Signal Processing Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
https://sps.ewi.tudelft.nl/

SPS-2025

M.Sc. Thesis

Automated Classification of Photic Stimulation EEG Responses for Improved Epilepsy Diagnosis

Giacomo Zanardini

Abstract

Epilepsy is a common neurological disorder, but its diagnosis remains difficult when screening EEGs lack interictal epileptiform discharges (IEDs). Intermittent photic stimulation (IPS) can reveal abnormal responses associated with epilepsy; however, its clinical interpretation is often subjective, inconsistent, and sometimes inconclusive. This thesis explores the automatic classification of EEG responses to IPS using machine learning to improve diagnostic accuracy and reliability.

Two datasets are analysed: the Temple University Hospital (TUH) Epilepsy Corpus and clinical recordings from Erasmus MC. A structured pipeline is developed, comprising preprocessing, feature extraction across temporal, spectral, wavelet, and connectivity domains, and classification with interpretable models such as XGBoost and ensemble approaches. To ensure robust generalization, leave-one-subject-out cross-validation is employed.

This work demonstrates that IPS EEG segments contain informative features capable of distinguishing epileptic from non-epileptic patients, even in the absence of IEDs, thereby aiding early diagnosis and reducing the risk of misdiagnosis. Furthermore, the use of explainability tools highlights candidate electrophysiological markers, providing valuable insights and suggesting new hypotheses for future investigation.

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Automated Classification of Photic Stimulation EEG Responses for Improved Epilepsy Diagnosis

Thesis

submitted in partial fulfilment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Giacomo Zanardini
born in Brescia, Italy

This work was performed in:

Signal Processing Systems Group
Department of Microelectronics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

TU Delft

Delft University of Technology

Delft University of Technology
Department of
Microelectronics

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Automated Classification of Photic Stimulation EEG Responses for Improved Epilepsy Diagnosis"** by **Giacomo Zanardini** in partial fulfilment of the requirements for the degree of **Master of Science**.

Dated: 30/09/2025

Chairman:

prof. dr. ir. Justin Dauwels

Advisor:

prof. dr. ir. Justin Dauwels

Committee Members:

prof. dr. ir. Justin Dauwels

dr. Robert van den Berg

dr. Nergis Tömen

Abstract

Epilepsy is a common neurological disorder, but its diagnosis remains difficult when screening EEGs lack interictal epileptiform discharges (IEDs). Intermittent photic stimulation (IPS) can reveal abnormal responses associated with epilepsy; however, its clinical interpretation is often subjective, inconsistent, and sometimes inconclusive. This thesis explores the automatic classification of EEG responses to IPS using machine learning to improve diagnostic accuracy and reliability.

Two datasets are analysed: the Temple University Hospital (TUH) Epilepsy Corpus and clinical recordings from Erasmus MC. A structured pipeline is developed, comprising preprocessing, feature extraction across temporal, spectral, wavelet, and connectivity domains, and classification with interpretable models such as XGBoost and ensemble approaches. To ensure robust generalization, leave-one-subject-out cross-validation is employed.

This work demonstrates that IPS EEG segments contain informative features capable of distinguishing epileptic from non-epileptic patients, even in the absence of IEDs, thereby aiding early diagnosis and reducing the risk of misdiagnosis. Furthermore, the use of explainability tools highlights candidate electrophysiological markers, providing valuable insights and suggesting new hypotheses for future investigation.

Acknowledgments

I would first like to express my sincere gratitude to my supervisor, prof. dr. ir. Justin Dauwels, for his invaluable guidance, feedback, and continuous support throughout the development of this thesis. His expertise and encouragement were essential in shaping the direction of this work. I am equally grateful to my advisor, Dr. Robert van den Berg, whose clinical insights consistently grounded the research in real-world scenarios and ensured its relevance to practical applications. I would also like to thank Dr. Nergis Tömen for kindly agreeing to be part of the committee and for dedicating her time to evaluate this work.

On a more personal note, I want to thank my family for their endless patience, love, and encouragement throughout this process. I am also deeply grateful to the people close to me: friends and colleagues, in the SPS section and beyond, in Delft, Brescia, and across the globe, for the laughs, discussions, and much needed distractions along the way. Your support made this journey not only possible but also truly enjoyable. No names mentioned, but you know who you are.

Giacomo Zanardini
Delft, The Netherlands
30/09/2025

Contents

List of Figures

List of Tables

Nomenclature

AUC	Area Under the Curve
EMC	Erasmus Medical Center
FN	False Negatives
FP	False Positives
IED	Interictal Epileptiform Discharge
LOSO	Leave One Subject Out
ML	Machine Learning
TN	True Negatives
TP	True Positives
TUH	Temple university Hospital
xAI	Explainable Artificial Intelligence
XGB	XGBoost/ Extreme Gradient Boosting

Introduction 1

1.1 Problem Statement

Epilepsy is a chronic neurological disorder characterized by recurrent, unprovoked seizures due to abnormal brain activity. It affects approximately 50 million people worldwide and significantly reduces their quality of life, as routine activities such as driving can often become unfeasible due to the constant risk of seizures [1]. The International League Against Epilepsy (ILAE) defines epilepsy as either having experienced two unprovoked seizures more than 24 hours apart or having a greater than 60% risk of seizure recurrence after a single unprovoked seizure [2]. While the first criterion is relatively straightforward, assessing recurrence risk introduces considerable complexity, especially in patients who have experienced only one seizure.

A key diagnostic tool is the electroencephalogram (EEG), which measures aggregate neuronal activity via electrodes placed on the scalp. EEGs are used to detect seizures and Interictal Epileptiform Discharges (IEDs) i.e. abnormal brain wave patterns that may indicate epileptogenic regions of the brain [3]. When such markers are observed, a diagnosis of epilepsy may be made, particularly under the second criterion of the ILAE definition. However, IEDs are not consistently present in all epileptic patients, and their interpretation can be difficult [4]. Consequently, many patients who are later confirmed to have epilepsy through clinical follow-up cannot initially be diagnosed on the basis of EEG findings alone.

Intermittent photic stimulation (IPS) is another non-invasive activation procedure routinely added to EEG recordings. During IPS the patient is exposed to stroboscopic flashes whose frequency is gradually swept, while brain responses are monitored in real time. Detecting epileptic responses (absent during baseline EEG) increases the likelihood of photosensitivity, thus refines the recurrence estimates after a first seizure.

To further enhance diagnostic yield, some patients undergo sleep deprivation before a second EEG, as this state can reduce inhibitory brain activity and increase the likelihood of detecting IEDs or inducing seizure events [5]. While this approach can improve detection rates, a significant portion of epilepsy cases remains undiagnosed after both standard and sleep-deprived EEGs. As a result, many patients are placed under a wait-and-see policy, contributing to prolonged diagnostic uncertainty and delayed treatment initiation [6]. This uncertainty is exacerbated by studies reporting that up to 50% of individuals who have experienced a first seizure will have a recurrence within a few years [7].

In the Netherlands and other countries, the standard diagnostic process often combines EEG with magnetic resonance imaging (MRI) to assess seizure risk and guide treatment decisions [8]. However, current methods are not always sufficient. Therefore, expanding diagnostic capabilities, particularly by integrating advanced computational approaches, could allow more patients to receive accurate and timely diagnoses.

In response to this need, recent research has increasingly applied machine learning (ML) and deep learning techniques to EEG analysis, with the goal of identifying epileptic patterns and improving diagnostic precision. This study contributes to this growing field by presenting a data-processing pipeline for automated epilepsy detection. Building upon methodologies from prior works [9–11], the pipeline includes four key stages: data acquisition, pre-processing, feature extraction, and classification. Raw EEG signals are obtained from clinical sources, cleaned through various signal processing techniques, analyzed for significant features, and ultimately classified to distinguish between epileptic and non-epileptic patients.

By leveraging such structured and automated approaches, this research aims to reduce diagnostic latency, improve treatment outcomes, and contribute to the broader objective of advancing epilepsy care and neurological research.

1.2 Previous Works

This research extends upon the findings of Thangavel et al. [9], who indicated the potential diagnostic value of EEG signals without interictal epileptiform discharges (IEDs) in epilepsy detection. Building on this foundation, Y. Mirwani's MSc thesis validated these findings by replicating the study using a dataset from Erasmus Medical Centre (EMC), Rotterdam [10]. Mirwani's research involved identifying optimal hyperparameters specific to each EEG feature set. His methodology included segmenting EEG montages into epochs, computing features individually for each epoch, and subsequently combining these epoch-level results using aggregation methods such as mean, median, and standard deviation. The best-performing EEG feature sets, based on Area Under the Curve (AUC) metrics derived from Leave-One-Subject-Out (LOSO) cross-validation, serve as the starting point for this current investigation.

The central aim of the current study is to evaluate whether analyzing data from IPS trials can enhance the diagnostic capability of existing ML models in distinguishing epilepsy from EEG signals without IEDs. Furthermore, the research explores the optimal interpretation of ML model predictions to ensure practical applicability within clinical settings.

1.3 Research Questions

This thesis investigates diverse strategies for feature extraction, ML based classification, as well as rigorous evaluation methods within EEG signal analysis. Specifically, this research addresses the following questions:

- **RQ1:** Is it possible to diagnose epilepsy by analyzing EEG segments recorded during IPS?
- **RQ2:** How do proposed classification methods compare across different EEG datasets, namely TUH versus EMC?
- **RQ3:** What are candidate electro-physiological markers for epilepsy?

These research questions guide the thesis, aiming to enhance the accuracy and reliability of EEG classification through sophisticated feature extraction, advanced machine learning approaches, and comprehensive evaluation methodologies.

1.4 Outline

The remainder of this thesis is structured as follows. Chapter 2 reviews the background on epilepsy diagnosis with machine learning, introduces intermittent photic stimulation, and highlights research gaps while also describing EEG montages, features, and evaluation methods. Chapter 3 presents the datasets and feature sets used in the study, while Chapter 4 details the methodological pipeline including preprocessing, cross-validation, model training, and ensemble strategies. Chapter 5 reports statistical analyses of the extracted features, followed by Chapters 6 and 7 which present the experimental results on the TUH and EMC datasets respectively. Chapter 8 discusses the findings in relation to existing literature and clinical practice, and Chapter 9 concludes the thesis by summarizing contributions, limitations, and directions for future work. Supplementary analyses and extended results are provided in the appendices.

Background 2

This chapter presents an up-to-date overview of the various techniques and methodologies utilized in this thesis. It begins by outlining the current state of knowledge regarding machine learning approaches to epilepsy diagnosis, allowing to highlight the knowledge gaps that reinforce the motivation for this research. The relevance of intermittent photic stimulation (IPS) is then introduced, followed by a discussion of EEG processing methods such as reference montages. The chapter continues with an explanation of feature extraction techniques, then details the machine learning algorithms employed and the corresponding evaluation metrics. Finally, the chapter introduces the statistical tests used in the analysis.

2.1 Epilepsy Diagnosis and Machine Learning

Current machine-learning (ML) research in EEG-based epilepsy largely focuses on detecting specific events, notably seizures or interictal epileptiform discharges, rather than making a definitive epilepsy diagnosis. Numerous studies have proposed pipelines to automatically detect seizures [12, 13] or to detect IEDs [14]. In practice, EEG analysis is indeed essential for the diagnostic workup of epilepsy, but diagnosing the disorder involves more than finding EEG abnormalities. A single routine EEG often fails to capture any seizures or epileptiform activity: in fact, up to 50% of people with proven epilepsy show a normal EEG between seizures [15]. EEG waveforms can sometimes appear deceptively normal or be obscured by artifacts, and even pathological patterns can overlap with benign variants.

Because of this, epilepsy diagnosis relies on a combination of clinical history and other tests (e.g. video-EEG monitoring or MRI), rather than EEG alone [16]. This explains why most ML studies tackle the subproblems of seizure or IED detection (which can support a diagnosis [17]) instead of attempting a standalone "epilepsy yes/no" classification from EEG. Detecting hallmark events in EEG is a more tractable and data-driven task, whereas a direct ML-based diagnosis of epilepsy is far more complex due to the intermittent nature of seizures and the need for clinical context beyond the EEG signals.

A critical concern in applying advanced ML to epilepsy is the interpretability of the models. Many high-performing models are deep neural networks that act as 'black boxes', making decisions that are not easily explainable to clinicians. This lack of transparency is problematic in a clinical setting where doctors must understand and trust the basis of a diagnosis. While deep learning (e.g. Convolutional Neural Networks or Recurrent Neural Networks, CNNs and RNNs, respectively) has achieved outstanding accuracy in seizure detection tasks [18, 19], these models offer little insight into what EEG features drive their predictions.

In contrast, classical ML approaches using hand-crafted features (e.g. spectral power, entropy) with algorithms like Random Forests or SVMs can be more interpretable: one can examine feature importance or decision rules, which align better with clinical reasoning. For example, a recent review noted that traditional ML allows development of interpretable clinical features and models, an aspect still highly valued in practice [18].

The trade-off is that simpler or constrained models may sacrifice some accuracy in exchange for transparency. There is active research into bridging this gap, such as designing deep models with built-in explanations or hybrid systems. Some works add attention mechanisms or annotate EEG channels to highlight why a deep model flags a seizure, incurring only a minor performance cost [20, 21].

Overall, improving explainability of EEG ML models is crucial: not only should an algorithm detect abnormalities, but it should also explain the patterns it finds, so clinicians can corroborate them. Only with clear interpretability will these models gain acceptance as decision support tools in epilepsy diagnosis.

An accurate evaluation of any ML system for epilepsy diagnosis hinges on two final considerations: the quality of the training data and, as discussed in the next section, the rigour of the validation protocol. A persistent pitfall is reliance on small, highly simplified datasets: most notably the single-channel Bonn University corpus, which contains recordings from just ten subjects and contrasts clear ictal segments with normal background rhythms [22]. Because the classes are so easily separable, models routinely score near-perfect accuracies, as highlighted in the systematic survey by Zendehbad et al. [23] and the earlier work of Acharya et al. [24]. Yet such results provide little insight into how well these models would perform on multi-channel, clinically realistic EEG.

Although Wong's recent review catalogues a broader suite of EEG datasets and their clinical characteristics [25], most of those collections are geared toward seizure detection rather than the more challenging goal of classifying seizure-free, inter-ictal recordings, the very scenario targeted in this project.

Even when researchers train on large, clinically realistic EEG corpora e.g. those used by Myers et al., [26], Reddy N. et al., [27] and Cao et al., [28], reported performance can still be misleading if the validation strategy is not designed to prevent data leakage. Conventional train–test splits, unless performed strictly at the subject level, allow recordings from the same patient to appear in both the training and test sets. The model then recognises patient-specific patterns rather than learning disease-relevant features, inflating accuracy and other metrics. Kunjan [29] and Tougui [30] provide empirical evidence of this effect: they show that even seemingly robust k-fold cross-validation remains overly optimistic for EEG because folds are typically drawn at the recording level, not the patient level. Both authors argue that, for neuroscience applications where inter-subject variability is high, leave-one-subject-out (LOSO) cross-validation or an external hold-out cohort is essential. These stricter protocols yield lower but far more realistic estimates of generalisation performance and are therefore the recommended standard for ML-based epilepsy diagnosis.

Table 2.1: Prior work in EEG-based epilepsy ML: datasets, validation (LOSO/LOIO/k-fold), and headline metrics.

Author	Model	Dataset	CV	Performance	Note
Thangavel et al. [31]	1/2D CNN	TUH, Others	LOSO, LOIO	LOIO: AUC=0.839, BAC=79.5% LOSO: AUC=0.856 BAC 78.1%	IED Based
Thangavel et al. [9]	XGBoost	TUH, Others	LOSO, LOIO	TUH with IEDs: AUC=0.790 BAC=71.7% TUH w/o IEDs: AUC=0.630 BAC=57.3%	Uses IED features
Myers et al. [26]	Logistic Regression	Others	10-Fold	AUC=0.940 Accuracy=90.4%	
Mirwani [10]	XGBoost, CNN	TUH, EMC	LOSO	TUH AUC=0.76 BAC=72% EMC AUC=0.7 BAC=67%	N/A
Van der Kleij [11]	XGBoost	EMC	LOSO	AUC=0.87	Incorrect Ensembling

2.2 Baseline EEG and Intermittent Photic Stimulation

Intermittent photic stimulation (IPS) enhances standard EEG by delivering a sequence of stroboscopic light flashes, typically ranging from 1 to 60 Hz, with the patient assessed under both eyes-open and eyes-closed conditions. In healthy individuals, IPS generally has no effect or produces a photic driving response (i.e. synchronization of the posterior alpha rhythm and its harmonics with the flashing light) without generating epileptiform activity. However, in certain susceptible people, IPS can trigger a photo-paroxysmal response (PPR), which is characterized by distinctive interictal epileptiform discharges (IEDs).

The presence of a PPR is especially significant, as it is strongly associated with epilepsy [32]. Its detection increases diagnostic confidence even when spontaneous IEDs are absent on the baseline EEG, and assists in clinical decision-making by identifying individuals at risk of visually-induced seizures. Not only is the PPR a frequent finding in such cases, but it is also highly diagnostically informative: its presence substantially raises the likelihood of epilepsy, even if no spontaneous IEDs are recorded during the baseline EEG.

In practice, the diagnostic yield of IPS depends on the specific stimulation protocol. To reduce false negatives and standardize results across centers, a European protocol was developed [33]. This protocol details the required light intensity, duration of stimuli, and an ascending–descending frequency sweep (using the following sequence: 1–2–6–8–9–10–13–15–18–20–23–25–30–40–50–60 Hz. If a generalized response is seen at any frequency, skip the remaining frequencies and proceed to 60 Hz, then sweep downward: 60, 50, 40, 30, 25 Hz, etc. until a PPR is obtained again). It also mandates recording under three eye conditions: eyes open, eyes closed, and during partial eye closure.

In summary, IPS provides a quick, standardized, and low-burden way to evaluate epileptiform networks. A normal response can reassure clinicians when the baseline EEG is inconclusive, while detection of a PPR strongly supports a diagnosis of epilepsy. Because IPS can yield critical diagnostic information in just a couple of minutes, it is now considered an essential part of first-line EEG evaluation for suspected seizure disorders in most European centers.

2.3 Research Gaps

Existing studies in EEG-based epilepsy diagnosis face several key limitations:

1. A scarcity of methods for classifying epilepsy from interictal EEG segments that lack obvious interictal epileptiform discharges (IEDs), making realistic but subtle cases difficult and often overlooked.

2. Insufficient use of rigorous cross-subject validation (e.g., leave-one-subject-out), with many works relying on k-fold evaluation that mixes data from the same patients and thus inflates performance on seen subjects.

3. An over-reliance on overly simplified or non-representative datasets (e.g., the Bonn corpus of isolated single-channel EEG segments) which do not capture the heterogeneity of routine clinical EEG and can yield over-optimistic results.

4. Poor model interpretability, especially in deep learning approaches that often function as black-boxes and thereby weaken clinicians' trust.

The present study addresses these gaps by targeting epilepsy detection in IED-free interictal EEG, leveraging a large multi-center clinical EEG dataset, rigorously evaluating generalizability with a leave-one-subject-out validation scheme, and favoring interpretable hand-crafted EEG features combined with transparent machine learning models.

2.4 Montage, References, Segment Lengths and Statistical Combiners

In order extract information from the EEG recordings, several features are extracted from the signals using sliding windows of varying lengths, ranging from 1 to 60[1] seconds. For each window, features are computed and then statistically summarized using several combiners: mean, median, standard deviation, skewness, and kurtosis. This statistical aggregation serves two purposes: it reduces the dimensionality of the resulting data and allows the analysis to capture both short-term and long-term temporal dynamics present in the data.

Montages refer to the way EEG electrodes are arranged on the scalp. Among the standard placement schemes we can find the International 10–20 system [34]. Beyond this spatial layout, one can choose different scalp sites as voltage references; configurations that specify such reference points are called referential montages. Because all data in this study were recorded with the 10–20 system, from now on, the term montage will denote a referential montage.

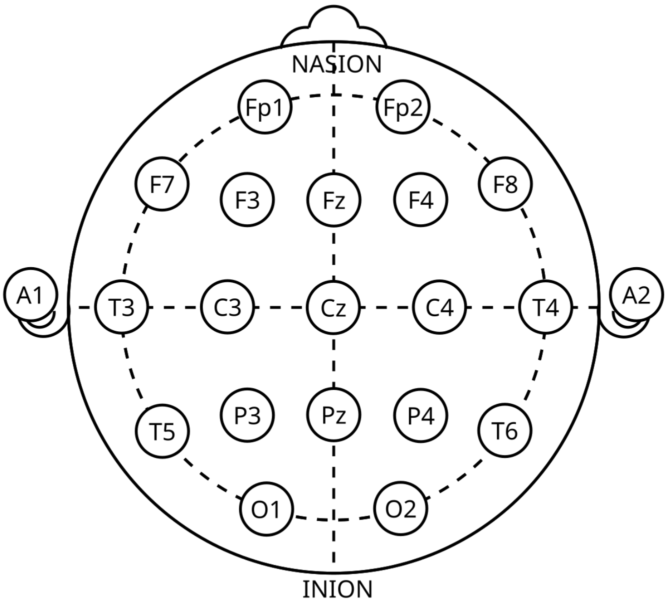

Figure 2.1: Electrode montage for the international 10–20 system [34] with key regions (frontal, temporal, central, parietal, occipital) indicated; used for all recordings in this thesis.

[1] Segment lengths of 20 and 60 seconds s are only used for whole-trial recordings.

Table 2.2 summarizes the montages, segment lengths, and statistical combiners used in the analysis. Specifically, four types of montages are examined: Common Average Reference (CAR), Cz Referential, Longitudinal Bipolar (also known as "Double Banana") [35], and Laplacian [36]. Feature extraction is performed over window lengths of 1, 2, 5, 10, 20[1] and 60[1] seconds, ensuring both fine-grained and more global patterns are considered. The use of multiple statistical combiners (mean, median, standard deviation, skewness, and kurtosis) further enriches the characterization of the data by summarizing different aspects of the feature distributions within each segment.

Table 2.2: Analysis design: referential montages, segment lengths (1–10 s; 20/60 s for whole trials), and five statistical combiners used to summarize windowed features.

Montages	**Segment lengths [s]**	**Statistical combiners**
	1	Mean
CAR	2	Median
Cz	5	Standard Deviation
Laplacian	10	Skewness
BipolarDB	20[1]	Kurtosis
	60[1]	

2.4.1 CAR

The CAR montage is a reference-free method that avoids complications related to using a physical reference [37]. In this technique, each electrode's potential is measured relative to the average of all electrodes, calculated as:

$$V_i^{\text{CAR}} = V_i^{\text{ER}} - \frac{1}{n}\sum_{j=1}^{n} V_j^{\text{ER}}, \tag{2.1}$$

where V_i^{ER} is the potential between the i^{th} electrode and the reference, and n is the number of electrodes in the montage [37].

