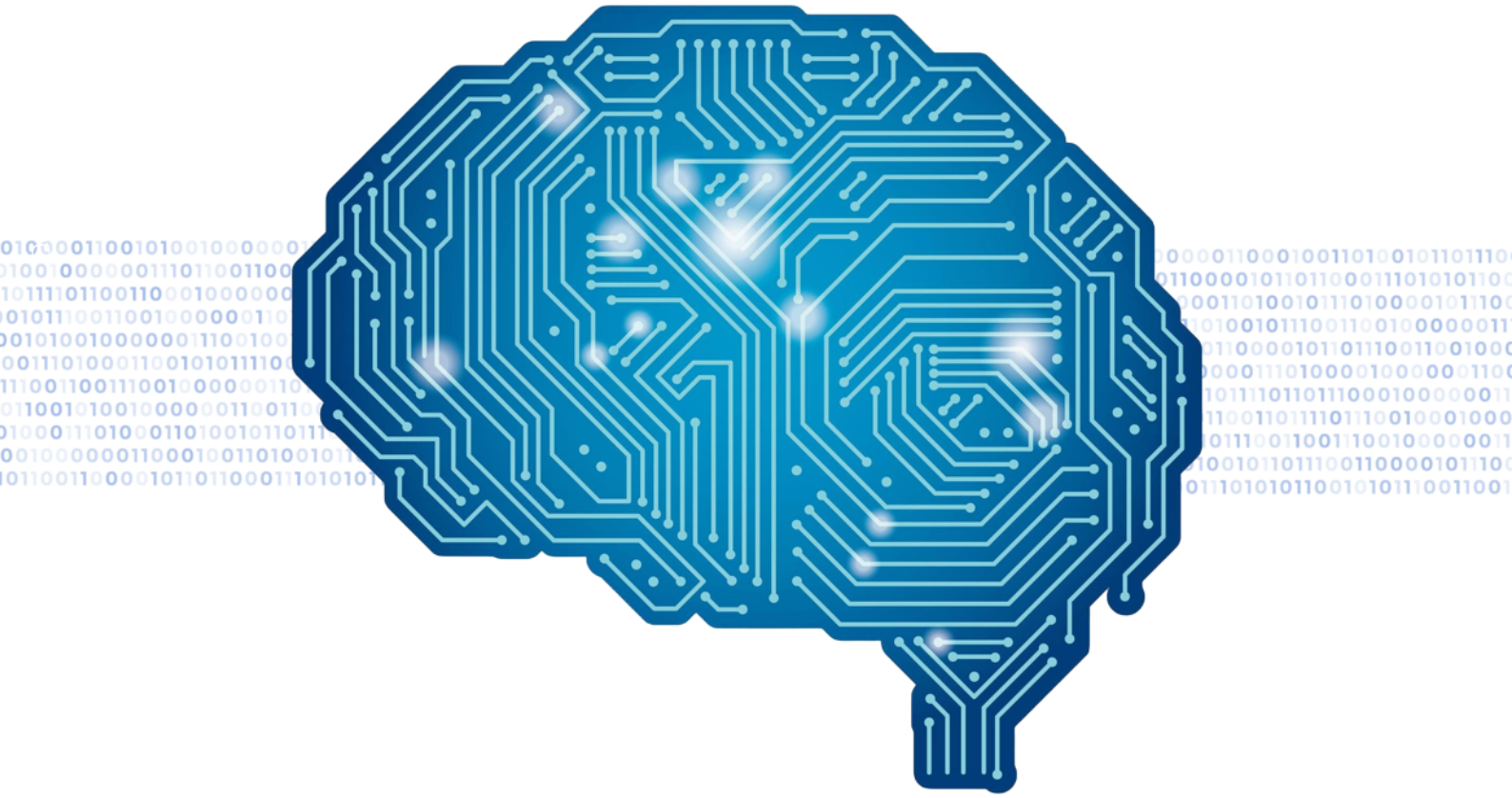
Using machine learning models trained on IED-free EEGs to support epilepsy diagnosis

TM30004: MSc Thesis

P.A. (Paul) van der Kleij



TU Delft Delft University of Technology

Erasmus MC Universitair Medisch Centrum Rotterdam

[This page was intentionally left blank]

Using machine learning models trained on IED-free EEGs to support epilepsy diagnosis

by

P.A. (Paul) van der Kleij

Student number: 4671015
Date: January 20th 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Delft University of Technology; Erasmus University Rotterdam; Leiden University

Master thesis project (TM30004; 35 ECTS)
Dept. of Biomechanical Engineering, TU Delft
March 5th 2024 - January 20th 2025

Supervisors:

Medical Supervision:	Dr. R. van den Berg	Neurologist, Erasmus Medical Center, Rotterdam
Technical Supervision:	Dr.ir. J.H.G. Dauwels	Associate Professor, Signal Processing Systems, TU Delft, Delft

Thesis committee members:

Dr. R. van den Berg	Erasmus MC
Dr.ir. J.H.G. Dauwels	TU Delft
Dr. R. Helling	Stichting Epilepsie Instellingen Nederland

Report Style: TU Delft Report, as modified by Daan Zwaneveld
An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

Erasmus MC Universitair Medisch Centrum Rotterdam

Nomenclature

Abbreviations

Abbreviation	Definition
ASM	Anti Seizure Medication
ATC(DDD)	Anatomical Therapeutic Chemical classification system (with Defined Daily Dose)
AUC	Area Under the (ROC-) Curve
CC	Cross-Correlation
CWT	Continuous Wavelet Transform
DBC	Diagnosis Treatment Combination
DWT	Discrete Wavelet Transform
EEG	ElectroEncephaloGram
EPD	Electronic Health Record
ER	Emergency Room
FPR	False Positive Rate
GCC	Graph measures of Cross-Correlation
GPLV	Graph measures of Phase Lock Values
IED	Interictal Epileptiform Discharge
ILAE	International League Against Epilepsy
LOSO (CV)	Leave-One-Subject-Out Cross Validation
mST	mean Stockwell Transform
PLV	Phase Lock Values
ROC	Receiver Operating Characteristic
S	Spectral
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
sST	square root Stockwell Transform
TC	Tonic-Clonic seizure
TPR	True Positive Rate
UTM	Univariate Temporal Measures
XGB(oost)	EXtreme Gradient Boosting

Abstract

Introduction and Research Goal: Epilepsy is a common neurological disorder that severely impacts patients' quality of life. Current diagnostic standards rely on the presence of seizures or interictal epileptiform discharges (IEDs) in the electroencephalogram (EEG). However, some patients who are ultimately diagnosed with epilepsy do not present with seizures or IEDs on their initial EEG, which delays their diagnosis and appropriate medical treatment. Previous studies by Thangavel et al. 2022 and Mirwani 2024 suggest the potential of machine learning methods applied to IED-free EEGs for classifying epilepsy. The aim of this study is to evaluate whether adding clinical characteristics and visual EEG interpretation to existing machine learning models based on quantitative EEG improves the performance of these models for the identification of epilepsy in IED-free EEGs. Additionally, the study explores model interpretability to promote clinical application.

