for a brain–computer interface". In: *IEEE Transactions on Rehabilitation Engineering* 8.4 (2000), pp. 447–456. DOI: 10.1109/86.895947.

[128] Huanhuan Lu et al. "Regularized common spatial pattern with aggregation for EEG classification in small-sample setting". In: *IEEE Transactions on Biomedical Engineering* 57.12 (2009), pp. 2936–2946. DOI: 10.1109/TBME.2010.2082540.

[129] Zhiyuan Tang, Chen Li, and Shufang Sun. "Single-trial EEG classification of motor imagery using deep convolutional neural networks". In: *Optik* 130 (2017), pp. 11–18. DOI: 10.1016/j.ijleo.2016.10.117.

[130] Jonathan M. Cassidy et al. "Low-frequency oscillations are a biomarker of injury and recovery after stroke". In: *Stroke* 51.5 (2020), pp. 1442–1450.

[131] Marco A. Cervera et al. "Brain–Computer Interfaces for Post-Stroke Motor Rehabilitation: A Meta-Analysis". In: *Annals of Clinical and Translational Neurology* 5.5 (2018), pp. 651–663. DOI: 10.1002/acn3.544.

[132] Vinay Jayaram and Alexandre Barachant. "Transfer learning in brain-computer interfaces". In: *IEEE Computational Intelligence Magazine* 11.1 (2016), pp. 20–31.

[133] Wojciech Samek, Florent Meinecke, and Klaus-Robert Müller. "Transferring subspaces between subjects in braincomputer interfacing". In: *IEEE Transactions on Biomedical Engineering* 60.8 (2013), pp. 2289–2298.

Declaration

During the preparation of this thesis, the author made use of OpenAI ChatGPT as a digital assistant to support tasks such as language editing, structural refinement, and latex formatting.

All AI-assisted content was carefully reviewed and validated by the author. The scientific content, methodology, analysis, and conclusions presented in this thesis are entirely the author's own work. The author takes full responsibility for the integrity and accuracy of all content.

Acknowledgments

This work would not have been possible without the guidance and support of my supervisors. I am deeply grateful to Prof. Giulia Cisotto for her constant presence from the very beginning of this journey for her mentorship, her patience, and the invaluable knowledge she has shared with me throughout this process. Equally, I owe a great deal to Prof. Leonardo Badia, who has been by my side from the moment I first considered pursuing a Master's degree in Italy all the way through to the final day. His guidance has been not only academically formative but deeply meaningful at every critical turning point of this experience.

On a personal note, I would like to express my heartfelt gratitude to my family, whose unconditional love and encouragement have carried me to where I stand today. To my mother Songül Petek, my father İmam Petek, and my sister Duygu Dilara Eldemir thank you for always believing in me and celebrating every step of my journey. I also wish to honour the memory of my grandmother, who, though no longer with us, played an irreplaceable role in shaping who I have become.I would also like to express my sincere thanks to my dear friend Süleyman Eşsiz, for always believing in me and supporting me throughout this journey.

My deepest thanks go to my husband, Barış Can Anar, my greatest supporter for never doubting me, for standing beside me through every challenge, and for making this path a shared one.

And to the little miracle growing quietly within me you are already my greatest achievement. Together, we have many beautiful moments and milestones yet to share, and I cannot wait to sign them all with you by my side.