2.4.2 Cz

In the Cz referential montage, the Cz electrode serves as a common reference for all channels. It is particularly useful for detecting widespread brain abnormalities, and differences between the hemispheres, though it is less effective for localizing focal activity [38]. The Cz-referenced potential is given by:

$$V_i^{\text{Cz}} = V_i^{\text{ER}} - V^{\text{Cz}}. \tag{2.2}$$

2.4.3 Bipolar DB

The Longitudinal Bipolar montage consists of a series of electrodes connected to their neighbouring electrodes in a chain-like pattern. It is favoured for its versatility and is commonly arranged from anterior to posterior across para-sagittal and temporal regions [35].

Table 2.3: Longitudinal bipolar electrode pairs mapped to approximate brain regions used in feature computation.

Electrode Pair	**Brain Region**
FP1 - F7	Frontal - Temporal
F7 - T3	Temporal
T3 - T5	Temporal - Parietal
T5 - O1	Parietal - Occipital
FP1 - F3	Frontal
F3 - C3	Frontal - Central
C3 - P3	Central - Parietal
P3 - O1	Parietal - Occipital
FP2 - F8	Frontal - Temporal
F8 - T4	Temporal
T4 - T6	Temporal - Parietal
T6 - O2	Parietal - Occipital
FZ - CZ	Frontal - Central
CZ - PZ	Central - Parietal
FP2 - F4	Frontal
F4 - C4	Frontal - Central
C4 - P4	Central - Parietal
P4 - O2	Parietal - Occipital

2.4.4 Laplacian

The Laplacian montage addresses the issue of referential contamination—where an outlier electrode skews the average—by using only the nearest surrounding electrodes as the reference. This technique is particularly effective for identifying focal brain abnormalities. According to Syam et al., Laplacian outperformed CAR in their brain abnormality classification study [36]. The Laplacian potential is calculated as:

$$V_i^{\text{Lap}} = V_i^{\text{ER}} - \frac{1}{n} \sum_{j \in S_i} V_j^{\text{ER}}, \tag{2.3}$$

where S_i represents the set of surrounding electrodes for the i^{th} electrode and j is a member of S_i.

Table 2.4: Laplacian montage neighbourhoods: reference sets per electrode and associated brain regions.

Electrode	**Reference (Average of Neighbours)**	**Brain Region**
FP1	F7, FP2, F3	Frontal
F3	FP1, F7, FZ, C3	Frontal
C3	F3, T3, CZ, P3	Central
P3	C3, T5, PZ, O1	Parietal
F7	FP1, F3, T3	Frontal-Temporal
T3	C3, F7, T5	Temporal
T5	T3, P3, O1	Temporal-Parietal
O1	T5, P3, O2	Occipital
FZ	F3, FP2, C3, F4, CZ	Frontal
CZ	C3, FZ, PZ, C4	Central
PZ	O2, P3, P4, O1, CZ	Parietal
FP2	FP1, F8, F4	Frontal
F4	FP2, F8, FZ, C4	Frontal
C4	F4, T4, CZ, P4	Central
P4	C4, T6, PZ, O2	Parietal
F8	FP2, F4, T4	Frontal-Temporal
T4	T6, C4, F8	Temporal
T6	O2, P4, T4	Temporal-Parietal
O2	P4, T6, O1	Occipital

2.5 Feature Extraction

Feature extraction involves converting raw data into numerical representations that retain the essential information from the original dataset. This process is fundamental to the success of machine learning (ML) applications, including those involving bio-signals for various neurological disorders. Choosing the right features is a critical step, as well-selected features can significantly enhance the accuracy of disease classification and prediction.

A feature typically captures a unique attribute, a measurable characteristic, or a functional aspect derived from a specific data segment. The goal of feature extraction is to preserve important embedded information in the signal while minimizing loss. Moreover, it facilitates dimensionality reduction, which helps decrease the need for extensive computational resources when processing large datasets.

2.5.1 Univariate Temporal measures

Univariate Temporal Measures (UTMs) are fundamental for analyzing the time-domain properties of EEG signals. These features provide valuable insights into the statistical characteristics of the signals, which play a crucial role in identifying patterns linked to epilepsy. This section outlines the specific UTMs utilized in our analysis pipeline and highlights their relevance to epilepsy diagnosis.

Let x(t) represent the EEG signal for a given channel. The following UTMs are computed:

- **Mean**: The average value of the EEG signal over a specified time window, providing a baseline level [39].

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x(i) \tag{2.4}$$

  where N is the number of samples in the window.

- **Median**: The median separates the higher half from the lower half of signal values. It is more robust to outliers than the mean, making it particularly useful for skewed distributions.

- **Standard Deviation (std)**: Measures the variability or dispersion of the signal around the mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x(i)-\mu)^2} \tag{2.5}$$

  A larger standard deviation indicates greater variability.

- **Kurtosis**: Assesses the 'tailedness' or presence of outliers in the signal distribution.

$$\text{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x(i)-\mu}{\sigma}\right)^4 - 3 \tag{2.6}$$

  where σ is the standard deviation. Higher kurtosis values indicate more frequent occurrence of sharp peaks.

- **Skewness**: Evaluates the asymmetry of the signal distribution.

$$\text{Skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x(i)-\mu}{\sigma}\right)^3 \tag{2.7}$$

  Non-zero skewness reflects deviations from symmetry.

- **Peak-to-Peak Amplitude (V_{pp})**: The difference between the maximum and minimum signal values.

$$V_{pp} = \max(x) - \min(x) \tag{2.8}$$

- **Number of Zero Crossings**: The count of times the signal crosses the zero voltage line.

- **Number of Peaks**: The number of local maxima, which may indicate spikes or artifacts, especially relevant in epileptic EEG patterns.

- **Non-linear Energy Operators (NLEO)**:

  - *Envelope Derivative (ED)*: Measures non-linearity in the signal envelope,

    $$\mathrm{ED}(x) = [x'(t)]^2 + [H[x'(t)]]^2 \tag{2.9}$$

    where H denotes the Hilbert transform [40].

  - *Teager-Kaiser (TK)*: Estimates instantaneous energy,

    $$\mathrm{TK}(x) = [x'(t)]^2 - x''(t)] \tag{2.10}$$

  Non-linear energy operators are valuable for capturing transient energy changes in EEG data [40].

- **Signal Energy**:

  - *Time Domain Energy (E_t)*: Total energy in the time domain,

    $$E_t = \log\left(\frac{1}{N}\sum_{i=1}^{N} x(i)^2\right) \tag{2.11}$$

  - *Frequency Domain Energy (E_f)*: Total energy in the frequency domain, calculated via the Discrete Fourier Transform (DFT),

    $$E_f = \log\left(\sum_{k=1}^{N} |X(k)|^2\right) \tag{2.12}$$

    where $X(k)$ is the DFT of $x(t)$ [41].

- **Shannon Entropy ($H(x)$)**: Quantifies the uncertainty or randomness of the signal,

  $$H(x) = -\sum_{i=1}^{N} p(x(i)) \log p(x(i)) \tag{2.13}$$

  where $p(x(i))$ is the probability of each signal value [42].

2.5.2 Spectral Features

Spectral features are assessed using the relative power (RP_f) calculated from five standard EEG frequency bands: delta (δ, 1–4 Hz), theta (θ, 4–8 Hz), alpha (α, 8–13 Hz), beta (β, 13–30 Hz), and gamma (γ, above 30 Hz) [9]. The relative power for each band is defined as:

$$\mathrm{RP}_f = \frac{P_f}{P_{\text{total}}}, \tag{2.14}$$

where

$$P_{\text{total}} = P_\delta + P_\theta + P_\alpha + P_\beta + P_\gamma \tag{2.15}$$

and f denotes the respective frequency band ($f \in \{\delta, \theta, \alpha, \beta, \gamma\}$). Thus, five relative power features are computed for each EEG channel segment.

2.5.3 Wavelet Features

Wavelet transforms provide a powerful approach for time-frequency analysis of EEG signals, enabling the extraction of features that capture both temporal and spectral information [43]. In this study, wavelet features are extracted using the Continuous Wavelet Transform and Discrete Wavelet Transform.

Continuous Wavelet Transform

The Continuous Wavelet Transform (CWT) of a signal $x(t)$ is defined as follows [44]:

$$W(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt \tag{2.16}$$

where a is the scale parameter, b is the translation parameter, $\psi(t)$ is the mother wavelet, and $\psi^*(t)$ denotes its complex conjugate.

For this analysis, the Morlet ('morl') wavelet is used as the mother wavelet, due to its effective balance between time and frequency localization, which is widely favored in EEG research [45]. Applying the CWT to a discrete signal $x[n]$ produces a matrix of coefficients:

$$\mathrm{CWT}_{\text{matrix}} = \begin{bmatrix} W(a_1, b_1) & W(a_1, b_2) & \cdots & W(a_1, b_N) \\ W(a_2, b_1) & W(a_2, b_2) & \cdots & W(a_2, b_N) \\ \vdots & \vdots & \ddots & \vdots \\ W(a_M, b_1) & W(a_M, b_2) & \cdots & W(a_M, b_N) \end{bmatrix} \tag{2.17}$$

To standardize feature extraction, the resulting CWT matrix is truncated to include the first 13 scales.

Discrete Wavelet Transform (DWT)

Unlike the CWT, which provides a continuous time-scale representation through the inner product of the signal with continuously scaled and translated wavelets, the DWT is specifically adapted for discrete signals [46, 47]. For a given discrete signal, the DWT involves iterative filtering through high-pass and low-pass filters, each followed by downsampling by a factor of two.

At each decomposition level ℓ, the input signal or previous approximation coefficients are filtered to produce approximation coefficients, representing the low-frequency signal features, and detail coefficients, capturing high-frequency localized details. Formally, these coefficients are calculated as:

$$D_{\ell+1}(k) = \sum_{n=0} g(2k-n)A_\ell(n) = D_{\ell+1} \tag{2.18}$$

$$A_{\ell+1}(k) = \sum_{n=0} h(2k-n)A_\ell(n) = A_{\ell+1} \tag{2.19}$$

The raw discrete-wavelet output for one channel is a collection of vectors containing coefficients at a given level:

$$\begin{gathered} \mathrm{DWT} = \Big\{A_L\Big\} \;\cup\; \Big\{\, D_\ell \in \mathbb{R}^{K_\ell} \;\mid\; j \in [L]\Big\}, \\ \text{with } A_L \in \mathbb{R}^{K_L},\ D_\ell \in \mathbb{R}^{K_\ell}, K_\ell = \left\lfloor \frac{N}{2^\ell} \right\rfloor \end{gathered} \tag{2.20}$$

where N is the input signal length in samples and L is the number of levels.

In this study, the Daubechies 4 (db4) wavelet is utilized, employing a 6-level decomposition [48], aligning with practical limits commonly observed due to signal length constraints. The resulting wavelet-based features enable comprehensive characterization of EEG signals, effectively capturing both temporal and spectral properties essential for robust epilepsy pattern detection.

Extracted DWT Features

Let c_ℓ be the wavelet coefficient obtained from level (or scale) $\ell \in \{1, \ldots, L\}$. For every level (or scale) we compress the entire set of coefficients into two energy statistics:

$$\mathrm{MSA}_\ell = \frac{1}{K_\ell} \sum_{k=1}^{K_\ell} |c_\ell|^2, \tag{2.21}$$

$$\mathrm{SSA}_\ell = \sqrt{\frac{1}{K_\ell} \sum_{k=1}^{K_\ell} \Big(|c_{\ell,k}|^2 - \mathrm{MSA}_\ell\Big)^2}. \tag{2.22}$$

Here MSA_ℓ (Mean-Square Amplitude) represents the average energy contained in the sub-band of level (or scale) ℓ, while SSA_ℓ (Square-Amplitude Standard Deviation) quantifies the temporal dispersion of that energy.

Because these four scalar features compress the time–frequency information into simple measures of energy and its spread, they offer a computationally efficient yet informative description of the EEG.

2.5.4 Stockwell (S-Transform) Features

The Stockwell transform (ST) augments the short-time Fourier transform with a frequency-dependent Gaussian window, thereby preserving absolute phase while achieving multi resolution time–frequency localisation [49]. For a discrete signal $x[n]$ the ST is

$$S(\tau, f) = \sum_{n=0}^{N-1} x[n] \, \frac{|f|}{\sqrt{2\pi}} e^{-\frac{(n-\tau)^2 f^2}{2}} e^{-j2\pi f n/N},$$

where τ is the time index and f the Fourier index. Two time-frequency descriptors obtained from the Stockwell Transform are particularly useful for EEG analysis:

- Mean Square Root of the Standard Deviations (mST) – This metric takes the average of the square-rooted standard deviations calculated over every element of the Stockwell-transform matrix. It summarises how much the signal varies across all time–frequency bins [50].

$$\text{mST} = \text{mean}\Big(\sqrt{\text{std}\big(\mathbf{ST}\big)} \Big) \tag{2.23}$$

- Skewness of the Sum of Powers (sST) – This statistic calculates the skewness of the cumulative power of the Stockwell coefficients, exposing any asymmetry in the distribution of spectral power [51].

$$\text{sST} = \text{skewness}\left(\sum_{f=0}^{F_s/2} \big| \, \mathbf{ST} \, \big| \right) \tag{2.24}$$

Before computing the Stockwell Transform, each EEG epoch is edge-trimmed and converted to its analytic form via the Hilbert transform to minimise boundary effects. In practice, mST describes overall variability, while sST highlights imbalances in power; together they help distinguish normal from abnormal brain activity [52].

2.5.5 Phase Locking Value

The phase locking value (PLV) measures the consistency of the phase difference between two signals over time. It quantifies the degree of phase synchronization and is commonly used to assess functional connectivity between brain regions. The PLV is computed as follows [53]:

$$\mathrm{PLV} = \frac{1}{N} \left| \sum_{n=1}^{N} \exp\left(i[\psi_x(n) - \psi_y(n)] \right) \right|, \tag{2.25}$$

where $\psi_x(n)$ and $\psi_y(n)$ are the instantaneous phase values of x_n and y_n at time n, obtained via the Hilbert transform.

The value of PLV ranges from 0 to 1, with 1 indicating perfect phase synchronization and 0 representing no phase relationship. PLV is especially useful for investigating functional connectivity in EEG data, helping to reveal the dynamic interactions between different brain regions. Visualization of connectivity matrices using PLV can illustrate differences between healthy and epileptic patients.

2.5.6 Cross-Correlation

The maximum normalized cross-correlation (C-C) quantifies the similarity between two input signals, x_n and y_n, as a function of the time lag between them. This metric evaluates how well the two signals are aligned at various time lags and provides insight into their relationship. The computation is defined as follows [54]:

$$\hat{R}_{xy,\max}(m) = \frac{1}{\sqrt{\hat{R}_{xx}(0)\hat{R}_{yy}(0)}} \hat{R}_{xy}(m), \tag{2.26}$$

where the cross-correlation function $\hat{R}_{xy}(m)$ is given by

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m}\, y_n^*, & \text{for } m \geq 0, \\ \hat{R}_{yx}^*(-m), & \text{for } m < 0, \end{cases} \tag{2.27}$$

with N representing the signal length and * denoting complex conjugation, $\hat{R}_{xx}(0)$ and $\hat{R}_{yy}(0)$ represent the autocorrelations of x_n and y_n at lag zero, respectively. This normalization ensures the cross-correlation is bounded and comparable across different signals. The maximum normalized cross-correlation is particularly useful for assessing the strength and temporal alignment of relationships between two EEG signals across different time lags [54].

2.5.7 Graph-based Features

Based on the connectivity metrics calculated in Sections 2.5.5 and 2.5.6, two graph representations of the channels are generated: GCC and GPLV [55]. From these graphs, we extract a range of features, including nodal features, edge features, and a single aggregate feature [56]. table 2.5 presents an summary of these network based features.

Table 2.5: Graph features derived from CC/PLV networks: node, edge, and aggregate measures (adapted from BCT) [56, 57].

Feature Name	**Description**
Nodal Features	
Degree	The number of connections a node has.
Strength	The sum of weights of connections a node has.
Assortativity	The tendency of nodes to connect to others that are similar.
Characteristic Path Length	The average shortest path length between nodes.
Local Efficiency	Efficiency of information transfer within a node's neighbourhood.
Eccentricity	Maximum distance between a node and any other node.
Betweenness Centrality	Count of shortest paths passing through a node.
Eigenvector Centrality	Strength of a node based on the importance of its neighbours.
Clustering Coefficient	The degree to which nodes cluster together.
Node Coreness	The level of connectivity of a node within the network core.
Participation Coefficient	Extent of a node's connections across different communities.
Diversity Coefficient	Diversity of a node's connections across communities.
Edge Features	
Assortativity Coefficient	Correlation between the degrees of connected nodes.
Global Efficiency	Efficiency of information transfer across the entire network.
Radius	Minimum eccentricity among all nodes.
Diameter	Maximum eccentricity among all nodes.
Transitivity	Ratio of triangles to triplets in the network.
Edge Neighbourhood Overlap	Overlap in Neighbours of connected node pairs.
Node Pair Degree	Product of degrees of connected nodes.
Aggregate feature	
Matching Index	Amount of overlap in the connection patterns.

2.6 Machine Learning

Machine learning is a branch of artificial intelligence that focuses on developing models capable of learning from data to make predictions or decisions. These methods are generally divided into supervised and unsupervised learning. In supervised learning, the model is trained on labelled data, where each example is paired with its correct output, to predict the labels of previously unseen data. In contrast, unsupervised learning aims to discover patterns or structure in data without access to predefined labels or groupings. In this thesis, a supervised learning approach is adopted. Each training instance is associated with a discrete label indicating its class, more specifically whether the instance corresponds to a healthy or an epileptic condition.

The internal parameters θ, of these models are usually trained by minimizing a cost function $J(\theta)$. Given that the task is a binary classification problem, this cost function is defined as the cross-entropy loss. Formally, each model can be expressed as a hypothesis function $h_\theta(\mathbf{x})$, which maps an input feature vector $\mathbf{x} \in \mathbb{R}^n$ from the feature space to the binary output space $y \in \{0, 1\}$. The training procedure involves finding the parameter set θ that minimizes the cross-entropy cost:

$$\min_\theta J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(\mathbf{x}^{(i)})) \right], \tag{2.28}$$

where m represents the total number of training examples, $y^{(i)}$ is the ground-truth binary label for the i-th example, and $h_\theta(\mathbf{x}^{(i)})$ is the model's predicted probability for that instance.

2.6.1 XGBoost

Extreme Gradient Boosting (XGBoost) [58] is an efficient and scalable implementation of gradient boosting machines. It constructs an ensemble of shallow decision trees sequentially, where each new tree aims to correct the errors of the previous ones. The prediction for the i-th instance is defined as the additive contribution of K trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \tag{2.29}$$

where $\mathcal{F}$ denotes the space of trees with limited depth.

Model training is guided by a regularized objective function that balances predictive accuracy and complexity:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \tag{2.30}$$

where $l(y_i, \hat{y}_i)$ is a differentiable loss (e.g., binary cross-entropy, Eq. (2.28)), and $\Omega(f_k)$ penalizes overly complex trees:

$$\Omega(f_k) = \gamma T_k + \tfrac{1}{2}\lambda \sum_j w_{k,j}^2. \tag{2.31}$$

Here, γ controls the minimum loss reduction required to split a node, while λ penalizes large leaf weights. XGBoost also incorporates advanced techniques such as shrinkage (learning rate), column subsampling (feature subsampling), and sparsity-aware splitting, making it particularly well-suited for modeling high-dimensional

EEG data. Its capacity to model complex, non-linear relationships between features, combined with its interpretability facilitated by the tree structure, makes XGBoost a robust method for producing accurate and clinically interpretable predictions.

2.6.2 Shapley Additive Values

Model interpretability is a key requirement for clinical adoption. To quantify how each feature contributes to the final epilepsy prediction, we employ *Shapley Additive exPlanations* (SHAP) [59]. Given a feature vector $\mathbf{x} = (x_1, \ldots, x_M)$, let $f(\mathbf{x})$ denote the model's predictive score (posterior probability of epilepsy). For every feature $i \in \{1, \ldots, M\}$ the Shapley value ϕ_i is defined as the average

$$\phi_i \;=\; \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!\,\big(|\mathcal{F}| - |S| - 1\big)!}{|\mathcal{F}|!} \Big[f_{S \cup \{i\}}\big(\mathbf{x}_{S \cup \{i\}}\big) \;-\; f_S\big(\mathbf{x}_S\big) \Big], \tag{2.32}$$

where $\mathcal{F}$ is the full set of features and f_S is the model evaluated with only the subset S present, the remaining features being integrated out with respect to their empirical background distribution.

A positive ϕ_i pushes the prediction towards the epileptic class, while a negative one favours the non-epileptic hypothesis. Because SHAP is additive, summing all ϕ_i plus the expected model output (i.e. bias) recovers $f(\mathbf{x})$ exactly:

$$f(\mathbf{x}) \;=\; \mathbb{E}[f(\mathbf{x})] \;+\; \sum_{j=1}^{M} \phi_i. \tag{2.33}$$

2.6.3 Accumulated Local Effects

Accumulated Local Effects (ALE) [60] describe how a feature influences the model on average. ALE plots show how the prediction changes as we gradually increase a feature, while carefully avoiding unrealistic value combinations. This makes ALE more reliable than the commonly used Partial Dependence (PD) plots when features are correlated, as is typical in EEG data. ALE look only at small intervals where data actually exist. For a feature X_j partitioned into bins $\mathcal{I}_k = (z_{k-1}, z_k]$, ALE measures the average local change

$$\Delta_j(z_{k-1}, z_k) \;=\; \mathbb{E}[f(X_{-j}, z_k) - f(X_{-j}, z_{k-1}) \mid X_j \in \mathcal{I}_k]\,, \tag{2.34}$$

and then accumulates these increments across bins. The curve is centred so its mean is zero:

$$A_j(x) \;=\; \sum_{\ell \le k} \Delta_j(z_{\ell-1}, z_\ell) \;-\; \mathbb{E}[\,\tilde{A}_j(X_j)\,]. \tag{2.35}$$

This ensures ALE shows only relative changes, not absolute offsets.

A flat ALE curve means the feature has little effect. A rising curve means higher values push the model more towards predicting positive label, while a falling curve means the opposite. Nonlinear shapes (e.g. thresholds or plateaus) reveal complex effects.

2.6.4 Evaluation Metrics

To quantitatively evaluate the performance of machine learning models, appropriate metrics are essential. In binary classification problems, model predictions can fall into one of four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These outcomes are typically presented in a confusion matrix for easy interpretation.

Confusion Matrix

The confusion matrix illustrates the number of correct and incorrect predictions made by the model for each class. Its structure is presented in Table 2.6.

Table 2.6: Binary classification confusion matrix used to derive ACC, Sens, Spec, F1, BAC, AUC, and AUPRC.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Each component of the matrix serves a specific role in evaluating classification performance:

- **True Positive (TP):** The model correctly predicts a positive case.
- **True Negative (TN):** The model correctly predicts a negative case.
- **False Positive (FP):** The model incorrectly classifies a negative case as positive.
- **False Negative (FN):** The model incorrectly classifies a positive case as negative.

Understanding these outcomes is critical. TP and TN reflect correct model behavior, while FP and FN highlight potential pitfalls. An FP occurs when a healthy individual is incorrectly flagged as having the condition, potentially leading to unnecessary treatment. Conversely, FN results when the model fails to detect the condition in an affected individual, which could result in missed diagnoses and delayed care.