Methods: We focus on subjects who presented at the emergency room following a first clinical seizure and were not diagnosed with epilepsy based on their initial EEG. Ten quantitative EEG feature sets based various mathematical transforms were readily available from previous research by Mirwani 2024. EXtreme Gradient Boosting (XGBoost) models were trained using Leave-One-Subject-Out Cross-Validation (LOSO CV) on these feature sets as benchmark models. Model performance was assessed using the Area Under the Curve (AUC). Clinical and EEG report features were added individually to each quantitative EEG feature set to evaluate their added value to model performance. The best performing clinical and report features for each quantitative EEG feature set were determined using a two-fold grid search, and significance was tested via Welch's t-test. Ensembles were created from the best performing models of each quantitative EEG feature set including their best performing clinical and report features. SHapley Additive exPlanations (SHAP) and XGBoost feature importance were used for model interpretation, while Bayesian statistics were applied to gain insight into clinical implementation.

Results: The best performing clinical and report features varied across EEG feature sets and did not consistently yield significant performance gains. Only the addition of EEG background to graph metrics of phase lock values showed a significant increase in model performance. SHAP analysis identified residual focal sharp activity as the primary contributor to this improvement. Combining individual models into ensembles substantially improved performance, achieving AUCs up to 0.870. To align model performance interpretation to the guidelines of the ILAE, sensitivity at $P(posterior) \geq 0.6$ was proposed as a key evaluation metric. The best XGBoost model with clinical and report features achieved 0.48 sensitivity at $P(posterior) = 0.6$, while the best ensemble attained 0.81.

Conclusion: Incorporating clinical and report features into XGBoost models based on quantitative EEG data does not consistently improve the detection of epilepsy based on IED-free EEGs. The variability of the best performing clinical and report features across quantitative EEG feature sets suggests that their impact is dependent on the quantitative EEG feature sets. Combining individual models into ensembles significantly enhances performance, achieving a sensitivity of 0.81 at $P(posterior) \geq 0.6$. However, external validation is required to confirm these findings. Future models should assess model performance according to the sensitivity at $P(posterior) \geq 0.6$ to comply with ILAE guidelines.

Contents

1 Introduction

1.1. Problem statement

Epilepsy affects approximately 50 million people worldwide and significantly reduces their quality of life.[1, 2] Activities that most individuals consider routine, such as driving, become unfeasible due to the constant threat of seizures. The International League Against Epilepsy (ILAE) has established a clinical definition of epilepsy, classifying an individual as epileptic if they have experienced two unprovoked seizures more than 24 hours apart or have a greater than 60% risk of seizure recurrence.[3] Diagnoses that fall within the first criterion are relatively straightforward; however, accurately assessing the risk of seizure recurrence introduces a considerable challenge, with many factors to take into account.[4, 5]

A key initial step in assessing seizure recurrence risk is performing an electroencephalogram (EEG), which records aggregate neuronal activity via electrode leads placed on the scalp.[6–8] If the EEG detects (sub-clinical) seizures, the patient can be classified as epileptic under the first criterion. Similarly, if Interictal Epileptiform Discharges (IEDs) are observed, the patient may be classified as having a greater than 60% risk of seizure recurrence, as these IEDs often indicate a potential epileptic hotspot in the cortex.[9] However, IEDs are not present in the EEGs of all epileptic patients, and their interpretation can be difficult.[10] Consequently, a substantial subset of patients who are later confirmed as epileptic through clinical follow-up cannot initially receive a diagnosis based on the EEG alone.[7]

To enhance the likelihood of detecting epileptic activity, patients may undergo sleep deprivation prior to a second EEG. Sleep deprivation reduces the brain's inhibitory mechanisms, thereby increasing the likelihood of observing IEDs or inducing seizure activity.[11] While some patients are diagnosed with epilepsy following this second EEG, a significant proportion of epileptic cases remain undiagnosed.[12]

Currently, the standard approach for patients who have experienced an unprovoked seizure but were not diagnosed with epilepsy after a standard and a sleep-deprived EEG is a wait-and-see policy.[13] Because even if the first diagnostic tests did not confirm an epilepsy diagnosis, epilepsy cannot be definitively excluded based on these tests. This creates substantial uncertainty for patients.[14] Especially since the percentage of individuals that experience a recurrent seizure within a few years fluctuates between 10-50 percent in literature, and strongly depends on in- and exclusion criteria of the population.[5, 15, 16] Expanding initial diagnostic capabilities could enable a larger proportion of epileptic patients to receive appropriate treatment, reducing the duration and uncertainty of the diagnostic process for epilepsy.

1.2. Previous research and research goal

This research builds upon the work of Thangavel et al. 2022, who demonstrated that there might be a role for EEGs that are free of IEDs in epilepsy diagnosis.[17]

Y. Mirwani's MSc thesis extended Thangavel's results by reproducing them on a dataset of subjects from the Erasmus Medical Centre, Rotterdam.[18] In his work, he identified the most suitable hyperparameters for each type of EEG feature set. His methodology involved segmenting specific EEG montages into epochs and calculating features for each epoch individually. The results at the epoch level were then combined using various combination techniques, including mean, median, and standard deviation calculations, to derive the final feature representation for the complete EEG dataset. The present study utilizes the EEG feature sets with the highest Area Under the Curve (AUC), as determined from the Leave-One-Subject-Out (LOSO) cross-validation.

This study aims to assess the added diagnostic value of incorporating clinical and EEG report features into these machine learning models for identifying epilepsy from IED-free EEGs. Additionally, it seeks to investigate how machine learning model predictions should be interpreted to promote applicability in clinical practice.

1.3. Thesis outline

This thesis begins by verifying the findings of prior studies conducted by Thangavel and Mirwani to ensure the robustness and reliability of their results for use in subsequent research. This is achieved by reproducing the ROC (Receiver-Operating-Characteristic) curves of single EEG feature set XGBoost models using LOSO cross-validation.

Next, the development of additional clinical and report-based features derived from EEG reports and electronic health records is discussed in detail. These features are then incorporated into the XGBoost models, and the results of their inclusion is evaluated using LOSO cross-validation. Throughout this thesis we will use the following definitions for different types of features:

EEG feature (set(s)) : Quantitative EEG-level feature set(s) as calculated in the previous research of Thangavel et al 2022 and Mirwani 2024. These feature sets were derived using mathematical transforms on EEG data, and are outlined in Chapter 4.1.

(EEG) **Report feature**(s) : Features that were derived from the EEG reports. These EEG reports were the result of the visual analysis of EEGs by neurologists. We use two types of report features in this research; individual EEG rhythms, and (overall) EEG background.

Clinical feature(s) : Features extracted from clinical or demographical data of the subjects. These include; age, sex, vigilance state, medical history, and medication use/history.

The results of the LOSO cross-validation are further evaluated using Bayesian statistics to gain more insight into model performance dynamics. Specifically, to integrate current clinical guidelines from the ILAE for epilepsy diagnosis within model performance metrics.

2
Background

2.1. Epilepsy

Epilepsy affects approximately 1% of the global population.[2] Individuals with epilepsy experience seizures that vary widely in recurrence rates and clinical presentations.[19, 20] The most well-known type is the Tonic-Clonic (TC) seizure, characterized by an initial phase of generalized muscle contractions followed by rhythmic muscle jerks, almost invariably accompanied by a loss of consciousness.[21] However, epilepsy can also manifest as focal seizures, where clinical presentations are highly variable depending on the affected brain region. These may include behavioral arrests, localized muscle spasms, auras, perseverations, and other symptoms. The underlying causes of epilepsy are diverse and include genetic, traumatic, immunological, structural, and other etiologies.[22, 23]

The International League Against Epilepsy (ILAE) has established a widely accepted clinical definition of epilepsy, designed to encompass all types of epilepsy regardless of seizure patterns or pathophysiology.[3] According to this definition, epilepsy is diagnosed in individuals who have experienced two unprovoked seizures more than 24 hours apart or in individuals who have had one unprovoked seizure with a recurrence risk of over 60%. The assessment of this recurrence risk is typically conducted by the treating neurologist but can be a challenging determination.[13]

Historically, the pathophysiology of epilepsy was attributed to focal brain damage leading to localized disruptions in brain function, which in turn caused seizures.[24] While focal epileptic hotspots remain a recognized potential cause, growing evidence supports the hypothesis that epilepsy is a network disorder characterized by increased excitability of the brain as a whole.[25, 26] Given the significant variability in causes, clinical presentations, and seizure frequencies among individuals, it is plausible that both pathophysiological mechanisms—focal epileptic hotspots and network dysfunction—may work in synthesis.[27]

Patients diagnosed with epilepsy are typically treated with Anti-Seizure Medication (ASM).[28] Even for patients who are promptly diagnosed, identifying the most effective ASM can be a lengthy and complex process.[29] With over 30 available ASMs, treatment effects and potential side effects vary significantly between individuals, making the selection of the most appropriate ASM a challenging endeavor.[30]

2.2. The electroencephalogram

To evaluate whether an individual has epilepsy or to gather information about the potential focal origin of seizures, neurologists commonly use the Electroencephalogram (EEG).[6] The EEG is a neuroimaging technique in which electrodes are attached to the scalp according to a standardized topology known as the 10-20 system. This system measures the distance from the inion (the bony prominence at the base of the skull) to the nasion (the bridge of the nose) and divides the vertical segment into intervals of 10%, 20%, 20%, 20%, 20%, and 10%. Similarly, horizontal measurements are taken from ear to ear and divided in the same proportions. Using these landmarks, a grid is created that accommodates 19 electrodes, each with a precise spatial location on the scalp.[31] Figure 2.1 Additionally, two electrodes—referred to as the 'ground' and 'referential' electrodes—are included to improve signal quality and facilitate the display of signals.[32] The electrodes are fixed to the scalp using an electrically conductive paste.[33, 34]

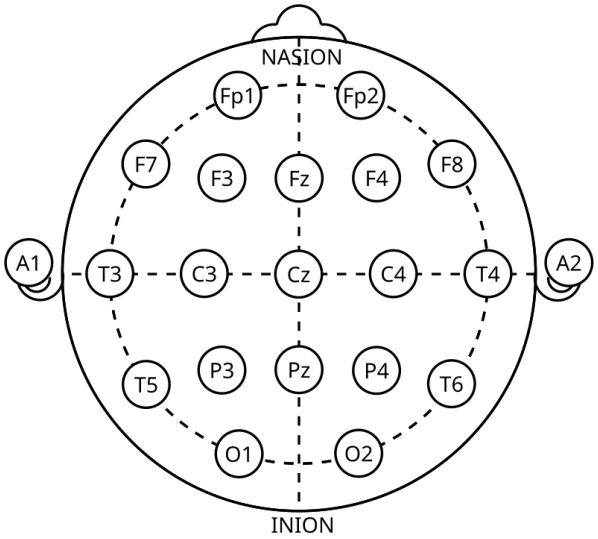


Figure 2.1: The 10-20 system for attaching electrodes to a subjects' scalp when performing an electroencephalogram

Neurons directly beneath each electrode generate small electrical pulses during axonal signal transmission. When numerous neurons follow similar signaling pathways, particularly along major brain signaling routes, their aggregate activity creates an electric dipole in the microvolt range.[35] This current is detected by the electrode positioned above. The amplitude of the registered signal depends on multiple factors; the proximity of the electrode to the source of neuronal activity, the composition of intermediate tissue, and the orientation of the electric dipole. By using multiple electrodes, the EEG can approximate the source of the recorded signals.[36, 37] However, this also highlights a key limitation of the EEG—it predominantly captures surface-level brain activity, as the electrodes are positioned closest to the outer cortical structures. Deeper brain regions, like the basal ganglia, thalamus, and brainstem, are not assessed directly with EEG due to their distance from scalp electrodes; instead, an EEG mostly reflects their impact on cortical rhythms.[38, 39]

Beyond documenting seizure events, EEGs are also capable of detecting interictal epileptiform discharges (IEDs), which occur in the intervals between seizures. These IEDs, characterized by distinct patterns such as spikes or spike-and-wave complexes, are clinically significant biomarkers.[40] Their presence is often associated with an increased risk of seizure recurrence and serves as an important observation to diagnose epilepsy.[10] Figure 2.2 shows an example of an EEG with spike-and-wave complexes. Not all patients with clinically confirmed epilepsy, as determined through follow-up, have observable seizures or IEDs in their initial EEG recordings.[41] This study focuses on this specific subset of patients, predicting their seizure recurrence risk through analysis of EEG with supervised machine learning techniques.