Sensitivity (Recall)

The sensitivity (or true positive rate (TPR), recall in ML) indicates the probability of a positive outcome conditioned on the individual being positive. It is calculated as:

$$\text{Recall} = \text{Sens} = \frac{TP}{TP + FN} \tag{2.36}$$

Specificity

Specificity (alternatively true negative rate, TNR), indicates the probability of an healthy subject to be predicted negative, conditioned on the subject being healthy. It is computed as:

$$\text{Spec} = \frac{TN}{TN + FP} \tag{2.37}$$

Precision

Precision, also Positive Predictive Value (PPV), allows to establish the percentage of true positives among all positive predictions.

$$\text{Prec} = \text{PPV} = \frac{TP}{TP + FP} \tag{2.38}$$

F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both. It is especially useful when the class distribution is imbalanced, as it takes into account both false positives and false negatives.

$$\text{F1} = \frac{2 \times \text{Prec} \times \text{Sens}}{\text{Prec} + \text{Sens}} \tag{2.39}$$

Accuracy

This is the ratio between the correctly predicted outcomes and total number of predictions. It is computed as follows:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.40}$$

Geometric Mean Score

The Geometric Mean Score is defined as the geometric mean of Precision and Recall [61]:

$$\text{GMean} = \sqrt{\text{PPV} \cdot \text{TPR}} \tag{2.41}$$

Balanced Accuracy

Traditional can be misleading when dealing with unbalanced datasets (e.g. high values can be achieved by predicting the majority class, however the model would exhibit low sensitivity or specificity.

$$\text{BAC} = \frac{\text{Sens} + \text{Spec}}{2} \tag{2.42}$$

It can be formulated for arbitrary values, by finding matching thresholds one can fix the sensitivity or specificity at a given value and compute the other, as described in Algorithm 1.

Algorithm 1 Balanced Accuracy with Sensitivity Constraint

```
Require: Labels $\mathbf{y} \in \{0,1\}^n$, prediction scores $\mathbf{s} \in [0,1]^n$, sensitivity threshold $\tau$
Ensure: Balanced Accuracy $BAC$ at specified Sens=$\tau$
$(\mathbf{FPR}, \mathbf{TPR}, \mathbf{T}) \leftarrow$ ROC_CURVE$(\mathbf{y}, \mathbf{s})$    ▷ Sweep all operating points
$k \leftarrow \min\{\, i \mid \mathbf{TPR}[i] \geq \tau \,\}$    ▷ First index that meets the sensitivity target
$\theta \leftarrow \mathbf{T}[k]$    ▷ Chosen score threshold
$\hat{\mathbf{y}} \leftarrow \mathbb{I}\{\mathbf{s} \geq \theta\}$    ▷ Hard predictions
$(TN, FP, FN, TP) \leftarrow$ CONFUSIONMATRIX$(\mathbf{y}, \hat{\mathbf{y}})$
Sensitivity $\leftarrow \dfrac{TP}{TP + FN}$
Specificity $\leftarrow \dfrac{TN}{TN + FP}$
$BAC \leftarrow \dfrac{\text{Sensitivity} + \text{Specificity}}{2}$
return $BAC$
```

Receiver Operating Characteristic (ROC)

The ROC curve plots the true-positive rate TPR $= \frac{TP}{TP+FN}$ against the false-positive rate FPR $= \frac{FP}{FP+TN}$ as the discrimination threshold sweeps the unit interval. Its area under the curve (AUC) summarises performance in a single, threshold-independent scalar: an AUC of 0.5 equals random guessing, whereas 1 denotes perfect separation. Because TPR and FPR are both insensitive to class prevalence, ROC analysis remains informative even when the epileptic and control cohorts are imbalanced, although it may obscure poor precision when controls dominate.

Clinically Relevant ROC Evaluation

Referring once more to the ILAE definition of epilepsy [62], the second diagnostic criterion requires a recurrence risk of 60% or greater to establish a diagnosis of epilepsy. Building on this, Van der Kleij [11] applied Bayesian inference to express the posterior probability of epilepsy as a function of model performance in ROC space:

$$P(A \mid B) = \frac{\text{sens} \cdot P(A)}{\text{sens} \cdot P(A) + (1 - \text{spec}) \cdot (1 - P(A))} \tag{2.43}$$

$P(A \mid B)$: Posterior probability that a subject has epilepsy (A) given a positive model prediction (B)

$P(A)$: Prior probability that a subject has epilepsy

sens : Sensitivity (TPR or recall) of the model

spec : Specificity (TNR) of the model

$P(A)$ is estimated from the cohort of interest, representing the distribution skewness in the available data. $P(A \mid B)$ is fixed at 0.6 by ILAE definition, we can rearrange the terms so that we produce a line in ROC space:

$$\text{TPR} = \frac{P(\text{posterior}) \cdot P(\text{healthy})}{P(\text{epileptic}) \cdot (1 - P(\text{posterior}))} \cdot FPR \tag{2.44}$$

Precision–Recall (PR) Curve

The PR curve depicts precision (PPV) versus recall (TPR). Unlike the ROC, precision directly incorporates the proportion of false alarms and is therefore more sensitive to class imbalance. We report the area under the PR curve (AUPRC) offering a complementary view that prioritises high-confidence predictions in low-prevalence settings such as routine EEG screening.

2.7 Mann-Whitney U-Test

After extracting features from both epileptic and healthy patients, the Mann–Whitney U-test is employed to compare the distributions of these features across the two groups. This non-parametric test is particularly useful for identifying differences between two independent samples without the assumption of an underlying normal distribution [63].

Its underlying assumptions instead are: observations must be independent both within and between groups; each feature should be measured on at least an ordinal scale, as the test operates on ranks rather than raw values; for the test to be interpreted strictly as a comparison of locations (medians), the two distributions should have similar shapes (i.e. variances and skewness) if this assumption is violated, the test may detect general distributional differences rather than pure shifts in central tendency; and finally, while the exact distribution of the test statistic assumes there are no tied values, in practice the presence of ties is addressed by the algorithms used to compute the values.

The test statistic U is computed using Eq. (3.4):

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1, \tag{3.4}$$

where n_1 and n_2 represent the sample sizes of the two groups, and R_1 is the sum of the ranks for the first group. For every feature the test evaluates

$$H_0: \; F_1(x) = F_2(x) \quad \forall x \qquad \text{(the two distributions are identical)}$$

$$H_1: \; P(X_1 < X_2) \; \neq \; \tfrac{1}{2} \qquad \text{(the two distributions differ in location)}$$

where F_1 and F_2 denote the cumulative distribution functions of the epileptic and healthy groups, respectively, and X_1, X_2 are random draws from those groups. A two-tailed alternative is used because either group could exhibit larger (or smaller) feature values.

Under the null hypothesis of identical distributions, the exact sampling distribution of U can be enumerated for small samples; the p-value is the proportion of all possible rank permutations whose U is at least as extreme as the observed one.

For moderate or large samples exact enumeration is impractical, so U is transformed to a normal variate:

$$z \; = \; \frac{U - \mu_U}{\sigma_U}, \qquad \mu_U = \frac{n_1 n_2}{2}, \qquad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}},$$

and the two-tailed p-value is obtained from the standard normal cumulative distribution function.

Because this test is repeated across m extracted features, the family-wise Type I error rate would inflate if we relied on the raw p-values. We therefore apply the Bonferroni correction, which in its equivalent "p-value adjustment" form multiplies each raw p-value by the number of tests:

$$p_{\text{adj}} \; = \; m \times p_{\text{raw}}, \quad \text{capped at 1.} \tag{2.45}$$

A feature is deemed significant if $p_{\text{adj}} < \alpha$, where α is the conventional threshold (in this study $\alpha = 0.05$). This conservative procedure ensures that the probability of making even a single false discovery across all m features does not exceed α.

Since p-values alone do not convey the magnitude of an effect, we calculate the effect size for each significant feature to better quantify its relationship with the categorical variable [64]. In this study, we use the rank-biserial correlation coefficient, defined as:

$$r = 1 - \frac{2U}{n_1 n_2} \tag{2.46}$$

This coefficient estimates both the strength and direction of the association between two independent groups and the outcome variable. Its values range from -1 to 1, where -1 indicates a perfect negative relationship, 0 indicates no relationship, and 1 represents a perfect positive relationship.

Datasets 3

The analyses in this thesis draw on two complementary electroencephalography (EEG) resources: the publicly released Temple University Hospital (TUH) Epilepsy Corpus and a clinically curated cohort from Erasmus MC (EMC). Together they provide a broad spectrum of interictal photic-stimulation (IPS) recordings that span different acquisition settings, demographics, and prevalence of epilepsy (Table 3.1). Leveraging both datasets allows us to (i) benchmark our methods on a widely used public corpus and (ii) assess their generalisability on an independent, real-world clinical cohort. The remainder of this chapter details the composition, selection criteria, and IPS protocols of each dataset, highlighting key similarities and differences that inform the experimental design in later chapters.

Table 3.1: Dataset composition: counts of IPS-segment samples and subjects (epileptic vs. non-epileptic) for TUH and EMC.

Dataset	TUH	EMC
Epileptic	145 (13)	1302 (40)
Non-Epileptic	316 (18)	3243 (101)
Total	461 (31)	4545 (141)

3.1 TUH Epilepsy Corpus

The Temple University Hospital EEG Data Corpus (TUEG) [65] is the largest publicly available collection of clinical EEG data. It contains 16,986 routine EEG recordings from 10,874 unique subjects, with data collection beginning in 2000. A subset of this dataset, the Temple University Epilepsy Corpus (TUEP), includes diagnostic labels indicating whether a subject has epilepsy. Within TUEP, there are 316 recordings from 100 healthy individuals and 1,389 recordings from 100 individuals diagnosed with epilepsy. From this subset, recordings that include photic stimulation annotations were identified. This yielded 40 recordings from 31 distinct subjects: 13 with epilepsy and 18 healthy. The ages of these subjects range from 12 to 88 years, with 12 females and 19 males. These particular recordings were collected between 2003 and 2011.

The photic stimulation protocol was determined by examining the ‘Photic PH’ channel, characterized by bursts of peaks corresponding to the photostimulation frequency, which was identified by computing the spectrogram of the ‘Photic PH’

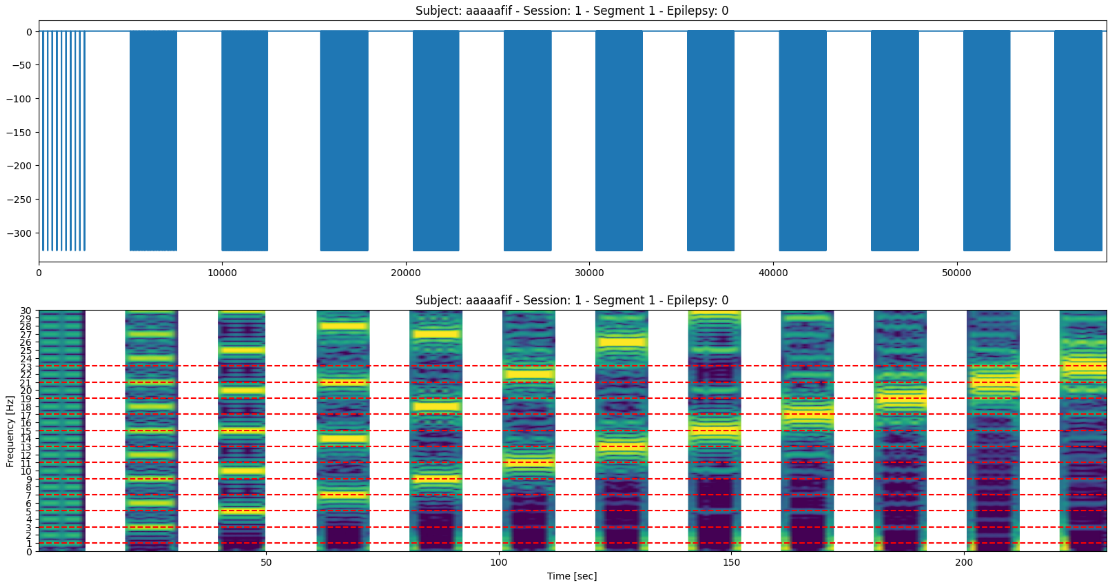

Figure 3.1: Example TUH IPS trial. Top: 'Photic PH' trigger channel; bottom: spectrogram showing stimulus frequency sweep and harmonics during intermittent photic stimulation.

channel. The protocol employed by TUH differs from that of EMC, described in section 2.2, and consists of 10-second periods of photostimulation followed by 10-second rest intervals. Frequencies were applied incrementally in 2 Hz steps, starting at 1 Hz and progressing up to 21 Hz, with occasional extension to 23 Hz for certain patients.

Among the 31 subjects, three have three recordings each, two have two recordings each, and the remaining subjects have one recording each. Notably, 26 of these 40 recordings correspond to the subject's first recording in the full TUEG dataset.

3.2 Erasmus MC

A total of 141 adult subjects were retrospectively included in this study, of whom 101 were classified as 'healthy' and 40 as 'epileptic', based on the occurrence of a recurrent seizure. All participants were seen in the emergency room (ER) following a first seizure. In accordance with standard protocol, an initial EEG was conducted by the Department of Clinical Neurophysiology to assess the risk of seizure recurrence. When the first EEG proved inconclusive, a second EEG after sleep deprivation had been performed to increase the likelihood of detecting epileptiform activity.

In this study, both the initial and follow-up (sleep-deprived) EEGs were inconclusive for all included subjects. Subject classification was based on a minimum of one year of clinical follow-up. Patients who experienced a recurrent seizure during this period were labelled ‘Epileptic’, whereas those who remained seizure-free were classified as ‘Healthy’.

If a subject initially labelled as healthy was later found to have had a recurrent seizure upon re-evaluation of their medical record for clarification purposes, their label was updated to ‘Epileptic’.

Only the first EEG, not the sleep-deprived follow-up EEG, was used in this study. These recordings were retrieved from the EEG archive of the Department of Clinical Neurophysiology at Erasmus Medical Center, Rotterdam.

In comparison with the datasets reported in the MSc theses by Mirwani and van der Kleij [10, 11], some subjects were excluded from this study due to annotation inconsistencies related to the IPS trials, after review and approval by Dr. Van den Berg. This study was reviewed and approved by the Medical Ethics Review Committee (Medisch Ethische Toetsings Commissie, METC) as non-WMO research, under case number MEC-2021-0145. All data were pseudonymized and accessed securely through the ‘my Digital Research Environment’ (anDREa B.V. 2021) [66].

3.3 Feature Sets

For each of the feature types described in Section 2.5, a feature vector is constructed for each data point (either a whole IPS trial EEG recording or a 10s epoch derived from it). Since the dimensionality of these feature vectors may vary depending on the montage of choice, a summary is provided in Table 3.2 to help to understand these sizes and the content of the each feature set.

Table 3.2: Feature-set sizes by montage: dimensionality per set as the number of channels N_{ch} varies.

Feature Type	**Feature details**	**Feature Vector Shape**
Spectral	Relative power in $\delta, \theta, \alpha, \beta$ and γ frequency bands	$1 \times (N_{ch} \times 5)$
UTM	Mean, Median, Std, Kurtosis, Skewness, peak-to-peak amplitude, zero crossings, number of peaks, NLEO-ED, NLEO-TK, Signal Energy, Shannon Entropy	$1 \times (N_{ch} \times 13)$
DWT	MSA and SSA of DWT coefficients at 6 levels	$1 \times (N_{ch} \times 6 \times 2 \times 2)$
CWT	MSA and SSA of CWT coefficients at 13 scales	$1 \times (N_{ch} \times 13 \times 2)$
CC	Cross-Correlation between channels	$1 \times \left(N_{ch} \times \frac{N_{ch}-1}{2}\right)$
sST	Skewness of sum of powers in $\delta, \theta, \alpha, \beta$ and γ bands, plus one for the whole spectrum	$1 \times (N_{ch} \times 6)$
mST	Mean square root of the standard deviation of powers in $\delta, \theta, \alpha, \beta$ and γ bands, plus one for the whole spectrum	$1 \times (N_{ch} \times 6)$
PLV	Phase Lock Value for each channel's $\delta, \theta, \alpha, \beta$ and γ bands, plus one for the whole signal.	$1 \times \left(N_{ch} \times \frac{N_{ch}-1}{2} \times 6\right)$
GCC & GPLV	Nodal and Edge features from CC and PLV networks	1×20 each

Methods 4

This chapter details the end-to-end methodology adopted in this study. Beginning with the acquisition of raw EEG recordings, we walk through each stage of the pipeline: from signal preprocessing and feature extraction to model training, cross-validation, and performance evaluation, providing all information necessary to reproduce our results.

4.1 Pre-Processing

Pre-processing of raw EEG data is a critical step to enhance signal quality and ensure reliable feature extraction and classification. The pipeline used in this study comprises the following stages:

1. Channels not included in the 10-20 system are discarded. Signals are converted to microvolts to maintain unit consistency across all recordings.

2. A notch filter (4$^{\text{th}}$ order Butterworth) is applied at 50 Hz for EMC data and 60 Hz for TUH data to eliminate power line interference.

3. A high-pass filter at 1 Hz (4$^{\text{th}}$ order Butterworth) to remove DC offset and low-frequency drifts. All filters are applied using zero-phase filtering to avoid phase distortion.

4. Segments containing extreme values, indicating of measurement artifacts such as electrode pop, are removed.

5. Data is downsampled to 200 Hz for EMC and 250 Hz for TUH to reduce data size and standardize temporal resolution for further processing.

6. The EEG signals are then segmented into 1-second windows using a sliding buffer. Artifact rejection is implemented as in [67], where each segment's root mean square (RMS) is computed. Segments exceeding a noise-based threshold are flagged and discarded.

7. The photostimulation segments are extracted, using annotation files for EMC data, and for TUH data by inspecting the 'Photic PH' channel, as previously illustrated in Section 3.1.

A summary of the preprocessing pipeline is shown in Figure 4.1.

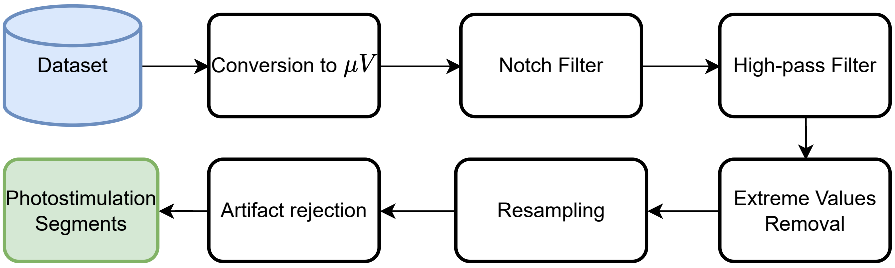

Figure 4.1: End-to-end preprocessing: channel selection, unit conversion, notch (50/60 Hz), 1 Hz high-pass, artifact removal, resampling (EMC 200 Hz; TUH 250 Hz), segmentation (1 s) with RMS-based rejection, and IPS segment extraction.

4.2 Cross-Validation: Leave one subject out

Cross-validation (CV) is a fundamental technique in machine learning as it robustly assesses a model's performance on unseen data, offering a realistic indication of how the model might generalize in practical applications [30]. Among various cross-validation methods, leave-one-subject-out cross-validation (LOSO CV) closely resembles real-world deployment scenarios. Specifically, LOSO CV involves iteratively removing one subject out of N subjects from the dataset, training the model(s) on the remaining N-1 subjects, and predicting the outcome for the excluded subject. The predicted outcome is stored, and this process repeats sequentially until each subject has served once as a test case. Finally, the aggregated stored predictions across all subjects provide performance metrics reflective of true generalization capabilities.

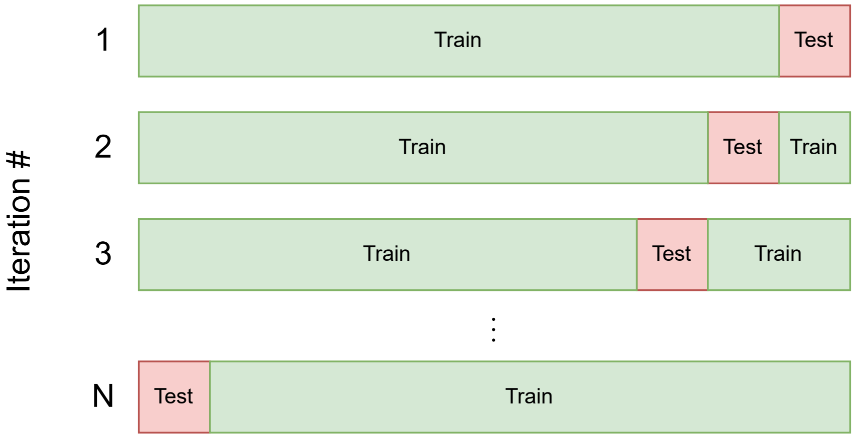

Figure 4.2: Leave-One-Subject-Out (LOSO) cross-validation: train on N-1 subjects and test on the held-out subject; repeat until every subject is tested once.

4.3 XGBoost Parameters

In our experiments we used the `XGBClassifier` implementation from the `xgboost` library. The main hyper-parameters were set as shown in Table 4.1

Table 4.1: XGBoost hyperparameters: roles and selected values for gradient-boosted trees used across experiments.

Parameter	**Description**
`n_estimators=100`	Number of boosting trees (iterations).
`max_depth=6`	Maximum depth of each decision tree; controls model complexity.
`subsample=0.9`	Fraction of samples used for training each tree.
`scale_pos_weight`	Balances positive/negative classes for imbalanced data. Dynamically computed for each fold
`gamma=0.1`	Minimum loss reduction required to split a node.
`learning_rate=0.1`	Shrinks contribution of each tree; lower values need more trees.

4.4 Feature Set Selection

As an initial step, LOSO-CV is applied to each individual feature set in order to identify for each feature type the best-performing combination of montage, segment length, and statistical combiner. As shown in Table 2.2, we evaluate four montages and six segment lengths, using five different statistical combiners. This results in $K = 4 \times (6 \times 5) = 120$ unique combinations, and therefore 120 LOSO CV results per feature set. For features extracted from 10-second epochs, only four segment lengths are considered, leading to $K = 4 \times (4 \times 5) = 80$ combinations for each set. For every combination, training is repeated five times and the results are averaged. Each of the ten feature types is then ranked based on the highest average LOSO CV AUC achieved among the 120 (or 80) tested combinations. We retain the information about the optimal montage, segment length, and statistical combiner for each of the ten feature types, so that they can be used in the next stage, when combining multiple features for the ensemble method.