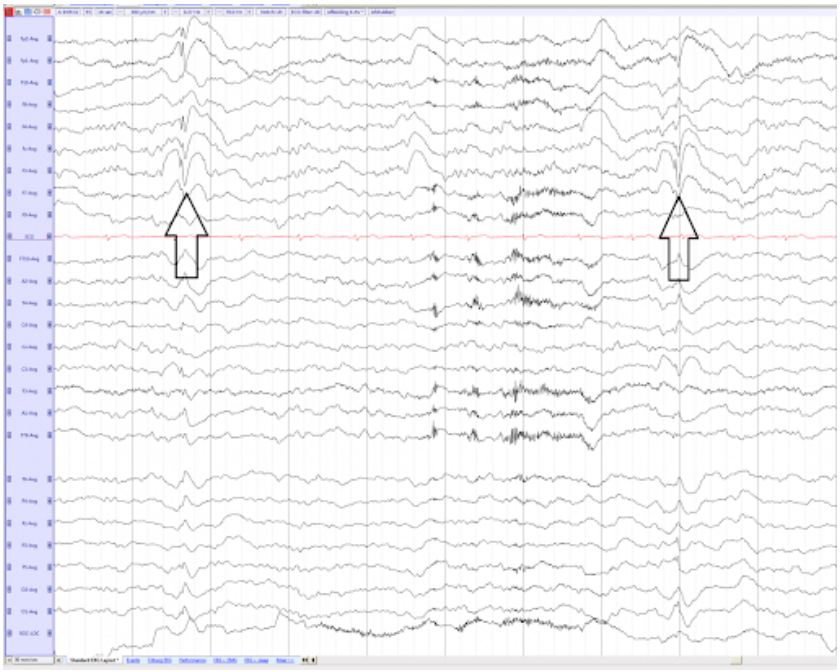


Figure 2.2: An electroencephalogram showing two instances of spike-and-wave complexes as indicated by the arrows.[42]

2.3. Machine learning and data encoding

Supervised machine learning is a method for building predictive models that learn from labeled datasets.[43] The core concept involves training a model on a dataset that includes:

Features: Independent variables or predictors that serve as inputs to the model. These can include numeric values (e.g., age, income) or categorical attributes (e.g., gender, occupation).

Labels: Dependent variables or targets that the model is designed to predict. For instance, in classification tasks, labels may represent categories such as “spam” or “not spam.”

The model’s goal is to minimize the difference between its predictions and the true labels. Metrics such as accuracy, precision, recall, or mean squared error are used to assess the model’s ability to generalize to new data. In many classification tasks, models output probabilities for each category, reflecting their confidence and enabling threshold-based decision-making.[44]

Through iterative parameter adjustments, supervised learning models learn a mapping from features to labels, striving to maximize prediction accuracy on unseen data.

Feature representation, especially for categorical data, is a critical aspect of supervised learning. While numeric data is often directly interpretable by machine learning algorithms, categorical data usually requires encoding to be effectively utilized.[45]

- **Ordinal:** Categories with a natural order (e.g., “low,” “medium,” “high”).
- **Non-Ordinal:** Categories without an intrinsic order (e.g., “red,” “blue,” “green”).
- **Single-Class and Multi-Class:** Features may represent one category per instance (e.g., “type of fruit”) or multiple categories simultaneously (e.g., “skills possessed”).

This research emphasizes the importance of encoding multi-class features, with one-hot encoding being a common method.[46] The process involves:

Assigning each unique category to a separate binary column. For each instance, placing a “1” in the column corresponding to the observed category and “0” in all other columns.

For example, consider a feature with four categories (A, B, C, D) across four subjects. The left table shows the original feature categories that are present in the subjects, the one-hot encoded feature is

shown on the right:

Subject	Feature
1	A, C
2	B
3	B, C
4	D

$\longrightarrow$

Subject	A	B	C	D
1	1	0	1	0
2	0	1	0	0
3	0	1	1	0
4	0	0	0	1

Each row represents a subject, and each column indicates the presence (1) or absence (0) of a category.

Accurate representation of multi-class categorical features is essential for model performance in this research. One-hot encoding effectively translates categorical data into a machine-readable numerical format. This ensures categorical distinctions are preserved and avoids introducing unintended ordinal relationships among categories.[47] This preprocessed data can then be used to train a machine learning model, in our case using an Extreme Gradient Boosting (XGBoost) algorithm.

2.4. Extreme Gradient Boosting

XGBoost is a powerful ensemble learning method based on decision trees. Understanding its mechanism requires understanding the core components of decision trees and how XGBoost uses them for advanced predictions.[48]

A decision is a model that makes predictions by sequentially splitting data based on specific features.[49] The key components of a decision tree include:

- **Node:** A decision point in the tree where a feature is evaluated to split the data.
- **Branch:** The outcome of a decision at a node, which leads to another node or a final prediction.
- **Leaf:** A terminal node where no further splitting occurs, representing the final group or prediction.

Each split within a decision tree aims to better classify subjects. However, a single decision tree often lacks the complexity needed to capture intricate patterns in data.

To overcome the limitations of a single decision tree, XGBoost constructs an ensemble of multiple decision trees. The final prediction is determined by aggregating the contributions of all trees in the ensemble. To enhance accuracy, XGBoost employs a process called gradient boosting, which iteratively minimizes the error of each subsequent tree using residuals; these are the differences between the predicted values of a tree and the actual known labels. They measure the error for a specific tree. By aiming to minimize the residuals, trees are made more accurate. This optimization process ensures that each new tree corrects the errors of its predecessors, gradually improving the ensemble's predictive performance.[50]

A more classical approach to ensemble learning is the Random Forest (RF) classifier. The main difference between RF and XGBoost lies in how the trees are built: RF constructs decision trees in parallel, with each tree trained independently on a random subset of the data and features. This independence promotes diversity among the trees, and their predictions are combined through averaging or voting to produce the final result. In contrast, XGBoost builds trees sequentially, where each tree learns from the errors of the previous ones by optimizing a gradient-based loss function. This sequential approach allows XGBoost to focus on harder-to-predict samples, often resulting in higher predictive accuracy, but at the cost of increased complexity compared to the more straightforward RF method.[51]

2.5. Model validation

The Area Under the Curve (AUC) is a performance metric used to evaluate the quality of a machine learning model, particularly for binary classifiers. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. An AUC value ranges between 0 and 1, where 1 indicates perfect classification, 0.5 reflects no discriminative ability (equivalent to random guessing), and values below

0.5 suggest performance worse than random.[52] Figure 2.3 illustrates how ROC curves visualize the performance of a model by showing the trade-off between TPR and FPR across various thresholds.