4.5 Ensembling

In the ensembling phase, LOSO CV is repeated by combining different feature sets. In this multi-feature scenario LOSO is performed as follows: A number of feature sets from 2 to 10 is chosen. For each selected feature combination, the data is split as train, validation, and test set. If there are N subjects in the dataset, the N-th subject is left for testing. From the remaining N-1 subjects, 70% of the subjects are used for training and 30% for finding the best combination of weights

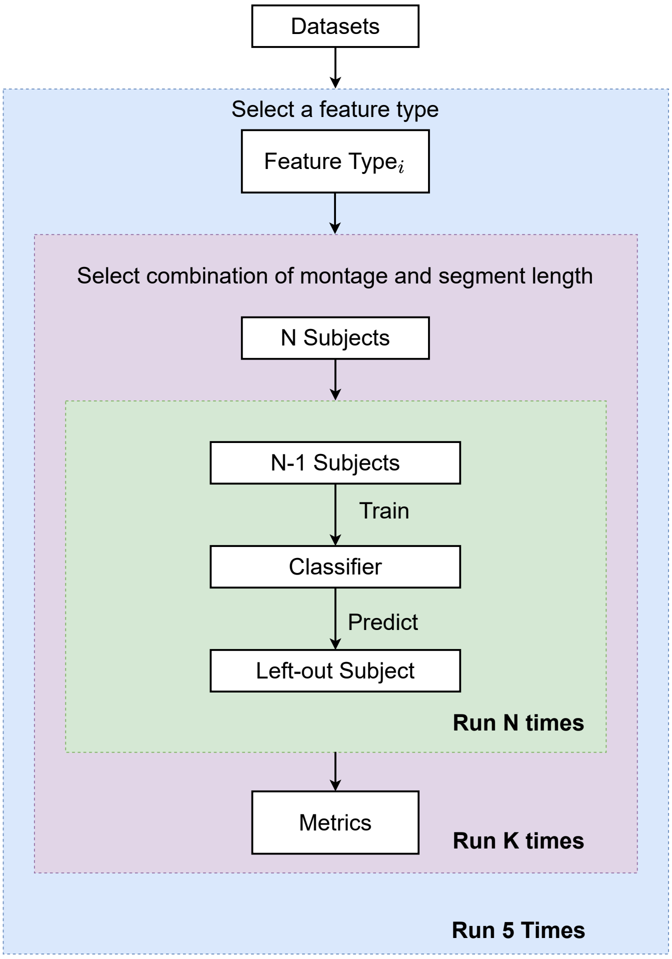

Figure 4.3: LOSO evaluation flow for individual feature sets: select montage/segment length/combiner per set, train XGBoost, retain the best model.

for each feature-specific classifier. Using the validation data, we also determine the optimal threshold that maximizes the Geometric Mean score ((2.41)). Finally, the models are re-trained, on the union of training and validation data, and the optimal threshold and weights are used to predict the test subject. This step is repeated N times to evaluate all the subjects. Then, metrics are computed.

Given the large number of possible combinations, considering different montages, features, segment lengths, and statistical measures, a guided approach is taken. Based on the feature rankings established in the initial selection stage, we systematically evaluate ensembles containing two, three, four, and more feature sets. For each feature set, we utilize the montage, EEG segment length, and statistical measure that best performed on that feature during the previous stage of feature set selection.

4.6 Stacking Classifier

To determine the best combination of weights coming from each individual classifier, a logistic regression (LR) model is used. The LR model is trained on the log-odds predictions of the validation data, and the optimized weights will then be applied to the left-out subject. This configuration is known as a stacking meta-classifier. For interpretability we impose two constraints on the weight vector: every weight must be non-negative and all weights must sum to one. The learning task therefore becomes the following constrained optimisation problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\sigma(z_i) + (1-y_i)\log\sigma(1-z_i)\right] - \alpha\sum_{k} \log\big(w_k\big) \\ \text{s.t.} \quad & \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0 \ \ \forall\, k, \end{aligned} \tag{4.1}$$

where

$$z_i = \sum_{k=1}^{K} w_k\, p_{ik},$$

is the weighted log-odd score, N is the number of validation samples, K the number of base classifiers and p_{ik} the log-odds from classifier k on sample i, $\sigma(\cdot)$ is the logistic function, and $\alpha \sum_k \log\big(w_k\big)$ is a penalty term to promote uniformity in the weights. All results will be reported at $\alpha = 0.05$, unless otherwise specified.

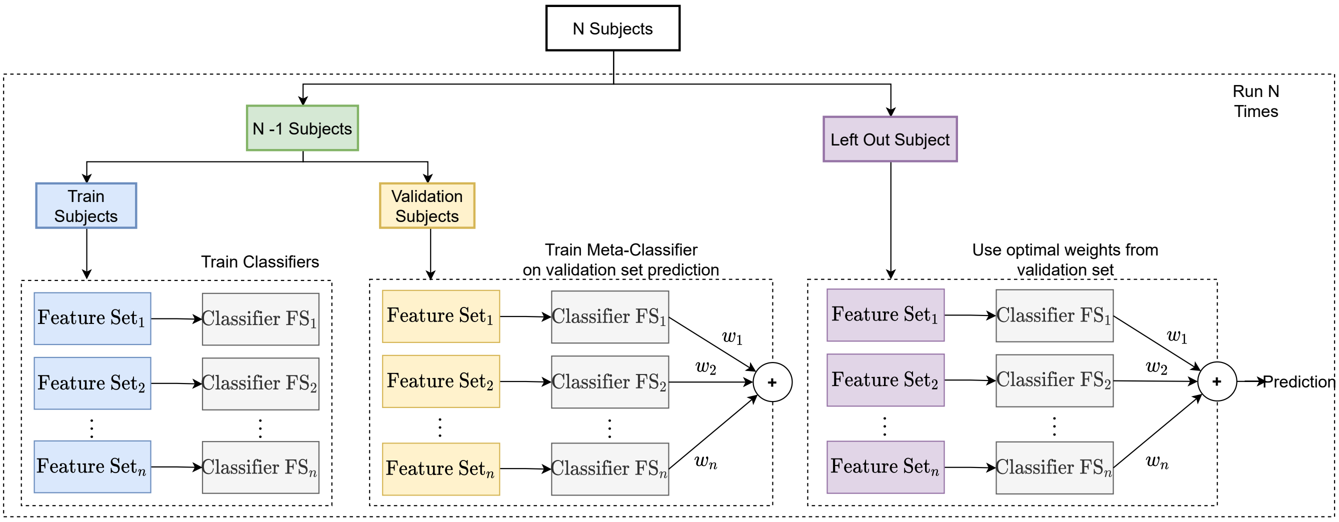

Figure 4.4: Ensemble LOSO scheme: base learners trained on selected feature sets; meta-classifier stacks base predictions to yield per-subject scores.

4.7 Subject Prediction

Since the ultimate goal of this thesis is to investigate machine learning's potential for epilepsy diagnosis, specifically aiming for predictions at the patient level rather than merely individual data segments, the methods utilizing data split into 10-second epochs or in the case a patient has more than one recording, we obtain predictions for each epoch/recording. These predictions are then averaged to determine whether the patient is ultimately epileptic or not. Additionally, the same metrics presented in Section 2.6.4 will be computed for the subject-level results, ensuring comparability between outcomes obtained from the 10-second epochs and those derived from entire subject recordings.

4.8 Feature Attribution with LOSO

Because leave-one-subject-out (LOSO) validation produces as many models as subjects, feature attribution methods must be adapted to this pipeline. This allows us to examine not only which features are important within a fold, but also how consistently decision rules generalise across different subjects.

To avoid data leakage, SHAP values for each left-out subject are computed using the explainer fitted to the model trained on the $N-1$ subjects of that fold. The resulting attributions are then aggregated across folds in the same way as the predictions. For visualisation, we employ beeswarm plots, which compactly display both the distribution and direction of SHAP values. To maintain interpretability, we restrict the analysis to the five features with the largest mean absolute SHAP values.

Accumulated Local Effects (ALE) are computed analogously within each LOSO fold. Feature bins are defined using only the training subjects, fold-specific ALE curves are centred, and then interpolated onto a common grid before averaging across folds. This produces a mean ALE curve with variability bands, offering both interpretability and robustness. Given the clinical origin of the EMC dataset, ALE analysis is reported specifically for this cohort, as the TUH data reflects a more heterogeneous patient population with less certain diagnostic context.

Statistical Analysis 5

This chapter presents the statistical analysis on the extracted feature sets for both TUH and EMC datasets. The Mann-Whitney U-test is performed on each individual feature, to check for significant differences between the two groups. For each dataset and feature set, the montage, segment length, statistical combiner and feature index with the lowest normalized p-value are reported. Only the features extracted over the whole IPS trials are analysed, to not violate sample independence assumptions.

5.1 TUH Dataset

Table 5.1 lists, for every feature set, the montage, segment length, combiner, and feature index associated with the lowest p-value. In total, 33 features in the TUH dataset proved significant after adjustment; none of the gPLV, PLV, or MST features met this threshold.

Full details including effect sizes appear in Table A.1, from which two patterns stand out:

- CC features: Feature 37 shows a consistently significant p-value across nearly all segment lengths when the bipolar montage is used.
- CWT features: Feature 194 becomes significant for longer segment lengths when either of the asymmetry combiners (skew or kurtosis) is applied.

Box-plots in Figures A.1 and A.2 illustrate the distribution of these significant features. It can be noted how, despite having significant p-values and comparatively large effect sizes, the feature distributions between the two groups tend to overlap, pointing to the absence of a simple linear separation.

5.2 EMC Dataset

Similarly, table 5.2, summarizes for each feature set, the montage, segment length, combiner, and feature index associated with the lowest p-value. In total, 16 features from the EMC dataset were below the significance threshold after probability adjustment; none of the GPLV features met this criterion.

Full details including effect sizes are reported in Table A.2,

The box-plots in Figure A.3 illustrates the distribution of these significant features. Once again, it can be noted how, despite having significant p-values and

Table 5.1: TUH feature significance: Mann–Whitney U results for features with between-group differences.

Feature Type	Montage	Segment Length [s]	Combiner	Index	P-value
cc	BipolarDB	2	mean	37	0.02663
cwt	CAR	5	skew	194	0.03506
dwt	Cz	20	skew	76	0.02409
gcc	CAR	1	std	19	0.03633
gplv	N/A	N/A	N/A	N/A	N/A
plv	N/A	N/A	N/A	N/A	N/A
mst	N/A	N/A	N/A	N/A	N/A
sst	Laplacian	20	skew	106	0.02794
spectral	CAR	60	std	37	0.02445
utm	Cz	1	mean	205	0.01476

Table 5.2: EMC feature significance: Mann–Whitney U results.

Feature Type	Montage	Segment Length [s]	Combiner	Index	P-value
cc	CAR	20	skew	59	0.01959
cwt	BipolarDB	20	skew	326	0.01750
dwt	Laplacian	2	kurt	6	0.002747
gcc	CAR	20	kurt	6	0.01697
gplv	N/A	N/A	N/A	N/A	N/A
plv	BipolarDB	1	skew	335	0.02298
mst	BipolarDB	1	skew	54	0.04513
sst	CAR	5	std	1	0.04602
spectral	Laplacian	1	std	18	0.03337
utm	Cz	20	kurt	63	0.02531

comparatively large effect sizes, the feature distributions between the two groups tend to overlap, pointing to the absence of a simple linear separation. If compared to the distribution of the TUH data this overlap is even greater, as in general the effect sizes are smaller for p-values of comparable magnitude.

Interestingly, no significant features arises from measures of central tendency such as the mean or median; instead, every significant feature emerges from dispersion combiners, namely standard deviation, skewness, and kurtosis.

Results - TUH Dataset 6

This chapter presents the experimental results on the TUH dataset in two versions: one including all patients and one excluding patients with IEDs. First, the individual feature sets are compared, followed by ensembles of varying lengths built from the best-performing sets identified in the previous stage. Additionally, for each of the top individual feature sets, an explanation of the predictive value of their features is provided.

6.1 Individual Feature Sets

6.1.1 With and Without IEDs

The optimal montage, segment length, and statistical combiner for each feature type are determined using the method described in Section 4.4. A summary of the results is provided in Table 6.1.

Table 6.1: TUH best individual feature sets (with & without IEDs): montage, segment length, combiner, and performance (% mean ± SD). Best performance in bold.

Feature	**Montage**	**Segment Length**	**Combiner**	**AUC**	**BAC$_{80}$**
Spectral	CAR	1	skew	81.10 ± 0.87	73.88 ± 3.29
CWT	BipolarDB	60	std	81.76 ± 9.43	81.81 ± 2.10
DWT	Cz	10	skew	76.35 ± 0.99	82.43 ± 2.18
MST	BipolarDB	10	skew	74.73 ± 2.10	73.79 ± 4.89
SST	Laplacian	20	skew	80.24 ± 1.28	77.49 ± 6.41
CC	Cz	10	skew	79.10 ± 0.82	83.19 ± 3.84
PLV	Laplacian	2	std	79.10 ± 1.43	81.29 ± 1.92
GCC	CAR	1	std	70.56 ± 0.87	69.47 ± 7.01
GPLV	BipolarDB	1	mean	52.70 ± 34.47	64.62 ± 19.99
UTM	Laplacian	60	mean	**87.46 ± 0.75**	**81.91 ± 2.22**

UTM achieves the highest AUC, surpassing all other feature sets by a clear margin: the next best, Spectral and CWT, score about 6% lower on average. Notably, CWT exhibits a higher AUC standard deviation, suggesting less consistent performance. Interestingly, both very short (1s) and long (60s) segment lengths are associated with top AUC results. The skewness combiner performs well across multiple feature sets, reinforcing the idea that robust statistics can outperform variance-based measures. Performance variability is evident, with CWT and GPLV showing particularly high standard deviations, whereas UTM remains highly stable. In terms of balanced accuracy (BAC), however, UTM ranks third with an average BAC of 81.91%, behind DWT at 82.43% and CWT at 83.19%.

The ROC curves for all feature sets are shown in Fig. 6.1.

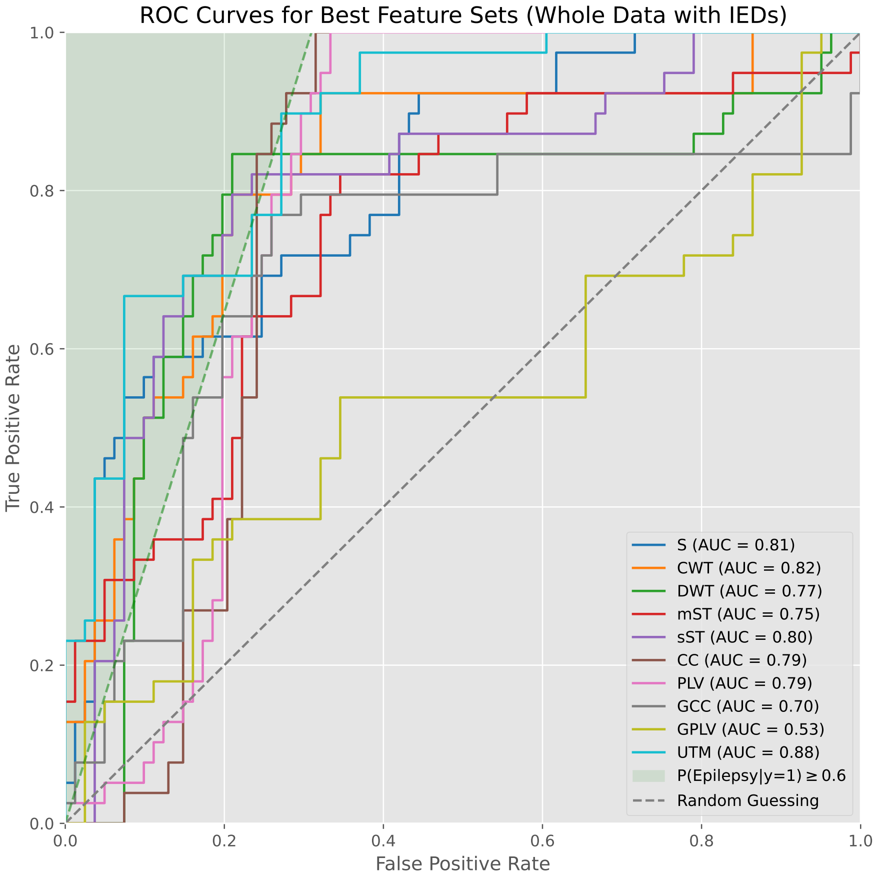

Figure 6.1: TUH (with & without IEDs): ROC curves for each feature set computed over whole IPS trials.

In addition to the cumulative statistics, average Shap values across the runs are computed for each of the optimal setting for each feature set. Fig. 6.2 shows the average SHAP value for the top 5 features in the UTM set.

6.1.2 Without IEDS

As before, the optimal set of parameters: montage, segment length, and statistical combiner, is identified for each feature set, this time considering only patients who do not exhibit IEDs in the recordings. The results for each feature set are presented in Table 6.2.

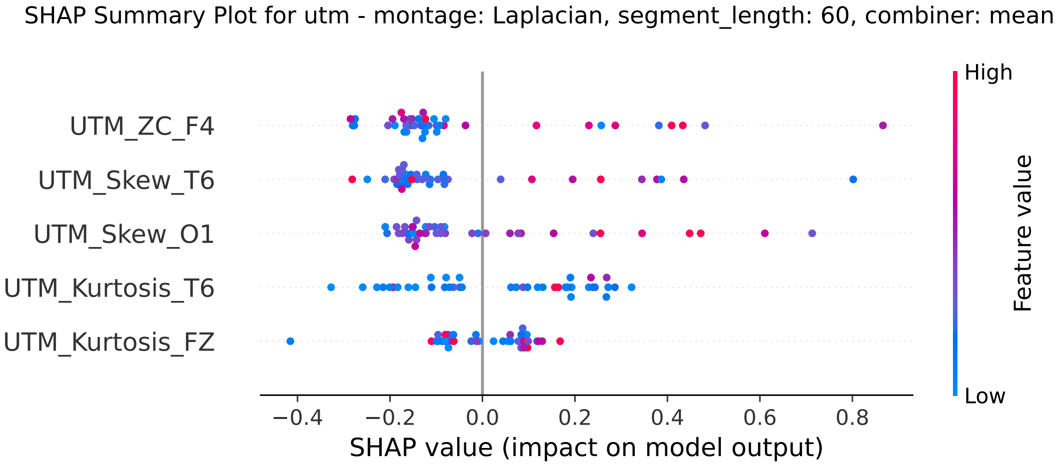

Figure 6.2: UTM set global feature attribution: top 5 mean absolute SHAP values for temporal features on TUH data.

In terms of AUC, the best-performing feature set is Spectral, achieving a consistent score of 84.18% with zero deviation across runs, followed closely by sST at 84.04% and PLV at 83.34%. When considering BAC, the Spectral feature set again ranks first, with a score of 84.68%, followed by sST at 82.27% and CC at 80.08%.
The ROC curves for all feature sets are shown in Fig. 6.3.

Similarly to before, we compute the average SHAP values across the runs for each of the optimal feature sets. Fig. 6.4 shows the average SHAP values for the top 5 features in the Spectral set.

Table 6.2: TUH best individual feature sets (IED-free): montage, segment length, combiner, and performance (% mean ± SD). Best performance in bold.

Feature	**Montage**	**Segment Length**	**Combiner**	**AUC**	**BAC$_{80}$**
Spectral	CAR	20	std	**84.18 ± 0.00**	**84.68 ± 0.01**
CWT	BipolarDB	60	std	76.32 ± 0.03	69.87 ± 0.02
DWT	Laplacian	20	std	75.31 ± 0.04	73.91 ± 0.01
MST	BipolarDB	10	skew	69.02 ± 0.03	69.92 ± 0.05
SST	CAR	20	mean	84.06 ± 0.03	82.27 ± 0.06
CC	Cz	60	median	82.72 ± 0.00	80.08 ± 0.02
PLV	BipolarDB	5	median	83.34 ± 0.02	77.95 ± 0.05
GCC	Laplacian	5	mean	72.16 ± 0.02	74.87 ± 0.01
GPLV	BipolarDB	10	median	63.30 ± 0.03	71.46 ± 0.01
UTM	CAR	1	median	82.38 ± 0.01	77.95 ± 0.02

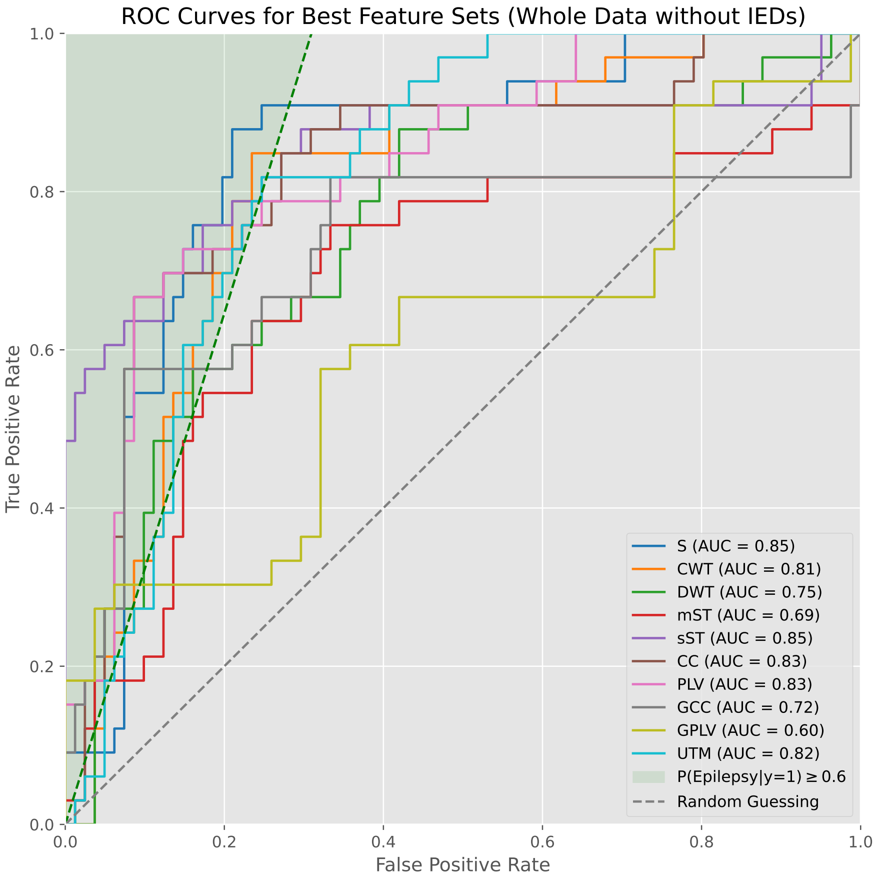

Figure 6.3: TUH (IED-free only): ROC curves per feature set on IPS trials with subjects containing IEDs removed.

6.2 Ensembles - With and Without IEDs

After assessing the performance of each individual feature set, we retain the configurations of the best-performing ones. Next, we construct ensembles of varying sizes using this selected set. Predictions are weighted using the stacking classifier described in Section 4.5.

The size 6 ensemble cc+plv+mst+utm+gcc+gplv achieves the highest AUC of 93.80%, edging out other combinations by about 0.2–0.4%. The simpler ensemble cc+plv+sst+utm attains the top BAC (90.20%), highlighting that smaller, more focused ensembles can outperform larger ones in balanced accuracy. Standard deviations vary, with some ensembles (e.g., cc+utm+gcc) showing higher variability, whereas the best-performing ensembles (e.g., cc+plv+sst+utm) remain relatively stable. Notably, all ensembles except the two-feature one, include the UTM set, highlighting its robustness as a core feature set for this dataset. The ROC curves of each ensemble model presented above are shown in Fig. 6.5.