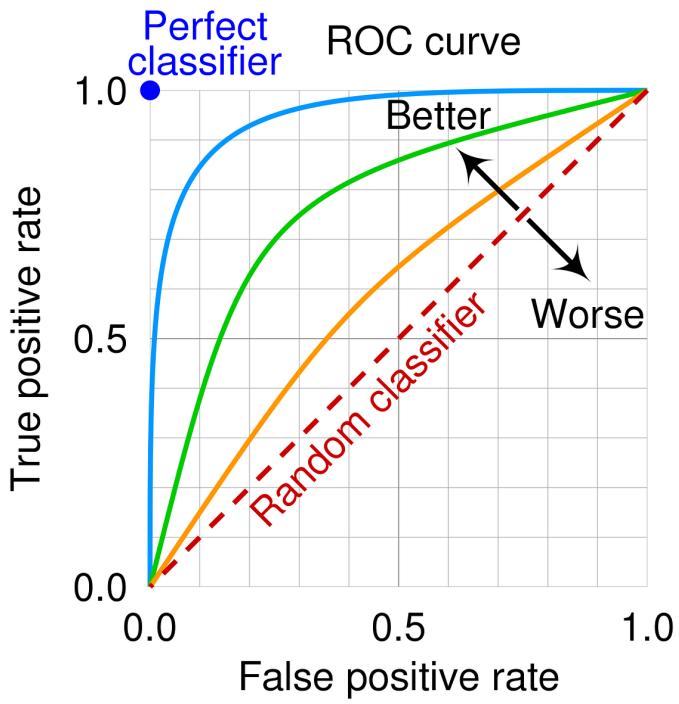


Figure 2.3: A few examples of ROC-curves with varying levels of model performance.
[ROC = Receiver-Operating Characteristic]

Leave-One-Subject-Out (LOSO) cross-validation uses data from all subjects except one as training data, and the remaining subject's data is used for testing. This process is repeated for each subject, ensuring that every individual contributes to both testing exactly once. LOSO is particularly effective in smaller datasets, because it leverages as much training data as possible, while still being able to cross-validate results.[53] Figure 2.4 shows that for each iteration a single subject functions as the test set, and the rest of the data can be used to train the model.

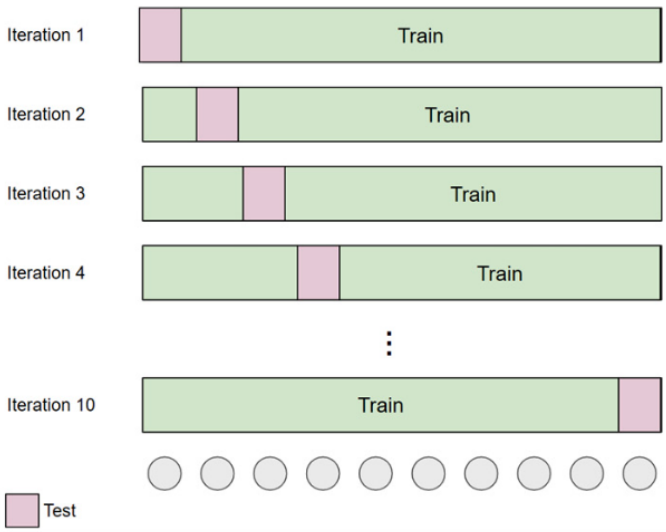


Figure 2.4: A schematic representation of Leave-one(-subject)-out cross validation for 10 subjects.

3
Data acquisition

3.1. Dataset

3.1.1. Subject population and non-WMO approval

In total, 143 people were retrospectively included in this research of which 104 were labeled as 'healthy' and 39 were labeled as 'epileptic', having had a recurrent seizure. All subjects were adults and seen in the emergency room (ER) after having had a first clinical seizure. As per protocol, after initial neurological assessment, the dpt. of clinical neurophysiology performed an EEG to estimate the recurrence risk of seizure. If this first EEG is inconclusive, a second EEG after sleep deprivation can be suggested, in order to provoke epileptiform activity in the brain. Both the first and second EEG (after sleep deprivation) were deemed inconclusive for all subjects in this research. The eventual labeling of subjects resulted from at least 1 year of clinical follow-up. If patients were reported to have had a recurrent seizure they were labeled 'Epileptic', if they remained seizure-free up to the moment of assessment in this research, they were labeled 'Healthy'. If a healthy subject's patient file was checked at a later time, for clarifying ambiguity elsewhere, and an instance of recurrent seizure was seen, the label of the patient was changed to 'epileptic'.

Compared to the previously reported dataset in the MSc thesis of Y. Mirwani, there have been 4 subject exclusions because epilepsy had already been diagnosed with epilepsy no more than 10 years prior.[3, 18] 1 subject's label was changed from 'healthy' to 'epileptic', because of a recurrent seizure.

The 'Medisch Ethische Toetsings Commissie' (METC) has approved this research as 'niet-WMO', under case number MEC-2021-0145.

3.1.2. Data sources and data safety

This data used in this research can be separated in three groups; EEG registration data, EEG report data, and clinical data.

The EEG registration data of subjects was exclusively sourced from the first EEG, and not from the second EEG after sleep deprivation. The EEG data was readily available from the EEG-archive of the dpt. of clinical neurophysiology at the Erasmus Medical Center, Rotterdam.

The EEG reports are a combination of initial patient information reported by neurologist on the ER, and the interpretation of the EEG by a neurologist after EEG registration. They contained information for both report and clinical features.

Clinical data from subjects' Electronic Patient Record (EPD) was obtained after a data issuance at the Erasmus MC Data Center.