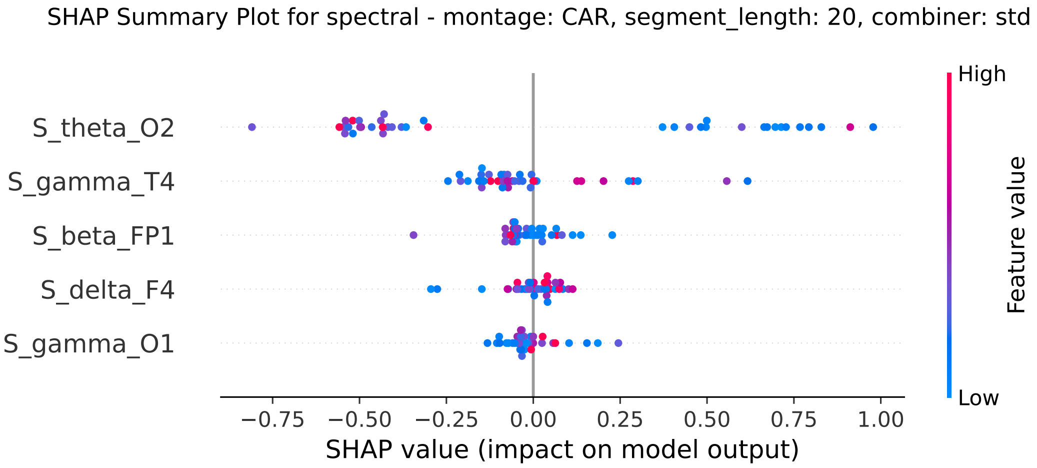

Figure 6.4: Spectral set global feature attribution: top 5 mean absolute SHAP values for band-power features on TUH.

Table 6.3: TUH ensembles (all data): performance metrics (mean ± SD over LOSO subjects) for stacking and baselines. Best performance in bold.

Combination	**AUC**	**BAC$_{80}$**
mst+gcc	85.00 ± 3.60	76.80 ± 7.50
cc+utm+gcc	90.20 ± 4.20	83.50 ± 6.90
cc+plv+sst+utm	93.60 ± 3.60	**90.20 ± 5.10**
cwt+plv+sst+utm+gplv	92.30 ± 3.30	86.90 ± 4.10
cc+plv+mst+utm+gcc+gplv	**93.80 ± 2.20**	80.10 ± 6.80
cc+cwt+dwt+mst+sst+utm+gcc	92.40 ± 3.10	69.30 ± 4.50
cc+plv+mst+sst+spectral+utm+gcc+gplv	93.40 ± 2.00	70.10 ± 3.60
cc+dwt+plv+mst+sst+spectral+utm+gcc+gplv	90.70 ± 1.20	71.60 ± 10.10
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	88.90 ± 6.60	69.60 ± 8.80

6.3 Ensembles - IED-free data

Similar to the single-set experiments described in Section 6.1.2, we retrain the ensemble models on the TUH dataset, this time excluding patients with IEDs. A summary of the results is presented in Table 6.4.

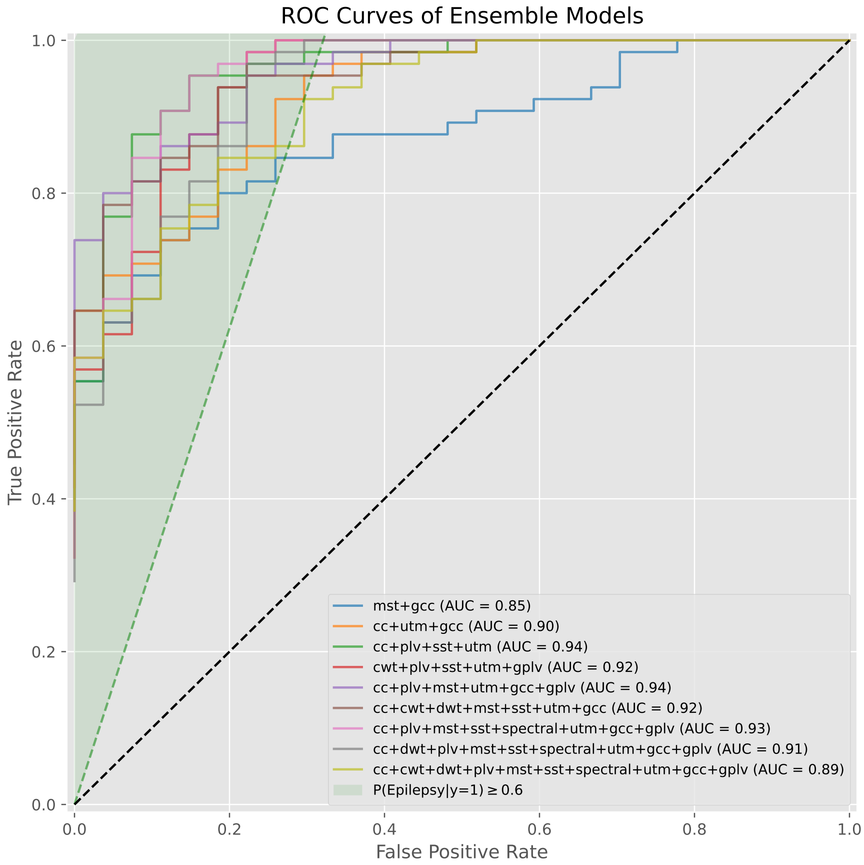

Figure 6.5: TUH ensembles (all data): ROC curves comparing stacking/other ensembles built from best individual sets.

Table 6.4: TUH ensembles (IED-free): performance metrics (mean ± SD) after removing subjects with IEDs. Best performance in bold.

Combination	**AUC**	**BAC$_{80}$**
plv+sst	88.50 ± 1.50	83.10 ± 5.10
mst+sst+spectral	89.70 ± 4.80	84.00 ± 5.40
plv+sst+spectral+utm	90.10 ± 5.60	84.00 ± 6.00
cwt+plv+sst+spectral+utm	92.90 ± 4.10	87.20 ± 5.70
plv+sst+spectral+utm+gcc+gplv	91.40 ± 2.50	83.80 ± 8.00
cc+cwt+plv+mst+sst+spectral+utm	**94.10 ± 3.10**	86.80 ± 2.40
cc+cwt+plv+mst+sst+spectral+gcc+gplv	92.50 ± 2.20	**89.70 ± 3.50**
cc+cwt+dwt+plv+sst+spectral+utm+gcc+gplv	91.60 ± 3.40	89.40 ± 2.90
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	89.80 ± 3.70	85.10 ± 4.40

The ensemble cc+plv+mst+sst+spectral+utm achieves the highest AUC (94.10%), standing out as the most effective combination under IED-free conditions. Meanwhile, the smaller ensemble plv+sst delivers the top BAC (83.60%), again suggesting that compact configurations can offer superior performance. Ensembles that include Spectral features are consistently strong across both metrics, aligning with findings in the IED-inclusive setting. Performance variability is relatively low overall, with standard deviations around 2–6%, indicating stable results even in the absence of IED-related activity. The ROC curves of each ensemble model presented above are shown in Fig. 6.6.

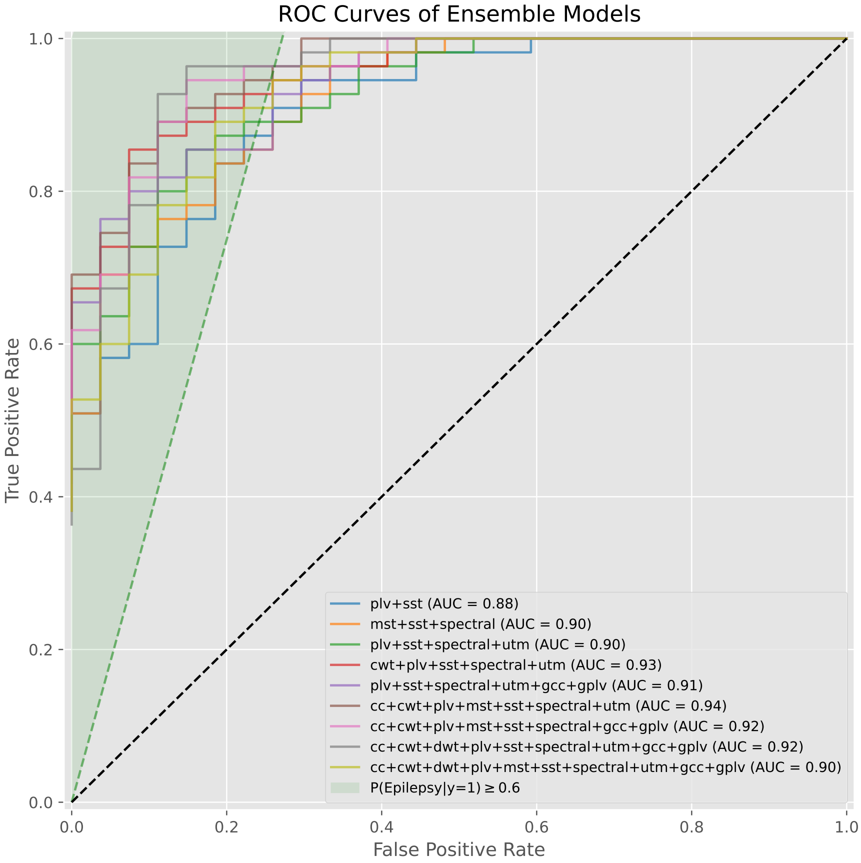

Figure 6.6: TUH ensembles (IED-free): ROC curves for ensemble models trained and evaluated after removing subjects with IEDs.

Results - EMC Dataset 7

This chapter presents the experimental results on the EMC dataset. Similarly to the TUH dataset, the individual feature sets are compared, followed by ensembles of varying lengths built from the best-performing sets identified in the previous stage. Additionally, the best performing individual feature set, an illustration of the predictive value of its features is provided.

7.1 Individual Feature Sets

Using the method describe in Section 4.4 the optimal montage, segment length and statistical combiner are determined for each feature set. A summary of the best performing hyper-parameters is provided in Table 7.1.

Table 7.1: EMC—individual feature sets on whole IPS trials: montage/segment/combiner with performance (mean ± SD).

Feature	**Montage**	**Segment Length**	**Combiner**	**AUC**	**BAC$_{80}$**
Spectral	Cz	10	std	67.08 ± 1.95	61.49 ± 3.02
CWT	BipolarDB	2	median	66.94 ± 1.41	67.38 ± 2.30
DWT	Laplacian	10	median	69.80 ± 1.26	63.17 ± 1.58
MST	BipolarDB	60	median	63.19 ± 1.59	65.15 ± 1.76
SST	CAR	10	median	70.43 ± 22.58	65.70 ± 1.98
CC	CAR	1	std	72.09 ± 1.46	63.27 ± 1.44
PLV	Laplacian	60	kurt	**75.01 ± 2.59**	**70.50 ± 3.53**
GCC	CAR	60	median	67.07 ± 0.61	64.66 ± 2.17
GPLV	Laplacian	2	std	70.18 ± 1.01	61.78 ± 2.10
UTM	BipolarDB	20	std	68.26 ± 2.94	63.47 ± 3.99

PLV achieves the highest AUC, surpassing all other feature sets by a clear margin: the next best, CC and DWT, score about 3% lower on average. Notably, SST exhibits a very high AUC standard deviation, suggesting inconsistent performance across trials. Interestingly, both very short (1s, CC) and moderate (10s, DWT) segment lengths are associated with relatively strong AUC results. The kurtosis combiner used with PLV appears especially effective, reinforcing the idea that higher-order statistics can outperform variance-based measures. Performance variability is evident, with SST and UTM showing particularly high standard deviations, while

PLV remains highly stable. In terms of BAC, PLV also ranks first with an average BAC of 70.5%, further confirming its robustness over competing feature sets.

ROC curves for all feature sets are shown in Fig. 7.1.

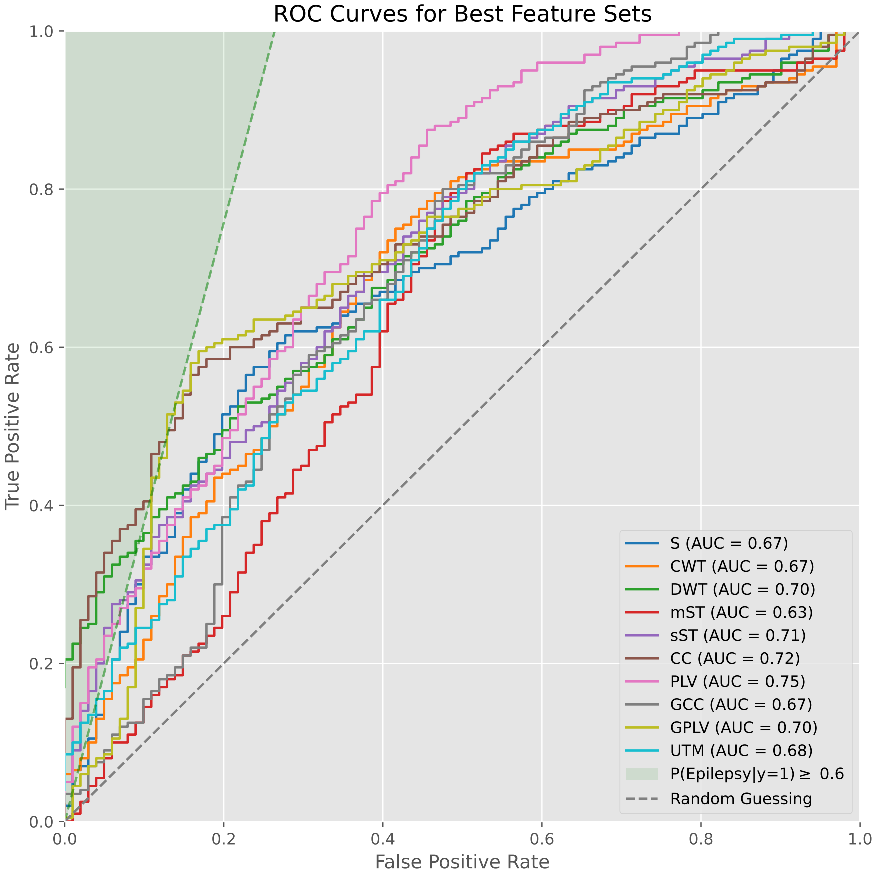

Figure 7.1: EMC: ROC curves for individual feature sets computed over whole IPS trials.

In the SHAP analysis, higher PLV values between O2 and P3 channels in the δ-band, are associated with negative SHAP values, indicating a strong contribution towards the healthy class. Conversely, higher values PLV between O1 and T5 β-band , are linked to positive SHAP values, favoring the epilepsy class. Similarly, higher values of δ-band PLV between FP2 and T5, δ-band PLV between PZ and FZ, and α-band PLV between F4 and FZ, also tend to push predictions towards epilepsy. However, for feature δ-band PLV between PZ and FZ, substantial overlap is observed in the medium-to-low feature range between the two classes, and for α-band PLV between F4 and FZ, in the medium-to-high range, suggesting that the separation between healthy and epilepsy is less clear-cut for these features.

To provide an additional perspective on feature attribution, we computed ALE values for the PLV features with the lowest p-values (Fig. 7.3). From these plots, several considerations emerge. First, the δ-band PLV between FP2–T5 and between

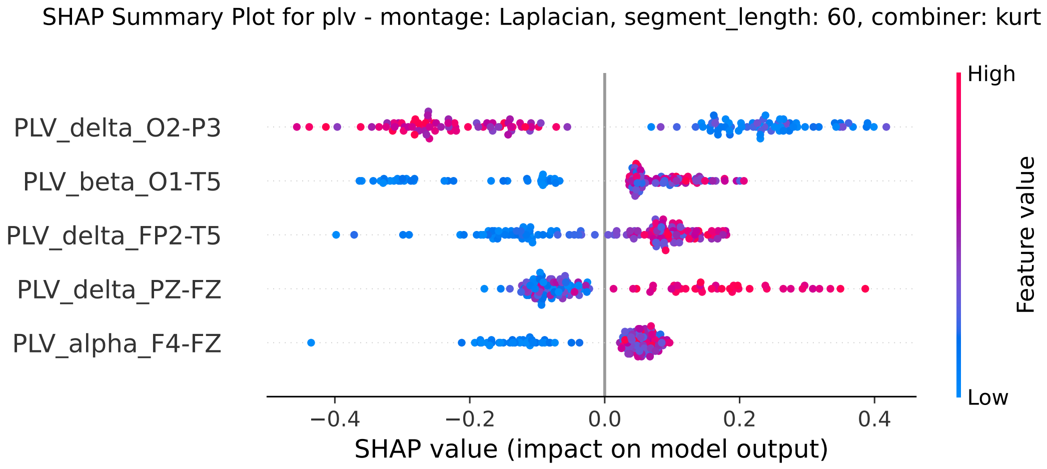

Figure 7.2: PLV set global feature attribution: SHAP values highlighting channel-pair phase-locking features most associated with epilepsy on EMC.

O2–P3 also appear in the SHAP analysis, which is partially consistent with their p-values. Second, the behaviour of these two features is equivalent across both attribution methods, reinforcing the hypothesis that they may be relevant to the classification task, even though their p-values fall below conventional significance thresholds.

7.2 Ensembles

Similarly to the TUH dataset, we construct ensembles of variable size using the best sets identified in the previous section. A summary of the results is shown inTable 7.2.

Table 7.2: EMC ensembles: performance metrics (mean ± SD) for the top stacked models.

Combination	**AUC**	**BAC$_{80}$**
cc+dwt	75.00 ± 3.50	66.30 ± 5.90
cc+mst+gplv	**79.40 ± 5.90**	**73.90 ± 6.80**
dwt+plv+gcc+gplv	77.80 ± 5.50	70.00 ± 11.60
cc+plv+utm+gcc+gplv	76.90 ± 4.00	69.70 ± 6.60
cc+dwt+plv+sst+gcc+gplv	78.90 ± 5.00	71.60 ± 3.70
cc+cwt+plv+sst+spectral+gcc+gplv	77.10 ± 5.10	71.50 ± 7.50
cc+cwt+dwt+plv+sst+spectral+gcc+gplv	76.00 ± 3.00	69.80 ± 4.30
cc+cwt+dwt+plv+mst+sst+spectral+utm+gplv	75.40 ± 2.40	67.20 ± 5.70
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	72.10 ± 6.60	67.20 ± 6.60

After evaluating the ensemble configurations, we observe that performance

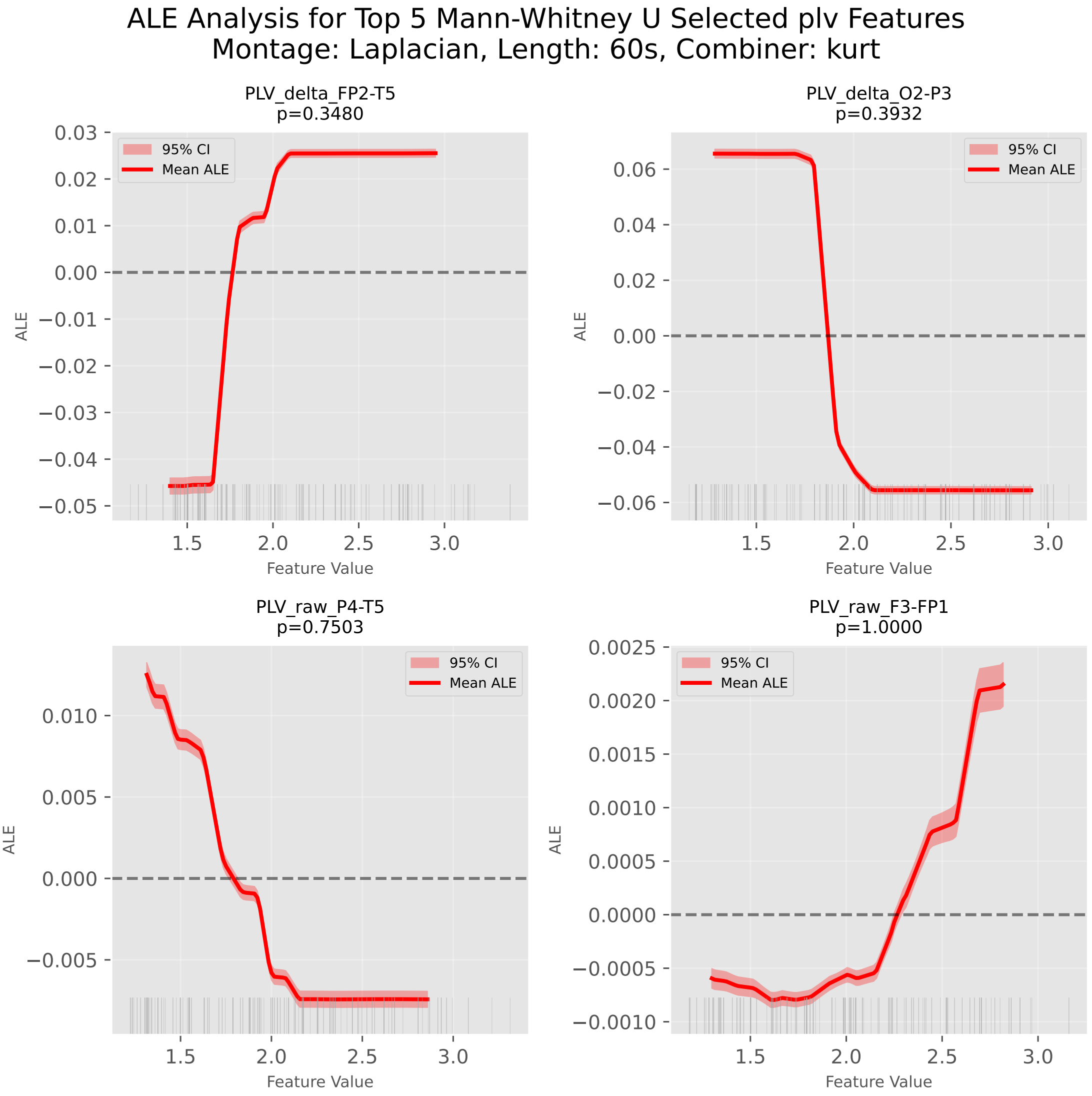

Figure 7.3: PLV set global feature attribution: ALE values for the features with the lowest p-values. Rug plot below the curve indicates the feature value distribution.

varies depending on the combination and size of the feature sets. The three-feature ensemble cc+mst+gplv achieves the highest AUC (79.40%), while the five-feature ensemble cc+plv+utm+gcc+gplv follows closely with an AUC of 76.90%. In terms of balanced accuracy at 80% specificity, the best-performing configuration is the same three-feature ensemble (cc+mst+gplv, 73.90%), slightly outperforming larger ensembles. Interestingly, increasing the number of feature sets does not consistently improve performance, as larger ensembles such as cc +cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv yield lower AUC (72.10%) and BAC (67.20%). Standard deviations remain moderate across ensembles, indicating stable performance with some variability for mid-sized combinations.despite being the strongest individual feature set, the PLV set is absent from the best ensemble configuration.

Overall, these results suggest that compact ensembles, particularly those including connectivity-based features such as cc and gplv, can provide more robust discrimination than more complex configurations. The ROC curves for the models discussed above are presented in Fig. 7.4.