All data was used after subject pseudonymization on a secure research server within the 'my Digital Research Environment (anDREa B.V. 2021).[54]

4

EEG feature sets from previous research

4.1. EEG feature sets in previous research

We did not start with raw EEG signal data in this thesis. Previous research by Thangavel et al. 2022 and Y. Mirwani, described, calculated and verified various types of EEG feature sets from the original EEG signals.[17, 18] The EEG feature sets were readily available for this research. The models trained on EEG feature sets act as the benchmark, whereupon the clinical and report features will be added to evaluate their (additional) predictive value. In total, 10 EEG feature sets were described in the preceding research.[18] Each of the EEG feature sets stems from a different method of signal analysis:

- **(Maximum normalized) Cross-Correlation**: This feature quantifies the level of similarity between different brain regions as a function of the time-lag that is applied to one of the input signals.
- **Continuous Wavelet Transform**: A time-frequency analysis method. Using the Morlet mother wave, it captures transient and non-stationary signal characteristics, providing information on the temporal and spectral domain.
- **Discrete Wavelet Transform**: Another wavelet-based method, which decomposes EEG signals into different frequency bands. It shifts the Daubechies mother wavelet in discrete steps to offer a computationally efficient representation temporal and spectral domain characteristics.
- **Graph Measures of Cross-Correlation**: This feature set uses graph theory to model the connectivity patterns from Cross-Correlation. They analyze properties such as clustering and centrality to interpret brain network dynamics.
- **Graph Measures of Phase Lock Values**: This feature set applies graph theory to Phase Lock Values, emphasizing phase synchrony-based network structures.
- **Mean Stockwell Transform**: The Stockwell Transform combines frequency properties of the Short-time Fourier Transform with the multiresolution capabilities of the Continuous Wavelet Transform. It is a measure of time-frequency energy distribution in EEG signals.
- **Phase Lock Values**: A measure of phase synchronization between EEG signals at different locations, often used to study brain coherence and connectivity.
- **Square Root Stockwell Transform** A variation of the Stockwell Transform that focuses on signal amplitude in time-frequency space.
- **Spectral Features**: These include power spectrum and frequency-domain features such as band power in five standard frequency ranges (i.e. delta, theta, alpha, beta, and gamma bands).
- **Univariate Temporal Measures**: Time-domain features derived from the raw EEG signals, such as amplitude, variance, and Shannon entropy.

Keeping these EEG featuresets separate when training the machine learning models gives insight in their individual performance, and their relation to the clinical and report features. In total 10 different types of EEG feature sets were evaluated, of which the abbreviations are shown in Table 4.1.

The features within each set were not calculated directly over the entire EEG. First all electrodes from the EEG were referenced according to a standardized montage. The EEG was split into epochs, where

Table 4.1: Overview of EEG feature sets from previous research and their abbreviations.

EEG feature set	Abbreviation
Cross-Correlation	CC
Continuous Wavelet Transform	CWT
Discrete Wavelet Transform	DWT
Graph measures of CC	GCC
Graph measures of PLV	GPLV
Mean Stockwell Transform	mST
Phase Lock Values	PLV
Square root Stockwell Transform	sST
Spectral	S
Univariate Temporal Measures	UTM

EEG = ElectroEncephaloGram

features were initially calculated. The resulting features from each epoch were then combined using a statistical combiner, to calculate the final features, representing the entire EEG. For all of these hyperparameters - montages, epoch lengths and statistical combiners - there were multiple options, as shown in Table 4.2. Calculations for all possible combinations of EEG feature set, montage, combiner, and segment length were already performed by Mirwani 2024.[18]

Table 4.2: Overview of EEG feature set hyperparameter options

EEG featureset hyperparameter	Options
Montage	[Common Average Reference, Laplacian montage, Cz-reference, Bipolar Double Banana]
Combiner	[Mean, Median, Standard Deviation, Skewness, Kurtosis]
Epoch segment length	[2, 5, 10, 20, 50, 120, 300]

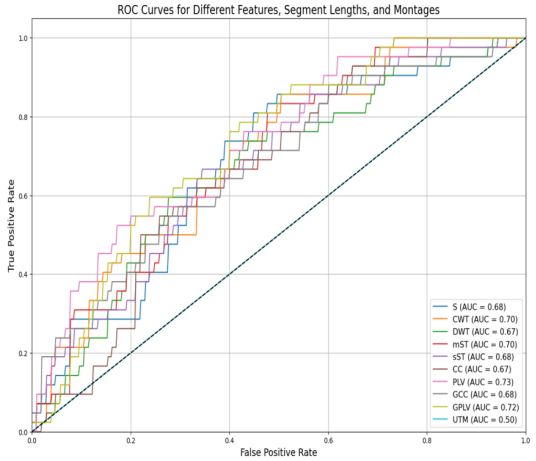
EEG = ElectroEncephaloGram

4.2. Verification on original dataset

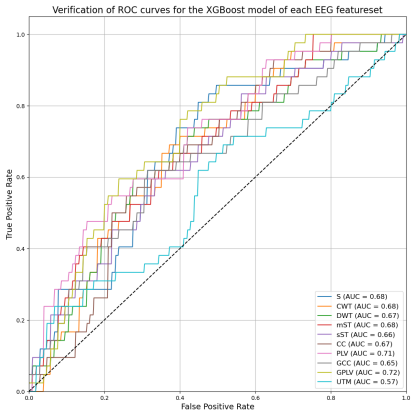
Before we used the EEG featuresets from previous research, their results were verified. In first instance, this was done on the original dataset (without the additional patient exclusions mentioned in Chapter 3). There were two main goals for the verification; 1) to check whether the settings of XGBoost resulted in reproducible results, 2) to check whether the optimal hyperparameters of EEG featuresets (montage, combiner, segment lengths) were still relevant. For the exact hyperparameters per EEG feature set, see Table A.1a in Appendix A. The methodology of training the XGBoost models was similar to previous research, using Leave-One-Subject-Out cross validation (LOSO (CV)).[17, 18]

Figure 4.1a shows the original results from the thesis of Y. Mirwani, Figure 4.1b shows the reproduction of these results. Some irregularities can be observed in the AUCs of the reproduced ROC-curves. In multiple instances, the reproduced ROC-curves yield an AUC that is 0.01-0.02 lower than their original counterparts. These discrepancies can be explained by the fact that these specific EEG featuresets already included either age and/or vigilance state as an encoded feature within the thesis of Y. Mirwani, whereas the verification ROC-curves were solely based on the EEG features.[18]

A second important observation is that the AUC of the UTM feature set equals exactly 0.5 in the original ROC. An AUC of 0.5 corresponds to pure chance, indicating zero predictive power of a model. After inspecting the UTM feature set data, it was found to be notated as imaginary numbers, which was not compatible with the selected method of loading data in Python. The imaginary parts of data were zero in >99% of data points, and negligibly small (<0.01%) compared to their real counterparts in all other data points. Based on this observation, it was concluded that the imaginary part of data did not hold any relevant info. To convert the imaginary numbers to the real domain we took the absolute value of



(a) ROC-curves of EEG featuresets for optimal hyperparameters from Y. Mirwani's thesis.