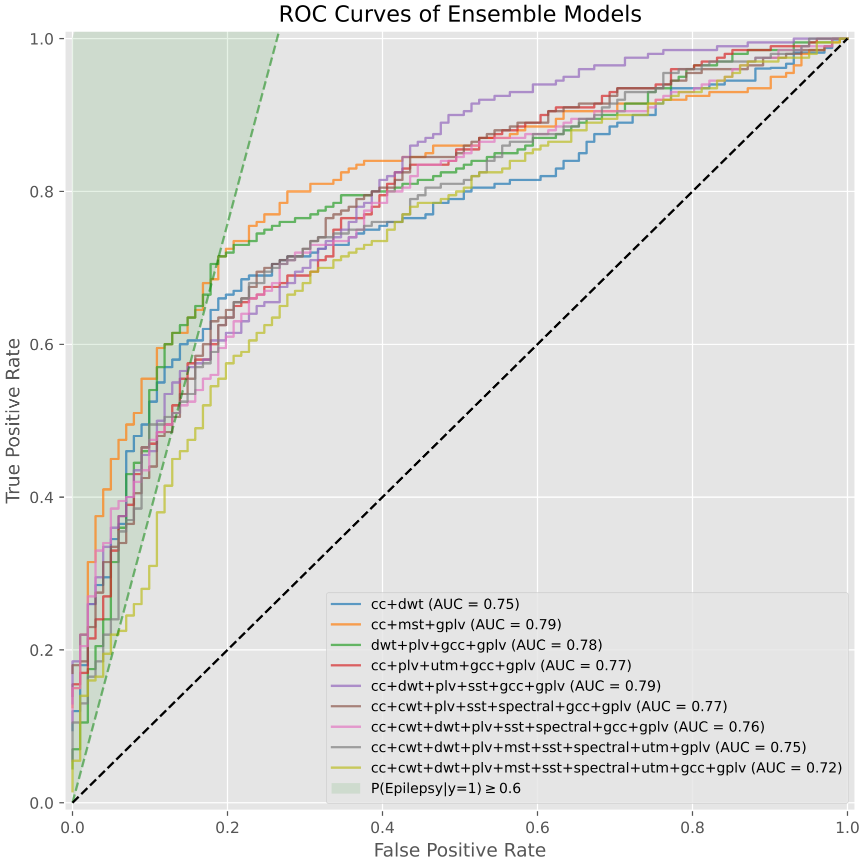

Figure 7.4: EMC ensembles: ROC curves for top-performing stacked models built from selected feature sets.

7.2.1 Unconstrained Ensemble Composition

We ran the ensemble pipeline with the meta-learner's regularization parameter set to $\alpha = 0$, meaning no uniformity constraint was imposed on the ensemble weights. This allowed the weights to potentially collapse onto one or a few dominant models. The training was repeated five times, and we recorded each classifier's weight in the ensemble for every run. We then report the average weights across the five repetitions, along with their respective standard deviations (Fig. 7.5).

On average, the ensemble assigns the greatest weight to the DWT feature set,

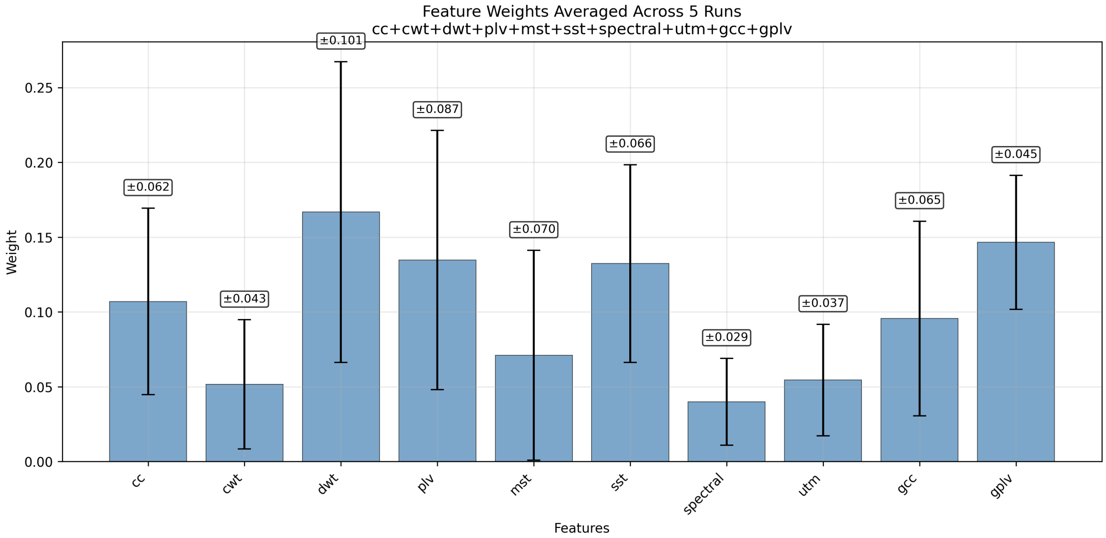

Figure 7.5: EMC ensembles: mean base-learner weights across five training repetitions ($\alpha = 0$); error bars show standard deviation.

followed by GPLV, PLV, SST, and CC. The remaining sets receive comparatively lower weights than expected under a uniform distribution, with notable deviations such as CWT and MST, which frequently receive a weight of zero. Interestingly, the GPLV set is consistently assigned a relatively high weight in the ensemble, despite lacking individually significant features.

Discussion 8

This chapter reflects on the implications of the results outlined in the previous chapter. Rather than focusing only raw numerical comparisons, the discussion later shifts toward clinical considerations. It begins with an evaluation of epoched data versus whole trials, providing justification for the exclusion of the former from the main results. Next, it examines the impact of excluding subjects who exhibit IEDs from the dataset. Finally, the chapter explores clinical perspectives on interpretability techniques and how they aid in understanding the models' decision-making processes, as well as potential integration of these techniques from clinical data.

8.1 Use of Epoched data

Using the features computed over 10 second epochs did not yield any significant performance increase when performing the search for the best individual set, as shown in Section 4.4. A comparison of the mean AUC and BAC, for both dataset is illustrated in Figs. 8.1 and 8.2. After observing this outcome, the use of features computed from 10s segments was not extended to the ensemble method, as it appeared to be an unviable direction. This decision was motivated not only by the lack of performance improvements but also by the substantial increase in sample number, which in turn increased training time more than twofold. Given that the ensemble models already required considerable computation under LOSO-CV, further extension was deemed impractical.

8.2 Data with IEDs vs IED-free

Since the EMC dataset is free of IEDs by design, it is instructive to compare performance on the TUH dataset with and without such discharges. As shown in Table 8.1, some notable differences emerge. It is important to clarify, however, that the IED-free condition was obtained by excluding only two subjects, both of whom were labeled as epileptic.

Overall, while most feature sets show a moderate positive shift in AUC after IED removal, the differences in BAC_{80} are considerably larger. This suggests that, under hypothetical class-balanced conditions, classification may in fact become more challenging when working with IED-free data, depending on the chosen feature representation. A particular case is the GPLV feature set, which shows a

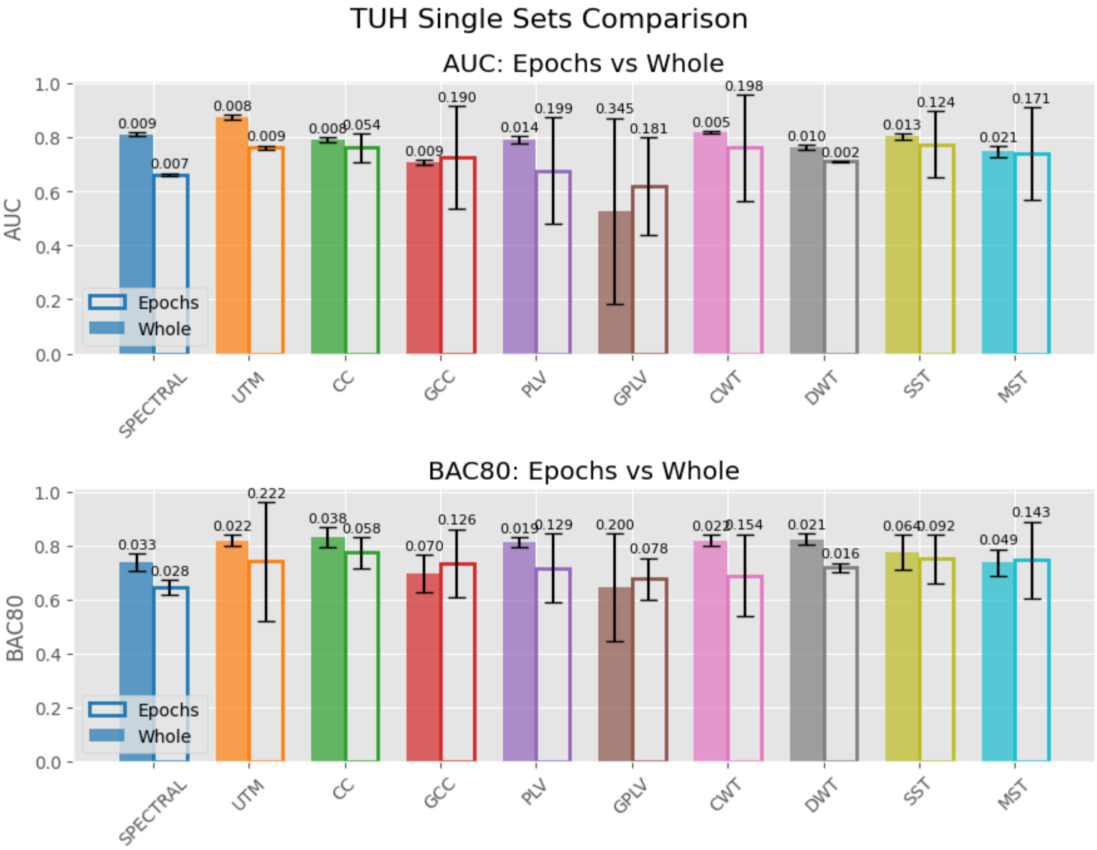

Figure 8.1: TUH: performance comparison (AUC/BAC80, etc.) between epoched IPS segments and whole-trial analysis.

Table 8.1: Change in performance when removing subjects with IEDs (Without – With). Positive values indicate improvement.

Feature	**Δ AUC (pts)**	**Δ BAC (pts)**
Spectral	+3.08	+10.80
UTM	–5.08	–3.96
CWT	–5.44	–11.94
DWT	–1.04	–8.52
MST	–5.71	–3.87
SST	+3.82	+4.78
CC	+3.62	-3.11
PLV	+4.24	–3.34
GCC	+1.60	+5.40
GPLV	+10.60	+6.84

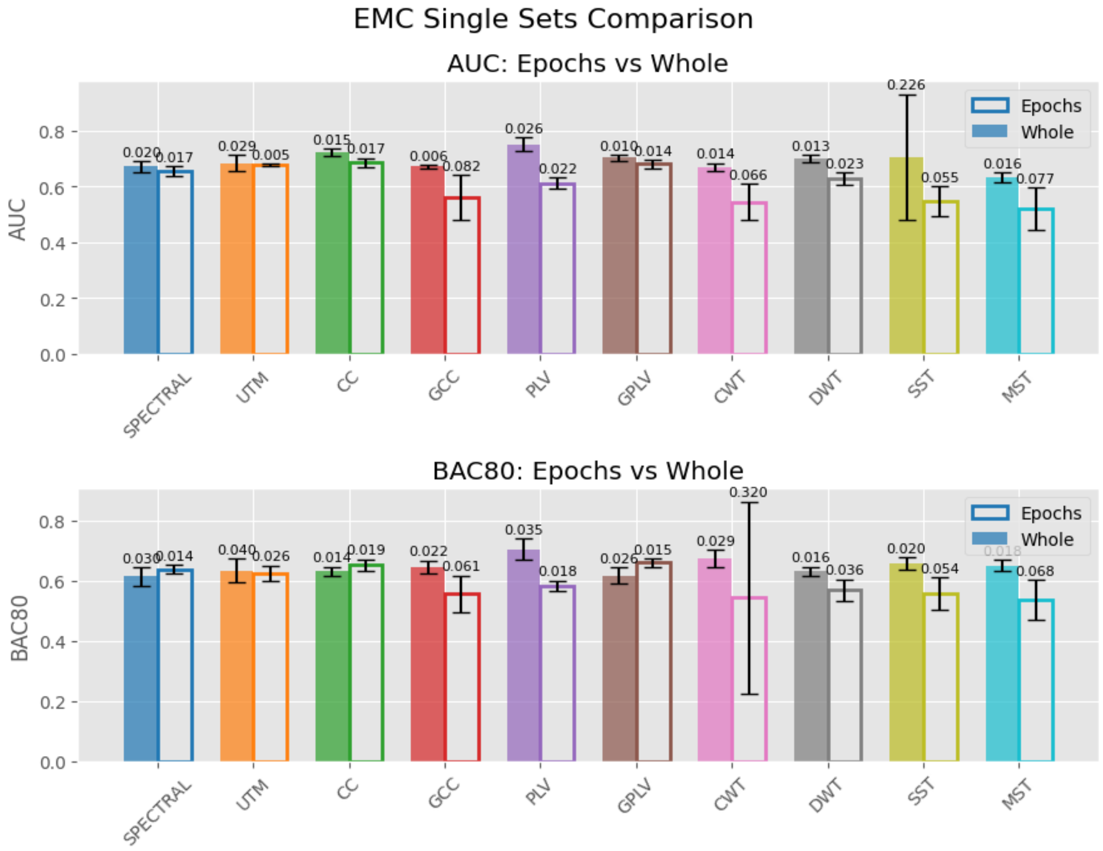

Figure 8.2: EMC: performance comparison between epoched IPS segments and whole-trial analysis.

large apparent improvement; however, this is primarily a consequence of its near chance-level performance in the mixed-data condition (Table 6.1).

Having compared individual feature sets, we now turn to ensembles of multiple feature representations. Tables 6.3 and 6.4 summarize the results for TUH data with and without IEDs, respectively. A direct comparison is provided in Table 8.2.

Beyond the numerical results, it is also important to acknowledge dataset limitations. Although TUH is the largest publicly available EEG repository, the present work employs only a small subset consisting of recordings with IPS procedures. As a result, a minor difference in the predictions has a greater impact on the metrics, therefore conclusions drawn from this comparison should be interpreted with caution.

8.3 TUH vs EMC

When looking at the difference in performance using the same methods for both datasets, a few consideration It is worth noting that TUH data presents much

Table 8.2: Change in ensemble performance when removing IEDs (Without – With). Positive values indicate improvement.

Ensemble Size	**Δ AUC (pts)**	**Δ BAC (pts)**
2	+3.50	+6.30
3	+4.70	+7.20
4	-3.50	-2.90
5	+0.60	+0.30
6	-1.00	+0.40
7	+1.70	+17.50
8	-0.90	+19.60
9	+0.90	+17.80
10	+0.90	+15.50

more separable features, as highlighted in the statistical analysis Chapter 5 and in the corresponding box plots in Appendix A. This difference can be explained by several factors. The EMC dataset was hand-picked and contains recordings from patients who, at the time of acquisition, were deemed not epileptic (these EEGs were obtained after ER access following a first seizure). These patients were only subsequently diagnosed with epilepsy after a second unprovoked seizure, in accordance with ILAE guidelines [62]. Therefore, even specialized neurologists were unable to detect clear abnormalities in the EEGs we analyzed.

The same cannot be said for TUH, which aggregates recordings of diverse origins. For instance, it is unknown whether some patients had previously undergone EEG screenings at other hospitals or had already been diagnosed with epilepsy prior to the recording.

Finally, despite the fact that up to 30% of people affected by psychogenic non-epileptic seizures (PNES) [68] also present comorbidity with epilepsy, a portion of the false positives for epilepsy may instead belong to the remaining 70% of this population. This hypothesis originated while reviewing the work by Faiman et al. [69], who, using ML on EEG features extracted from IED-free intracranial recordings (similar to the case of EMC), concluded that it was not possible to reliably distinguish between the two conditions.

Consequently, correctly classifying the EMC dataset was, by design, a substantially harder task than classifying TUH.

8.4 The Value of IPS

The works summarized in Table 2.1 primarily focus on background EEG analysis, recorded during resting state without external stimulation. In particular, the studies of Thangavel et al. [9], Mirwani [10], and Van der Kleij [11] serve as natural benchmarks: not only do they employ the same datasets, but their pipelines also share a high degree of methodological similarity. This allows us to directly quantify the additional diagnostic value that the IPS procedure can provide when analysed using machine learning techniques.

Regarding the TUH dataset, for data including IEDs, Mirwani's best feature set (spectral features with age and vigilance encoded as additional categorical variables) achieved an AUC of 0.871 and a BAC of 79.1%. In comparison, the best feature set in this work achieved an AUC of 0.875 and a BAC of 81.91%.
For IED-free data, Mirwani reports an AUC of 0.770 and a BAC of 71.6%, while the proposed method achieves an AUC of 0.842 and a BAC of 84.68%.
Thangavel did not report results for individual feature sets on the TUH data.

Looking at the ensemble methods, Thangavel et al. report an AUC of 0.790 and a BAC of 71.7% for recordings including IEDs. By comparison, the proposed method in this work achieves an AUC of 0.936 and a BAC of 90.2% (with the second-best ensemble in terms of AUC, only slightly lower than the best, but achieving the highest BAC). For IED-free data, Thangavel reports an AUC of 0.630 and a BAC of 57.3%, while Mirwani achieves an AUC of 0.760 and a BAC of 72.0%. In contrast, the proposed method reaches a best AUC of 0.941 with a BAC of 86.8%, and, for the model optimized for balanced accuracy, an AUC of 0.925 with a BAC of 89.7%.

It is important to note that in the version of the TUH dataset used in their work, the epileptic class constitutes the majority. In contrast, in the subset of recordings containing the IPS procedure, the epileptic class is a minority, as illustrated in Table 3.1. This class imbalance has a direct impact on performance metrics and makes comparisons across studies less straightforward, as classifiers may behave differently depending on whether the target class is under- or over-represented.

Regarding the EMC dataset, the best single feature set reported by Mirwani achieved an AUC of 0.720, while Van der Kleij reported an AUC of 0.714. Neither of these works reported BAC. In comparison, the proposed method achieves an AUC of 0.750 and a BAC of 70.5%.

For the EMC data, Mirwani reports an AUC of 0.760 and a BAC of 72%, while Van der Kleij reports an AUC of 0.870 (BAC not reported). However, the ensembling method used in the latter suffered from data leakage, as predictions from the best-performing single sets during LOSO CV (equivalent to stage one of this work, described in Section 4.4) were combined by post-hoc averaging . In comparison, the proposed approach achieves an AUC of 0.794 and a BAC of 73.9%.

From these comparisons, several conclusions can be drawn. First, the proposed method consistently outperforms previous approaches on both the TUH and EMC datasets, across single feature sets as well as ensemble methods. This performance gap is particularly pronounced for IED-free data, where the classification task is inherently more challenging: here, the proposed method achieves substantially higher AUC and BAC values compared to both Mirwani and Thangavel, demonstrating the added diagnostic value of incorporating IPS responses into machine learning pipelines.

Second, while single feature sets already show moderate improvements over prior work, the largest performance gains arise from ensemble approaches, highlighting the benefit of combining diverse feature domains in a flexible way. The proposed ensembles not only surpass prior benchmarks but also maintain robustness under leave-one-subject-out validation, ensuring that improvements are not limited to isolated feature sets.

Taken together, these findings provide strong evidence that IPS responses, when analysed with robust methods, can substantially enhance the diagnostic accuracy of EEG classification, particularly in difficult cases such as IED-free recordings.

8.5 Clinical Considerations on ROC Analysis and Feature Attribution

It is important to recall that for each fold of the LOSO cross-validation, the decision threshold was selected to maximize the G-mean score. As a consequence, the recalls and sensitivities achievable with ROC analysis and *a posteriori* threshold tuning differ from those obtained with “hard” predictions, such as the results reported in the appendices. This distinction should be kept in mind when interpreting the following analyses.

According to ROC analysis, it is possible to produce a classifier with a desired sensitivity–specificity trade-off by adjusting the decision threshold at which predicted probability scores are converted into class labels. In this way, the ROC curve effectively defines the set of achievable operating points for a given model.

Starting with the TUH dataset (Figs. 6.1 and 6.3), we observe that all individual feature models, as well as the ensemble models shown in Figs. 6.5 and 6.6, reach the “clinically relevant” region of ROC space, in which the probability of epilepsy conditioned on a positive prediction is greater than 60%.

Regarding the EMC dataset, we observe that for the single feature sets (Fig. 7.1), only the MST set does not reach the clinically relevant region of the ROC space. By contrast, all ensemble models (Fig. 7.4) fall within this meaningful region, demonstrating their added diagnostic value.

However, some models only enter this region near the lower left corner (i.e., at higher thresholds), indicating limited discriminative power in practice. By contrast, models whose curves start at FPR = 0 and TPR > 0 are preferable, as they guarantee full specificity while still achieving clinically meaningful recall.

From a clinical perspective, this flexibility is particularly valuable: depending on the context, thresholds can be chosen to prioritize minimizing false positives (e.g., avoiding unnecessary treatments or follow-up examinations) or maximizing sensitivity (e.g., ensuring early detection of epilepsy even at the cost of some false alarms). Which criterion should be prioritized ultimately depends on the policies and practices of individual clinical centres. A false positive diagnosis may lead to the unnecessary initiation of anti-seizure medication, which is associated with potential adverse effects, while a missed case can severely impact a patient's quality of life if further seizures occur in daily life. The decision on where to place this balance lies beyond the scope of this work and should remain a matter of careful debate within clinical environments.

Feature attribution methods such as SHAP and ALE can indeed provide useful insights into how different input features influence the model's predictions. However, it is essential to emphasize that these methods capture associations between feature values and predicted probabilities at the data-point level: they describe how the model behaves given certain inputs, but they do not establish a causal relationship between features and the predicted outcome.

A further challenge arises when features are correlated. In such cases, attribution methods may distribute "importance" across correlated variables in ways that can be misleading, obscuring the actual mechanisms driving the model's behaviour. SHAP, in particular, may attribute importance to feature combinations that were never present in the training data. ALE mitigates this by dividing the data into quantiles and relying on the observed marginal distributions of feature values.

Approaches such as removing correlated features or projecting them onto an uncorrelated basis, for example using principal component analysis, can reduce this issue and yield clearer patterns of attribution. Nevertheless, these transformations come at the cost of interpretability, and they still do not overcome the fundamental limitation that such methods cannot explain causality.

Another important limitation is that the insights obtained from attribution methods are tied to the data distribution used for training and analysis. If new data points fall outside this distribution, the previously drawn conclusions may no longer hold, and interpretations based on feature importance can easily be proven wrong. This is particularly problematic in clinical applications, where patient heterogeneity and variability across cohorts are to be expected.