(b) Reproduced ROC-curves on same dataset and hyperparameters of EEG featuresets

Figure 4.1: Side-by-side comparison of EEG feature set ROC-curves for original subject population and EEG feature set hyperparameters.
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

each data point. The ROC-verification in Figure 4.1b was plotted after fixing the UTM data set. The inconsistent morphology of the UTM ROC-curve compared to the other EEG feature sets indicates that the current hyperparameters for the UTM feature set might not be optimal.

4.3. New EEG benchmark ROC-curves

We decided to do a new grid search for the optimal hyperparameters of EEG feature sets, because of the following; 1) the subject exclusions mentioned in Chapter 3 accounted for almost 10% of the epileptic cases in the original dataset. These subject were excluded because they were already diagnosed with epilepsy at the time the EEG was performed. This could heavily skew feature representations in the epileptic subset. 2) There was a lack of stochastic variables within the former settings of XGBoost, resulting in near-deterministic results 3) The UTM feature set did not show a credible optimum.

A 2-fold grid search was performed for each EEG featureset, trying all combination of hyperparameters shown in Table 4.2. Stochasticity was introduced by setting the XGB settings 'subsample', 'colsample_bytree', 'colsample_bylevel' to 0.9 and using the 'random_state' variable to define the sample splits. These stochastic parameters constrained the XGBoost algorithm to optimize over a subset of the total available data during each iteration of the LOSO CV, promoting variability of used data and reducing potential overfitting. For each EEG feature set, the highest average AUC over the 2-fold search was chosen to be the new benchmark. This benchmark was then bootstrapped 10 times to gain insight in the average AUC and its standard deviation (SD). The resulting new benchmarks for EEG feature sets are shown in Figure 4.2. The hyperparameters for each EEG featureset are shown in Table A.1b in Appendix A

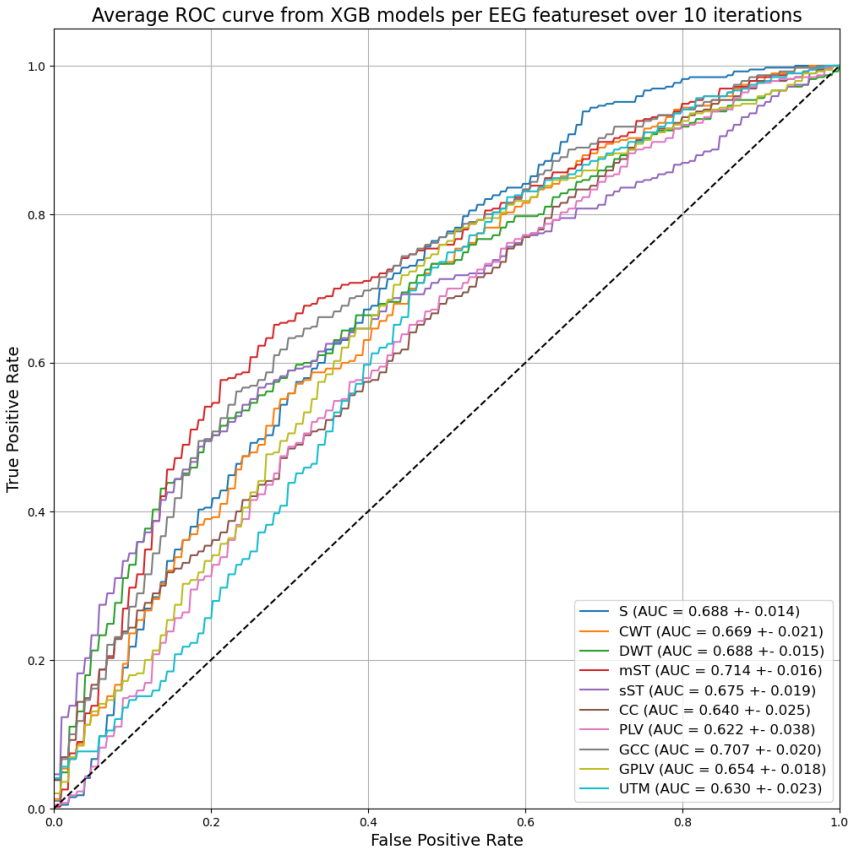


Figure 4.2: ROC-curves for EEG feature sets after patient exclusions and new hyperparameters with performance indicated as AUC +- SD

[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, SD = Standard Deviation, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

5 Development of clinical and report features

One of the primary research objectives is to evaluate the added value of incorporating clinical and report features alongside existing EEG feature sets for training an XGBoost model for the detection of epilepsy. The clinical and report features were developed according to data from the electronic health record (EPD) and EEG reports.

5.1. Clinical features

5.1.1. Age, sex, vigilance

The age of a subject is notated both in the EEG Report and EPD. Age was split into 4 groups: 18-30, 30-50, 50-70 and 70+ years old. There were no pediatric subjects (<18 years old) in the population, as one of the inclusion criteria was being an adult at the time of their first seizure. Epilepsy is known to be bimodally distributed, appearing mostly at young age and in the elderly.[55, 56] Splitting subjects based on 4 age-groups aids anonymity of subject data, but still aims to capture this distrubution. Finally, the age-groups were numerically encoded to be compatible with machine learning techniques.

The sex of subjects was also notated both in the EEG Report and EPD. The sex of a subject was numerically encoded, with 0 corresponding to male, and 1 to female.

The vigilance state during the EEG was extracted from the EEG report. EEG background can significantly change during drowsiness or sleep, compared to the EEG in awake patients.[57] The 'patient state' and 'sleep' data fields describe to which extent a patient presented with drowsiness or (intermittent) sleep during the EEG recording. According to the available information, a practicing neurologist classified subjects to be in one of the following categories; awake, drowsy, (intermittent) sleep. The vigilance state was 1-hot encoded for further use (refer to Chapter 2.3 for an explanation on 1-hot encoding). Since drowsiness precedes sleep, subjects that presented with (intermittent) sleep were simultaneously categorized as being drowsy. An advantage of one-hot encoding is that it allows subjects to be assigned multiple feature categories, a characteristic which has also been used in other features.

5.1.2. Medical history

The medical history of a subject was collected through the EPD and EEG reports. Dutch hospitals use 'Diagnosis Treatment Combinations' (DBCs) to track the diagnoses, treatments, and costs of patients. These DBCs are mainly a means of systematically documenting medical treatment in order to receive reimbursement from healthcare insurance companies.[58] All DBCs of a subject leading up to the date of EEG registration were taken into account for this research. An additional source of subjects' medical history was found within the EEG reports. The neurologist at the ER who filed the inquiry for a first EEG usually includes relevant background information. This background information stems from initial anamnesis at the ER with the subject and/or acquaintances that were present.