These considerations are especially relevant in the present setting, where the features are handcrafted, statistically aggregated descriptors rather than direct physiological measurements. Their clinical meaning is therefore not straightforward, and any discovery-oriented interpretation must be approached with caution. For instance, within the PLV feature set, the kurtosis-based combiner yielded the best predictive performance. However, it would be an overstatement to claim that kurtosis itself, as a statistical measure of Gaussianity, constitutes a biomarker for epilepsy. At most, these results suggest that the model exploits differences in synchronization patterns between specific brain regions. For example, changes in PLV in the δ band between O2 and P3 (Fig. 7.2) may be informative: lower kurtosis values, indicating greater variability or presence of outliers, appear more characteristic of epilepsy, while higher kurtosis values, reflecting stable synchronization around the mean, seem more prominent in healthy subjects.

Thus, the findings should not be interpreted as direct clinical markers but rather as model-driven signals that suggest potentially relevant neurophysiological mechanisms/hypotheses that require independent validation.

9 Conclusion

This chapter discusses limitations and directions for future work and summarizes the main findings of the thesis in relation to the initial research questions, highlighting their scientific and clinical implications.

9.1 Limitations of this study and future directions

Several limitations should be acknowledged, which in turn suggest concrete directions for improvement. First, the study cohorts provide only a restricted view of the broader epilepsy population. Although we used two datasets with different acquisition protocols (TUH and EMC), the analysis remains limited in size and variety. We did not implement leave-one-institution-out validation, and certain subgroups are likely under-represented, for example, patients with different epilepsy types, those under heavy medication, or those with noisy recordings. As a result, the reported performance may overestimate what is achievable in broader clinical practice. Access to data from multiple centres and a broader variety of epilepsy cases would likely improve the generalizability of performance.

Second, there are intrinsically difficult cases in which IPS responses are weak, ambiguous, or strongly contaminated by artifacts such as eye blinks or muscle activity. More sophisticated preprocessing could help by improving the clarity of the EEG signal and better isolating underlying brain activity. However, these approaches are not only computationally costly but also labour-intensive, since they typically require careful tuning and manual supervision, and they carry the risk of discarding meaningful physiological information along with noise.

Another limitation lies in the balance between model capacity and explainability. The use of tree ensembles and linear stacking allowed us to apply attribution methods such as SHAP and ALE, making the models more interpretable. However, these methods trade off raw capacity to capture complex temporal dynamics and high-order interactions in EEG. More powerful models, such as deep temporal networks, could potentially extract richer structure but at the cost of transparency and a higher risk of overfitting given the limited data.

A related limitation concerns the interpretation of feature attributions. Methods such as SHAP and ALE provide associational explanations tied to the fitted model and the observed data distribution, but they do not imply causality. For this reason, biological interpretations based on attribution tools should be regarded as hypothesis-generating rather than conclusive evidence. They are highly sensitive to feature correlations and may yield misleading importance rankings under distribution shifts, meaning that results may not generalize well to other populations.

Multicollinearity remains an unresolved challenge: many features (i.e. adjacent frequency bands or related graph metrics) are strongly correlated. This issue could be addressed with dimensionality reduction techniques; for instance, principal component analysis produces uncorrelated orthogonal components, though this comes at the expense of interpretability since the composite features no longer map directly to physiological constructs.

Model development can also be refined. Fine-tuning hyperparameters systematically could reduce variance and squeeze more performance from simpler models.

Finally, there is a need to test attribution-derived hypotheses in a more causality-aware framework. This may involve the hand-picking of individual feature from each feature family to create a heterogeneous set of candidate,possibly uncorrelated, markers and the use of more reliable causal xAI/feature attribution methods.

In short, while this thesis shows that EEG, in particular IPS segments, can be used effectively for epilepsy classification, progress will depend on specific preprocessing, more careful handling of feature redundancy, and rigorous multi-site validation. Attribution tools remain useful but should be treated as hypothesis generators until supported by causal evidence.

9.2 Summary

This thesis demonstrates that EEG recorded during intermittent photic stimulation (IPS) contains relevant information for diagnosing epilepsy, even in the absence of IEDs. Using subject-level pipelines across two independent cohorts (TUH and EMC), the analysis showed that spectral features, measures of synchronization (i.e. phase-locking value), and simple temporal statistics consistently differentiated between epileptic and non-epileptic subjects. On TUH, compact ensemble models achieved strong discrimination (AUC = 0.94; BAC = 90%), while on EMC the same framework reached AUC = 0.79 and BAC = 74%, outperforming prior work under comparable validation strategies.

These results confirm that IPS-driven responses generalize across datasets, though with expected differences in performance due to cohort characteristics. Feature attribution further suggested candidate electrophysiological markers, such as altered synchronization in posterior brain networks and band-specific spectral modulations, though these should be interpreted as model-driven hypotheses rather than established biomarkers. Overall, the work advances interpretable approaches to automated diagnosis, supporting its potential to enrich routine diagnostic workflows, while emphasizing the importance of external validation and careful clinical translation.

Bibliography

[1] *Epilepsy*, en. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy (cit. on p. 1).

[2] E. Hirsch *et al.*, 'ILAE definition of the Idiopathic Generalized Epilepsy Syndromes: Position statement by the ILAE Task Force on Nosology and Definitions,' en, *Epilepsia*, vol. 63, no. 6, pp. 1475–1499, Jun. 2022, ISSN: 0013-9580, 1528-1167. DOI: 10.1111/epi.17236 (cit. on p. 1).

[3] R. S. Fisher *et al.*, 'Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE),' en, *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005, ISSN: 1528-1167. DOI: 10.1111/j.0013-9580.2005.66104.x (cit. on p. 1).

[4] R. Basiri, A. Shariatzadeh, S. Wiebe and Y. Aghakhani, 'Focal epilepsy without interictal spikes on scalp EEG: A common finding of uncertain significance,' *Epilepsy Research*, vol. 150, pp. 1–6, Feb. 2019, ISSN: 0920-1211. DOI: 10.1016/j.eplepsyres.2018.12.009 (cit. on p. 1).

[5] J. T. Dell'Aquila and V. Soti, 'Sleep deprivation: A risk for epileptic seizures,' *Sleep Science*, vol. 15, no. 2, pp. 245–249, 2022, ISSN: 1984-0659. DOI: 10.5935/1984-0063.20220046 (cit. on p. 1).

[6] R. M. Epstein, 'Facing epistemic and complex uncertainty in serious illness: The role of mindfulness and shared mind,' eng, *Patient Education and Counseling*, vol. 104, no. 11, pp. 2635–2642, Nov. 2021, ISSN: 1873-5134. DOI: 10.1016/j.pec.2021.07.030 (cit. on p. 1).

[7] A. T. Berg, 'Risk of recurrence after a first unprovoked seizure,' en, *Epilepsia*, vol. 49, no. s1, pp. 13–18, 2008, ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2008.01444.x (cit. on p. 2).

[8] *Epilepsie - elektrofysiologisch onderzoek bij epilepsie*, en. [Online]. Available: https://richtlijnendatabase.nl/richtlijn/epilepsie/elektrofysiologisch_onderzoek_bij_epilepsie.html (cit. on p. 2).

[9] P. Thangavel *et al.*, 'Improving automated diagnosis of epilepsy from EEGs beyond IEDs,' en, *Journal of Neural Engineering*, vol. 19, no. 6, p. 066 017, Dec. 2022, ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ac9c93 (cit. on pp. 2, 6, 14, 54).

[10] Y. Mirwani, 'Automated Epilepsy Diagnosis beyond IEDs by Multimodal Features and Deep Learning,' en, M.S. thesis, TU Delft, Delft, 2024. [Online]. Available: https://resolver.tudelft.nl/uuid:c829feac-3482-47a3-9c3e-2e27e89056c0 (cit. on pp. 2, 6, 28, 54).

[11] P. A. van der Kleij, 'Using machine learning models trained on IED-free EEGs to support epilepsy diagnosis,' en, M.S. thesis, TU Delft, Delft, 2025. [Online]. Available:

https://repository.tudelft.nl/record/uuid:e89c0857-496b-40a4-9361-c5a94680b908 (cit. on pp. 2, 6, 23, 28, 54).

[12] J. P. Carvajal-Dossman *et al.*, 'Retraining and evaluation of machine learning and deep learning models for seizure classification from EEG data,' en, *Scientific Reports*, vol. 15, no. 1, p. 15 345, May 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-98389-y (cit. on p. 4).

[13] Y. Shin *et al.*, 'Using spectral and temporal filters with EEG signal to predict the temporal lobe epilepsy outcome after antiseizure medication via machine learning,' en, *Scientific Reports*, vol. 13, no. 1, p. 22 532, Dec. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-49255-2 (cit. on p. 4).

[14] M. C. Tjepkema-Cloostermans *et al.*, 'Expert level of detection of interictal discharges with a deep neural network,' *Epilepsia*, vol. 66, no. 1, pp. 184–194, Jan. 2025, ISSN: 0013-9580. DOI: 10.1111/epi.18164 (cit. on p. 4).

[15] *UC Davis Department of Neurology - Epilepsy FAQs.* [Online]. Available: https://health.ucdavis.edu/neurology/subspecialties/epilepsy_faqs.html (cit. on p. 4).

[16] S. R. Benbadis, 'The Best Seizure Diagnostic Tool Is Not a Medical Device,' *Neurology: Clinical Practice*, vol. 13, no. 1, e200117, Feb. 2023, ISSN: 2163-0402. DOI: 10.1212/CPJ.0000000000200117 (cit. on p. 4).

[17] J. Pillai and M. R. Sperling, 'Interictal EEG and the Diagnosis of Epilepsy,' en, *Epilepsia*, vol. 47, no. s1, pp. 14–22, 2006, ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2006.00654.x (cit. on p. 4).

[18] Z. Zou, B. Chen, D. Xiao, F. Tang and X. Li, 'Accuracy of Machine Learning in Detecting Pediatric Epileptic Seizures: Systematic Review and Meta-Analysis,' EN, *Journal of Medical Internet Research*, vol. 26, no. 1, e55986, Dec. 2024. DOI: 10.2196/55986 (cit. on p. 5).

[19] Z. Khan, A. Dayal and H.-C. Kim, 'An Attention-Enhanced 3D-CNN Framework for Spectrogram-Based EEG Analysis in Epilepsy Detection,' *IEEE Access*, pp. 1–1, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3574646 (cit. on p. 5).

[20] T. Tuncer and S. Dogan, 'An explainable EEG epilepsy detection model using friend pattern,' en, *Scientific Reports*, vol. 15, no. 1, p. 16 951, May 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-01747-z (cit. on p. 5).

[21] S. Wong *et al.*, 'Channel-annotated deep learning for enhanced interpretability in EEG-based seizure detection,' *Biomedical Signal Processing and Control*, vol. 103, p. 107 484, May 2025, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2024.107484 (cit. on p. 5).

[22] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David and C. E. Elger, 'Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state,' *Physical*

Review E, vol. 64, no. 6, p. 061 907, Nov. 2001. DOI: 10.1103/PhysRevE.64.061907 (cit. on p. 5).

[23] S. A. Zendehbad, A. S. Razavi, N. Tabrizi and Z. Sedaghat, 'A systematic review of artificial intelligence techniques based on electroencephalography analysis in the diagnosis of epilepsy disorders: A clinical perspective,' *Epilepsy Research*, vol. 215, p. 107 582, Sep. 2025, ISSN: 0920-1211. DOI: 10.1016/j.eplepsyres.2025.107582 (cit. on p. 5).

[24] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis and J. S. Suri, 'Automated EEG analysis of epilepsy: A review,' *Knowledge-Based Systems*, vol. 45, pp. 147–165, Jun. 2013, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2013.02.014 (cit. on p. 5).

[25] S. Wong *et al.*, 'EEG datasets for seizure detection and prediction— A review,' *Epilepsia Open*, vol. 8, no. 2, pp. 252–267, Feb. 2023, ISSN: 2470-9239. DOI: 10.1002/epi4.12704 (cit. on p. 5).

[26] P. Myers *et al.*, 'Diagnosing Epilepsy with Normal Interictal EEG Using Dynamic Network Models,' en, *Annals of Neurology*, vol. 97, no. 5, pp. 907–918, May 2025, ISSN: 0364-5134, 1531-8249. DOI: 10.1002/ana.27168 (cit. on pp. 5, 6).

[27] G. R. N, C. Debnath, D. Guha, A. Adhya and M. Mahadevappa, 'Epileptic EEG Signals Classification Based on Multi-Criteria Decision Aid Classifier Ensemble Approach,' in *2024 National Conference on Communications (NCC)*, Feb. 2024, pp. 1–6. DOI: 10.1109/NCC60321.2024.10485927 (cit. on p. 5).

[28] J. Cao *et al.*, 'Using interictal seizure-free EEG data to recognise patients with epilepsy based on machine learning of brain functional connectivity,' en, *Biomedical Signal Processing and Control*, vol. 67, p. 102 554, May 2021, ISSN: 17468094. DOI: 10.1016/j.bspc.2021.102554 (cit. on p. 5).

[29] S. Kunjan *et al.*, 'The Necessity of Leave One Subject Out (LOSO) Cross Validation for EEG Disease Diagnosis,' en, in *Brain Informatics*, M. Mahmud, M. S. Kaiser, S. Vassanelli, Q. Dai and N. Zhong, Eds., Cham: Springer International Publishing, 2021, pp. 558–567, ISBN: 978-3-030-86993-9. DOI: 10.1007/978-3-030-86993-9_50 (cit. on p. 5).

[30] I. Tougui, A. Jilbab and J. El Mhamdi, 'Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications,' *Healthcare Informatics Research*, vol. 27, no. 3, pp. 189–199, Jul. 2021, ISSN: 2093-3681. DOI: 10.4258/hir.2021.27.3.189 (cit. on pp. 5, 31).

[31] P. Thangavel *et al.*, 'Time–frequency decomposition of scalp electroencephalograms improves deep learning-based epilepsy diagnosis,' *International journal of neural systems*, vol. 31, no. 08, p. 2 150 032, 2021 (cit. on p. 6).

[32] P. Jayakar and K. H. Chiappa, 'Clinical correlations of photoparoxysmal responses,' *Electroencephalography and Clinical Neurophysiology*, vol. 75, no. 3, pp. 251–254, Mar. 1990, ISSN: 0013-4694. DOI: 10.1016/0013-4694(90)90178-M (cit. on p. 7).

[33] D. Kasteleijn-Nolst Trenité *et al.*, ‘Methodology of photic stimulation revisited: Updated European algorithm for visual stimulation in the EEG laboratory,’ en, *Epilepsia*, vol. 53, no. 1, pp. 16–24, Jan. 2012, ISSN: 0013-9580, 1528-1167. DOI: 10.1111/j.1528-1167.2011.03319.x (cit. on p. 7).

[34] *10–20 system (EEG)*, en, Jul. 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=10%E2%80%9320_system_(EEG)&oldid=1234311940 (cit. on p. 8).

[35] O. Mecarelli, Ed., *Clinical Electroencephalography*, en. Cham: Springer International Publishing, 2019, ISBN: 978-3-030-04572-2. DOI: 10.1007/978-3-030-04573-9 (cit. on pp. 9, 10).

[36] S. H.-F. Syam, H. Lakany, R. B. Ahmad and B. A. Conway, ‘Comparing Common Average Referencing to Laplacian Referencing in Detecting Imagination and Intention of Movement for Brain Computer Interface,’ en, *MATEC Web of Conferences*, vol. 140, p. 01 028, 2017, ISSN: 2261-236X. DOI: 10.1051/matecconf/201714001028 (cit. on pp. 9, 10).

[37] D. J. McFarland, L. M. McCane, S. V. David and J. R. Wolpaw, ‘Spatial filter selection for eeg-based communication,’ *Electroencephalography and clinical Neurophysiology*, vol. 103, no. 3, pp. 386–394, 1997 (cit. on p. 9).

[38] J. N. Acharya and V. J. Acharya, ‘Overview of EEG Montages and Principles of Localization,’ en-US, *Journal of Clinical Neurophysiology*, vol. 36, no. 5, p. 325, Sep. 2019, ISSN: 0736-0258. DOI: 10.1097/WNP.0000000000000538 (cit. on p. 9).

[39] A. Othmani, A. Q. M. Sabri, S. Aslan, F. Chaieb, R. Rameh Halaand Alfred and D. Cohen, ‘EEG-based neural networks approaches for fatigue and drowsiness detection,’ *Neurocomputing*, vol. 557, no. C, Nov. 2023. DOI: 10.1016/j.neucom.2023.126709 (cit. on p. 12).

[40] J. M. O’Toole, A. Temko and N. Stevenson, ‘Assessing instantaneous energy in the EEG: A non-negative, frequency-weighted energy operator,’ in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 3288–3291. DOI: 10.1109/EMBC.2014.6944325 (cit. on p. 13).

[41] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid and J. Picone, ‘Improved EEG Event Classification Using Differential Energy,’ eng, *... IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE Signal Processing in Medicine and Biology Symposium*, vol. 2015, Dec. 2015, ISSN: 2372-7241. DOI: 10.1109/SPMB.2015.7405421 (cit. on p. 13).

[42] D. Phung, D. Tran, W. Ma, P. Nguyen and T. Pham, ‘Using Shannon Entropy as EEG Signal Feature for Fast Person Identification,’ en, *Computational Intelligence*, 2014 (cit. on p. 13).

[43] N. Bajaj, ‘Wavelets for EEG Analysis,’ en, in *Wavelet Theory*, IntechOpen, Nov. 2020, ISBN: 978-1-83881-948-4. DOI: 10.5772/intechopen.94398 (cit. on p. 14).

[44] A. K. Singh and S. Krishnan, 'Trends in EEG signal feature extraction applications,' English, *Frontiers in Artificial Intelligence*, vol. 5, Jan. 2023, ISSN: 2624-8212. DOI: 10.3389/frai.2022.1072801 (cit. on p. 14).

[45] Ö. Türk and M. S. Özerdem, 'Epilepsy Detection by Using Scalogram Based Convolutional Neural Network from EEG Signals,' en, *Brain Sciences*, vol. 9, no. 5, p. 115, May 2019. DOI: 10.3390/brainsci9050115 (cit. on p. 14).

[46] E. Khoursheed and A. Eesa, *EEGs Feature Extraction by Multi-Level DWT with Different Numbers of Principal Components.* Apr. 2019. DOI: 10.1109/ICOASE.2019.8723789 (cit. on p. 15).

[47] O. Rioul and P. Duhamel, 'Fast algorithms for discrete and continuous wavelet transforms,' *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 569–586, Mar. 1992, ISSN: 1557-9654. DOI: 10.1109/18.119724 (cit. on p. 15).

[48] K. P. Indiradevi, E. Elias, P. S. Sathidevi, S. Dinesh Nayak and K. Radhakrishnan, 'A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram,' *Computers in Biology and Medicine*, vol. 38, no. 7, pp. 805–816, Jul. 2008, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2008.04.010 (cit. on p. 15).

[49] R. Stockwell, L. Mansinha and R. Lowe, 'Localization of the complex spectrum: The S transform,' *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 998–1001, Apr. 1996, ISSN: 1941-0476. DOI: 10.1109/78.492555 (cit. on p. 16).

[50] M. Hariharan, V. Vijean, R. Sindhu, P. Divakar, A. Saidatul and S. Yaacob, 'Classification of mental tasks using stockwell transform,' *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1741–1749, Jul. 2014, ISSN: 0045-7906. DOI: 10.1016/j.compeleceng.2014.01.010 (cit. on p. 16).

[51] A. H. Mooij, B. Frauscher, J. Gotman and G. J. M. Huiskamp, 'A skew-based method for identifying intracranial EEG channels with epileptic activity without detecting spikes, ripples, or fast ripples,' *Clinical Neurophysiology*, vol. 131, no. 1, pp. 183–192, Jan. 2020, ISSN: 1388-2457. DOI: 10.1016/j.clinph.2019.10.025 (cit. on p. 16).

[52] H. Kalbkhani and M. G. Shayesteh, 'Stockwell transform for epileptic seizure detection from EEG signals,' *Biomedical Signal Processing and Control*, vol. 38, pp. 108–118, Sep. 2017, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2017.05.008 (cit. on p. 16).

[53] J. P. Lachaux, E. Rodriguez, J. Martinerie and F. J. Varela, 'Measuring phase synchrony in brain signals,' eng, *Human Brain Mapping*, vol. 8, no. 4, pp. 194–208, 1999, ISSN: 1065-9471. DOI: 10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c (cit. on p. 17).

[54] P. G. Stoica, R. Moses, P. Stoica and R. L. Moses, *Spectral analysis of signals*, en. Upper Saddle River, NJ: Pearson, Prentice Hall, 2005, ISBN: 978-0-13-113956-5 (cit. on p. 17).

[55] E. D. Fagerholm, P. J. Hellyer, G. Scott, R. Leech and D. J. Sharp, ‘Disconnection of network hubs and cognitive impairment after traumatic brain injury,’ eng, *Brain: A Journal of Neurology*, vol. 138, no. Pt 6, pp. 1696–1709, Jun. 2015, ISSN: 1460-2156. DOI: 10.1093/brain/awv075 (cit. on p. 17).

[56] M. Rubinov and O. Sporns, ‘Complex network measures of brain connectivity: Uses and interpretations,’ *NeuroImage*, Computational Models of the Brain, vol. 52, no. 3, pp. 1059–1069, Sep. 2010, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2009.10.003 (cit. on pp. 17, 18).

[57] aestrivex, *Aestrivex/bctpy - brain connectivity toolbox for python*, Jun. 2025. [Online]. Available: https://github.com/aestrivex/bctpy (visited on 25/06/2025) (cit. on p. 18).

[58] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System,’ in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, arXiv:1603.02754 [cs], Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. [Online]. Available: http://arxiv.org/abs/1603.02754 (visited on 28/09/2025) (cit. on p. 19).

[59] S. M. Lundberg and S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions,’ in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017 (cit. on p. 20).

[60] D. W. Apley and J. Zhu, ‘Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models,’ *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020, ISSN: 1369-7412. DOI: 10.1111/rssb.12377. [Online]. Available: https://doi.org/10.1111/rssb.12377 (visited on 25/09/2025) (cit. on p. 20).

[61] A. Tharwat, ‘Classification assessment methods,’ en, *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021, ISSN: 2634-1964, 2210-8327. DOI: 10.1016/j.aci.2018.08.003 (cit. on p. 22).

[62] R. S. Fisher *et al.*, ‘ILAE Official Report: A practical clinical definition of epilepsy,’ en, *Epilepsia*, vol. 55, no. 4, pp. 475–482, 2014, ISSN: 1528-1167. DOI: 10.1111/epi.12550 (cit. on pp. 23, 54).

[63] P. Wilson, ‘An Updated Wilcoxon–Mann–Whitney Test,’ en, in *Developments in Statistical Modelling*, Springer, Cham, 2024, pp. 159–165, ISBN: 978-3-031-65723-8. DOI: 10.1007/978-3-031-65723-8_25 (cit. on p. 24).

[64] K. J. Berry, J. E. Johnston and P. W. Mielke, ‘Ordinal-Level Variables, I,’ en, in *The Measurement of Association: A Permutation Statistical Approach*, K. J. Berry, J. E. Johnston and J. Mielke Paul W., Eds., Cham: Springer International Publishing, 2018, pp. 223–295, ISBN: 978-3-319-98926-6. DOI: 10.1007/978-3-319-98926-6_5 (cit. on p. 25).

[65] I. Obeid and J. Picone, 'The Temple University Hospital EEG Data Corpus,' English, *Frontiers in Neuroscience*, vol. 10, May 2016, ISSN: 1662-453X. DOI: `10.3389/fnins.2016.00196` (cit. on p. 26).

[66] *myDRE platform*, nl-NL. [Online]. Available: `https://andrea-cloud.com/mydre-platform/` (cit. on p. 28).

[67] J. Thomas *et al.*, 'Automated Detection of Interictal Epileptiform Discharges from Scalp Electroencephalograms by Convolutional Neural Networks,' *International Journal of Neural Systems*, vol. 30, no. 11, p. 2 050 030, Nov. 2020, ISSN: 0129-0657. DOI: `10.1142/S0129065720500306` (cit. on p. 30).

[68] S. Popkirov *et al.*, 'The aetiology of psychogenic non-epileptic seizures: Risk factors and comorbidities,' en, *Epileptic Disorders*, vol. 21, no. 6, pp. 529–547, 2019, ISSN: 1950-6945. DOI: `10.1684/epd.2019.1107` (cit. on p. 54).

[69] I. Faiman *et al.*, 'Limited clinical validity of univariate resting-state EEG markers for classifying seizure disorders,' *Brain Communications*, vol. 5, no. 6, fcad330, Dec. 2023, ISSN: 2632-1297. DOI: `10.1093/braincomms/fcad330` (cit. on p. 54).

Statistical Analysis - Supplementary Tables and Figures

A

Table A.1: TUH statistical details: feature-wise effect sizes and p-values.

Feature	Montage	Segment Length	Combiner	Index	P-value	Effect Size
cc	BipolarDB	2	mean	37	0.0266	-0.74
cc	BipolarDB	5	mean	37	0.0375	-0.73
cc	BipolarDB	10	mean	37	0.0375	-0.73
cc	BipolarDB	20	mean	37	0.0338	-0.73
cc	BipolarDB	60	mean	37	0.0375	-0.73
cwt	CAR	5	skew	194	0.0305	0.79
cwt	CAR	5	kurt	194	0.0390	0.78
cwt	CAR	20	skew	194	0.0481	0.78
cwt	BipolarDB	5	skew	298	0.0453	0.77
cwt	BipolarDB	20	kurt	350	0.0453	0.77
cwt	Cz	20	kurt	246	0.0478	0.77
cwt	Laplacian	5	median	481	0.0376	0.78
dwt	Cz	10	skew	76	0.0442	0.77
dwt	Cz	20	skew	76	0.0204	0.80
dwt	Cz	60	skew	77	0.0392	0.80
gcc	CAR	1	std	4	0.0363	0.62
sst	Cz	10	std	86	0.0436	0.70
sst	Cz	10	std	100	0.0436	0.70
sst	Cz	10	std	100	0.0436	0.70
sst	Laplacian	1	kurt	4	0.0313	-0.72
sst	Laplacian	5	skew	105	0.0487	0.70
sst	Laplacian	10	std	47	0.0391	0.79
sst	Laplacian	20	skew	23	0.0436	0.70
spectral	CAR	10	std	37	0.0245	-0.72
spectral	CAR	60	std	12	0.0452	-0.69
spectral	Laplacian	5	mean	40	0.0391	-0.69
spectral	Laplacian	10	mean	37	0.0245	-0.72
spectral	Laplacian	60	skew	70	0.0364	-0.70
utm	BipolarDB	1	median	140	0.0342	-0.75
utm	BipolarDB	20	std	95	0.0456	-0.74
utm	Cz	1	mean	40	0.0148	-0.79
utm	Cz	1	median	205	0.0182	-0.78

Table A.2: EMC statistical details: feature-wise effect sizes and p-values

Feature	Montage	Segment Length	Combiner	Index	P-value	Effect Size
cc	CAR	20	skew	59	0.01956	0.42
cc	BipolarDB	10	std	36	0.04670	0.39
cwt	BipolarDB	20	skew	326	0.01746	0.45
cwt	Cz	20	skew	222	0.01746	0.45
dwt	Laplacian	2	skew	6	0.006158	0.47
dwt	Laplacian	2	skew	7	0.02294	0.43
dwt	Laplacian	2	kurt	6	0.002747	0.49
gcc	CAR	20	kurt	6	0.01698	-0.36
plv	BipolarDB	1	skew	335	0.02298	0.46
plv	Cz	60	std	330	0.02275	-0.45
mst	BipolarDB	1	skew	54	0.04513	0.38
sst	CAR	5	std	1	0.04602	-0.34
spectral	Cz	5	kurt	24	0.03508	0.34
spectral	Laplacian	1	std	18	0.03337	-0.39
utm	CAR	60	skew	63	0.04907	-0.40
utm	Cz	20	kurt	63	0.02531	0.42

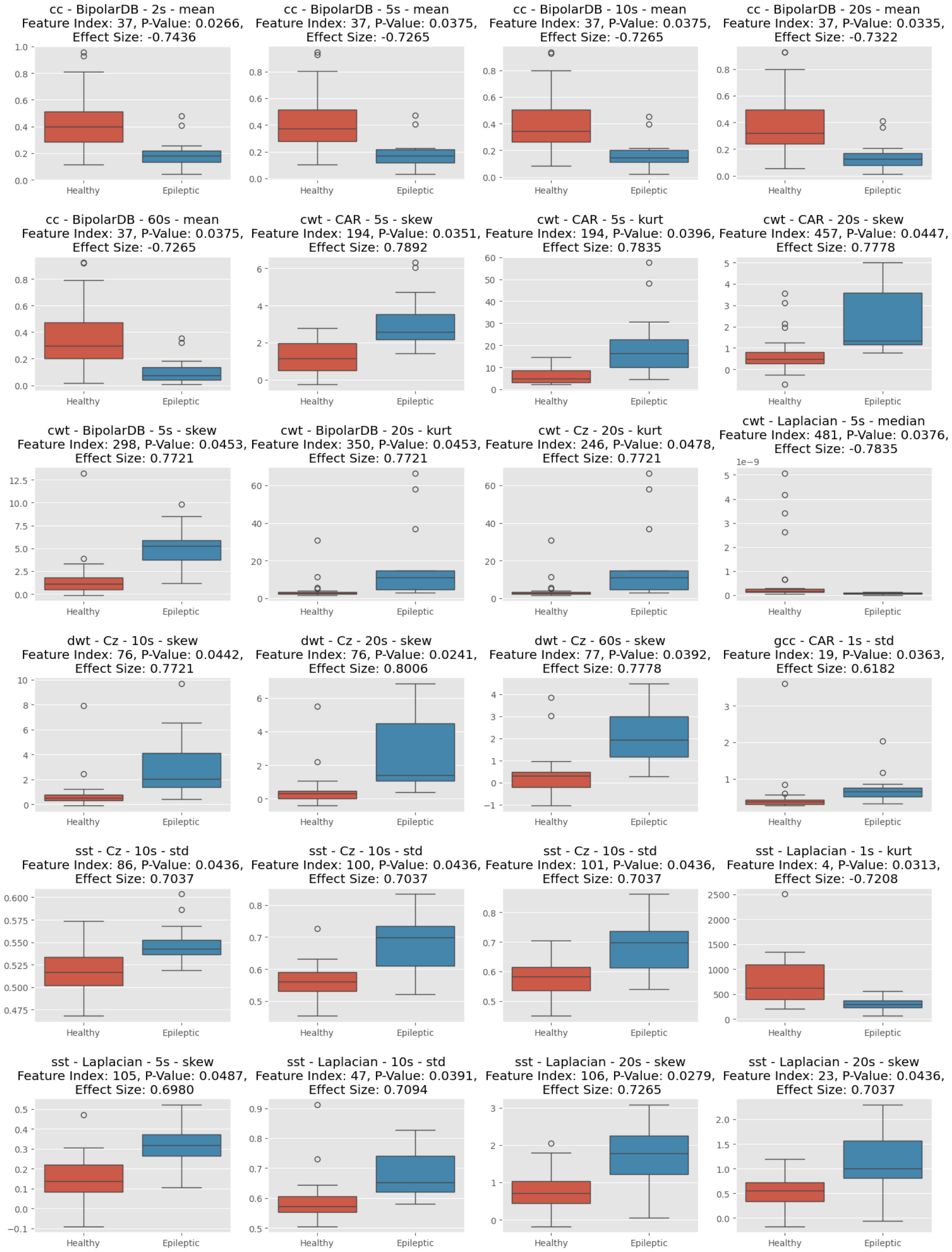

Figure A.1: TUH statistical analysis: box-plots for the first 24 significant features (effect sizes & p-values in Table A.1).

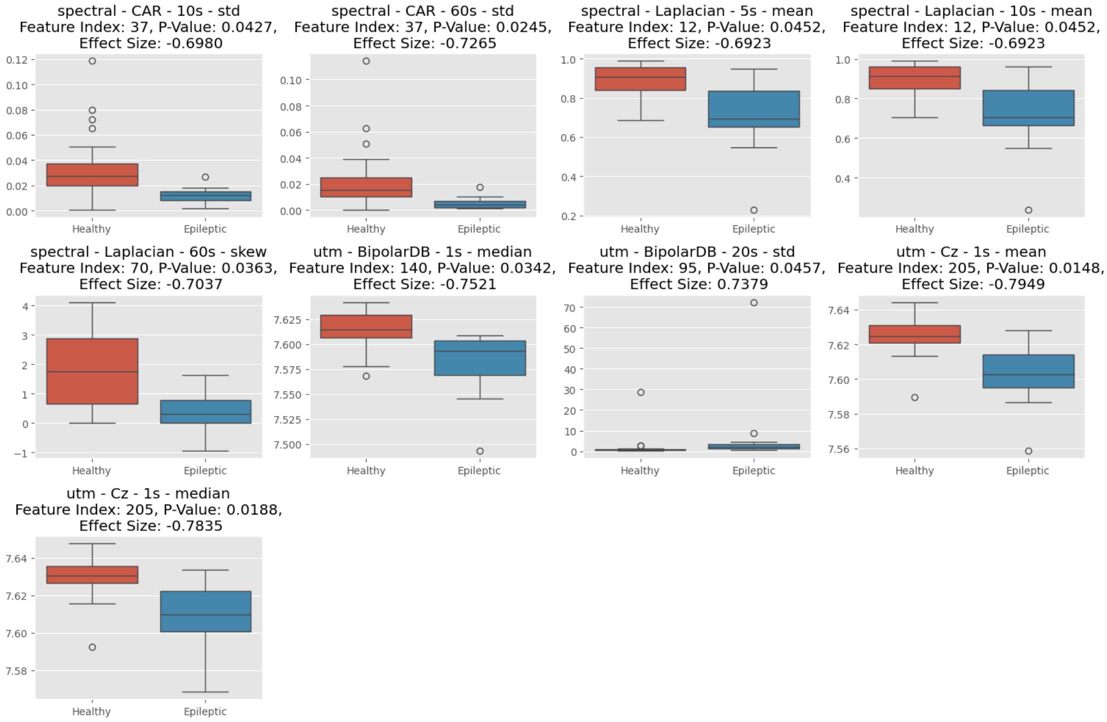

Figure A.2: TUH statistical analysis: box-plots for the remaining 9 significant features.

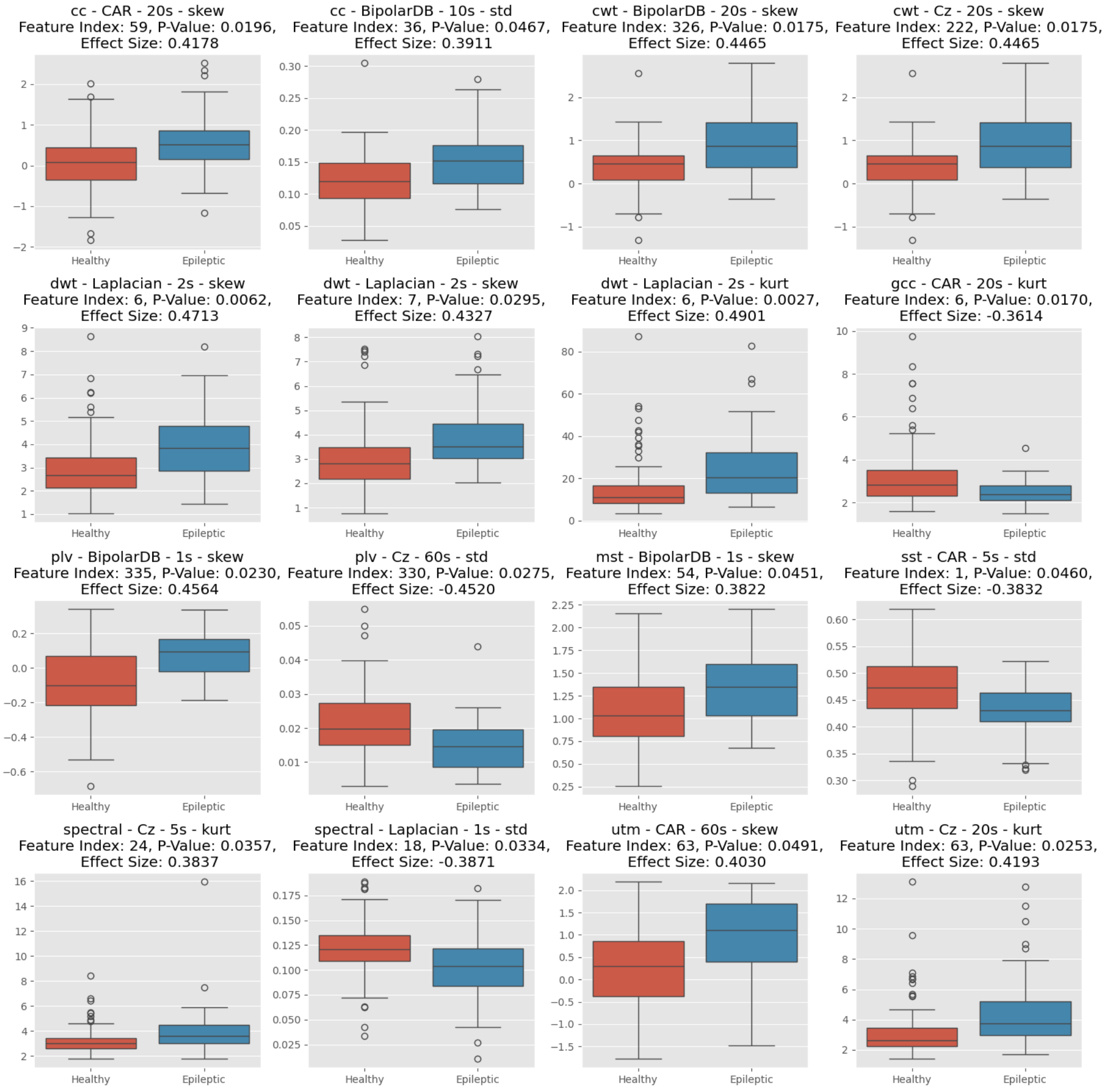

Figure A.3: EMC statistical analysis: box-plots for all significant features (see Table A.2).

B TUH Dataset - Supplementary Results

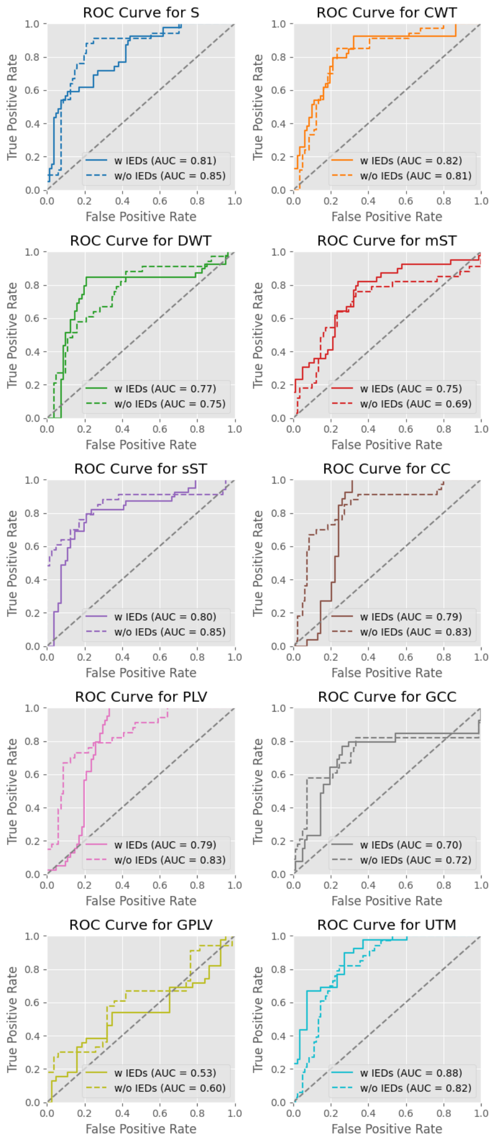

Figure B.1: ROC comparison of feature sets with IEDs vs. without

Table B.1: TUH (all data) extended ensemble metrics: BAC80, F1, Precision, Recall, AUC, AUPRC for combinations of feature sets.

Combination	BAC80	F1-score	Precision	Recall	AUC	AUPRC
mst+gcc	76.80 ± 7.50	68.30 ± 5.30	69.60 ± 8.60	67.70 ± 6.40	85.00 ± 3.60	79.70 ± 4.50
cc+utm+gcc	83.50 ± 6.90	74.60 ± 7.60	70.40 ± 12.00	**80.00 ± 4.20**	90.20 ± 4.20	79.70 ± 10.30
cc+plv+sst+utm	**90.20 ± 5.10**	69.40 ± 15.00	87.20 ± 8.60	60.00 ± 19.90	93.60 ± 3.60	84.60 ± 10.80
cwt+plv+sst+utm+gplv	86.90 ± 4.10	72.20 ± 13.40	89.80 ± 13.70	63.10 ± 18.40	92.30 ± 3.30	84.00 ± 8.50
cc+plv+mst+utm+gcc+gplv	87.50 ± 5.30	**73.80 ± 10.00**	81.30 ± 9.80	67.70 ± 10.00	**93.80 ± 2.20**	**87.30 ± 7.60**
cc+cwt+dwt+mst+sst+utm+gcc	89.10 ± 4.00	55.60 ± 8.60	**93.30 ± 9.10**	40.00 ± 8.40	92.40 ± 3.10	87.00 ± 4.40
cc+plv+mst+sst+spectral+utm+gcc+gplv	88.70 ± 3.00	57.50 ± 7.20	83.50 ± 15.10	46.20 ± 12.20	93.40 ± 2.00	84.80 ± 5.90
cc+dwt+plv+mst+sst+spectral+utm+gcc+gplv	84.60 ± 2.40	58.80 ± 17.30	92.70 ± 10.10	44.60 ± 19.20	90.70 ± 1.20	79.40 ± 7.40
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	82.70 ± 8.40	54.80 ± 18.00	96.70 ± 7.50	40.00 ± 17.50	88.90 ± 6.60	80.60 ± 11.90

Table B.2: TUH (IED-free) extended ensemble metrics: same metrics as Table B.1 on the IED-free cohort.

Combination	BAC80	F1-score	Precision	Recall	AUC	AUPRC
plv+sst	83.10 ± 5.10	73.70 ± 8.20	69.30 ± 12.00	**80.00 ± 7.60**	88.50 ± 1.50	78.50 ± 5.90
mst+sst+spectral	84.00 ± 5.40	70.00 ± 9.10	84.90 ± 10.50	60.00 ± 10.40	89.70 ± 4.80	79.50 ± 8.10
plv+sst+spectral+utm	84.00 ± 6.00	73.40 ± 5.80	81.80 ± 12.60	69.10 ± 13.80	90.10 ± 5.60	82.10 ± 7.50
cwt+plv+sst+spectral+utm	87.20 ± 5.70	**80.80 ± 9.40**	96.00 ± 8.90	70.90 ± 13.50	92.90 ± 4.10	86.60 ± 6.40
plv+sst+spectral+utm+gcc+gplv	83.80 ± 8.00	76.40 ± 5.50	92.80 ± 6.60	65.50 ± 7.60	91.40 ± 2.50	82.90 ± 4.40
cc+cwt+plv+mst+sst+spectral+utm	86.80 ± 2.40	70.00 ± 11.00	82.00 ± 14.60	61.80 ± 11.90	**94.10 ± 3.10**	**87.60 ± 6.60**
cc+cwt+plv+mst+sst+spectral+gcc+gplv	**89.70 ± 3.50**	72.30 ± 6.80	90.50 ± 14.70	61.80 ± 10.00	92.50 ± 2.20	86.00 ± 1.40
cc+cwt+dwt+plv+sst+spectral+utm+gcc+gplv	89.40 ± 2.90	60.40 ± 13.20	85.80 ± 9.10	47.30 ± 13.50	91.60 ± 3.40	80.70 ± 8.50
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	85.10 ± 4.40	63.80 ± 15.20	**96.70 ± 7.50**	49.10 ± 16.50	89.80 ± 3.70	79.70 ± 7.20

C

EMC Dataset - Supplementary Results

Table C.1: EMC extended ensemble metrics: BAC80, F1, Precision, Recall, AUC, AUPRC across feature-set combinations.

Combination	BAC80	F1-score	Precision	Recall	AUC	AUPRC
cc+dwt	66.30 ± 5.90	55.20 ± 4.50	50.80 ± 11.10	62.50 ± 4.30	75.00 ± 3.50	59.10 ± 7.10
cc+mst+gplv	**73.90 ± 6.80**	63.20 ± 5.50	58.90 ± 11.30	**70.50 ± 7.40**	**79.40 ± 5.90**	**65.60 ± 11.80**
dwt+plv+gcc+gplv	70.00 ± 11.60	**64.50 ± 5.30**	**61.10 ± 7.60**	69.00 ± 5.50	77.80 ± 5.50	58.00 ± 5.70
cc+plv+utm+gcc+gplv	69.70 ± 6.60	57.50 ± 3.20	53.50 ± 4.90	63.00 ± 6.90	76.90 ± 4.00	59.90 ± 8.10
cc+dwt+plv+sst+gcc+gplv	71.60 ± 3.70	58.70 ± 6.60	53.20 ± 5.80	65.50 ± 8.00	78.90 ± 5.00	61.70 ± 8.60
cc+cwt+plv+sst+spectral+gcc+gplv	71.50 ± 7.50	56.60 ± 7.90	52.80 ± 4.30	62.50 ± 14.90	77.10 ± 5.10	61.00 ± 9.20
cc+cwt+dwt+plv+sst+spectral+gcc+gplv	69.80 ± 4.30	54.20 ± 5.80	49.70 ± 4.80	60.50 ± 10.10	76.00 ± 3.00	59.90 ± 6.80
cc+cwt+dwt+plv+mst+sst+spectral+utm+gplv	67.20 ± 5.70	58.90 ± 5.00	57.20 ± 5.20	61.00 ± 7.20	75.40 ± 2.40	56.00 ± 4.40
cc+cwt+dwt+plv+mst+sst+spectral+utm+gcc+gplv	67.20 ± 6.60	55.30 ± 7.20	52.70 ± 7.70	59.00 ± 9.80	72.10 ± 6.60	49.20 ± 10.70