These two sources of medical history were first summarized in keywords, to create an overview of all available history for a given subject. Keywords were then grouped based on organ systems or similarity of etiology. The eventual grouping of keywords was discussed with a neurologist, as a means of verification. Data on seizure semiology was not taken into account, because there was a large expected reporting bias, both by subject/bystanders and by ER physicians. The final medical history

groups were 1-hot encoded, of which an overview can be found in Appendix A.2.

5.1.3. Medication

Medication (use and history) of a subject was extracted from the EPD. This included both active medication prescriptions at the time of EEG registration and medication that was administered or prescribed in the past. All medication use had to predate the EEG registration. Medication was grouped based on the Anatomical Therapeutic Chemical Classification System (ATC).[59] This classification system was founded as a tool for drug utilization monitoring and research in order to improve quality of drug use. It classifies medication on different sub-levels, of which the first three will be taken into account for this research. The 1st order ATC-code of medication refers to their anatomical target, the 2nd order describes the therapeutic subgroup and the 3rd order refers to their pharmacological subgroup. The subgroups are notated in an additive manner; the 1st order is represented by a letter, the 2nd order by a (zero-padded) digit, the 3rd order by a letter again. As an example: N05A, describes medication that target the nervous system (1st order 'N'), is a psycholeptic ('05'), and falls within the antipsychotics group ('A').

The medication groups were 1-hot encoded twice, to experiment with the different levels of ATC-codes. There was no correction used for patients that use multiple types of medication that fall within the same 1-hot category. The first encoding method consisted of only the 1st order of medication ATC codes. The second encoding method used the 1st and 2nd order of medication ATC codes, except for the medication group 'N' corresponding to the nervous system. This group was deemed most relevant, since epilepsy is also a nervous system disease. All medication that corresponded to the nervous system was classified based on 1st, 2nd, and 3rd order ATC codes, to gain a more detailed view of this medication. An overview of ATC groups that were included in both versions of 1-hot encoding can be found in Appendix A.3.

5.2. Report features

5.2.1. EEG background and individual rhythms

The report features that were developed aim to describe the qualitative interpretation of an EEG by the neurologist. The EEG background and rhythms were also extracted in two levels of detail, and were both 1-hot encoded. The first describes all reported individual EEG rhythms and electrographical findings. These rhythms and findings were extracted from standardized fields within the EEG reports. An overview of all included individual rhythms and electrographical findings, is reported in Appendix A. The second version of 1-hot encoding used all present rhythms, findings and the conclusion of the neurologist to classify the overall EEG background. The categories for classifying the EEG background were; normal, focal fast, focal slow, focal sharp, diffuse fast, diffuse slow, diffuse sharp, and are outlined in Appendix A. For both versions of the 1-hot encodings, subjects can be assigned multiple categories.

5.3. Clinical and report feature characteristics

Before the XGBoost models were trained, the characteristics of the clinical and report features were assessed to gain insight into the distribution of data. The relative counts of categories within each feature were compared between the epileptic and healthy groups. A 2x2 contingency table was created for each category within a clinical or report feature, as illustrated in Figure 5.1

	Epileptic group	Healthy group
Category present (+)	#A	#B
Category not present (-)	#C	#D

Figure 5.1: A contingency table for a category within a clinical or report feature with counts (#) A, B, C, D to be used in further χ^2-test calculations

The χ^2-test statistic was calculated through the 'SciPy.stats' Python package.[60] Here the χ^2-test statistic is calculated according to Equation (5.1).[61] If any of the frequencies in a contingency table are ≤ 10, the Yates contuity correction will be used to prevent overestimating statistical significance, as shown in Equation (5.2).[62] The p-value corresponding to the χ^2-test statistic can be determined by using a χ^2 lookup table. The 'SciPy.stats' package has this lookup table built-in, which will be used for finding the p-value. A significance level of $\alpha = 0.05$ is used for evaluating the p-values.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.1}$$

$$\chi^2_{Yates} = \sum \frac{(\mid O_{ij} - E_{ij} \mid - 0.5)^2}{E_{ij}} \tag{5.2}$$

where;

χ^2 : The χ^2-test statistic.

O_{ij} : Observed count of cell (i,j) in the contingency table.

E_{ij} : Expected count of cell (i,j), as calculated from marginal totals; $E_{ij} = \frac{Row_i\ Total \cdot Column_j\ Total}{Table\ Total}$

Additionally, the co-occurrence was visualized for medical history, medication, individual EEG rhythms, and EEG background. If we take the individual EEG rhythms as an example, the co-occurrence describes which EEG rhythms are seen in the same patient. If 2 rhythms co-exist in all subjects, they contain the same information, and adding both rhythms as features individually might be redundant. In general terms, the co-occurrence of 2 feature categories is determined by counting the amount of subjects in which both categories are present.

6
XGBoost Methods

6.1. Model training

6.1.1. XGBoost and leave-one-subject-out cross validation

All models trained in this thesis use EXtreme Gradient Boosting (XGBoost), which is explained in more detail in Chapter 2. The settings of XGBoost were largely similar to previous research by Y. Mirwani.[18] The most important change to the XGBoost settings is found within the stochastic variables. The 'subsample', 'colsample_bytree', 'colsample_bylevel' were all set to 0.9, to introduce stochasticity on different levels within the XGBoost model. If there is not enough stochasticity within the presented portion of data to an XGBoost model, it can work in a deterministic way since the algorithm aims to minimize the residual error compared to the previous decision tree. Deterministic models are more likely to hone in on local optimums. Literature shows machine learning models with stochastic variables tend to generalize better than deterministic models.[63, 64]

For all models that were trained, Leave-One-Subject-Out (LOSO) cross validation was employed. This cross validation method trains a model on all but one subject for each iteration. The left-out subject functions as the test set for that iteration. By saving the prediction probabilities of test subjects over all iterations, model performance metrics can be calculated.[53]

6.1.2. Benchmarks

Before adding clinical and report features to models, a benchmark is required to test their performance against. We use XGBoost models from a single EEG feature set as benchmark. In total there are 10 EEG feature sets, with corresponding XGBoost models. Chapter 4 describes how the optimum hyper-parameters of EEG feature sets were found. By keeping the EEG feature sets separate in this stage, we can evaluate whether there is an association between clinical or report features and specific EEG feature sets. It also prevents making the feature space too elaborate and overly complex. The XGBoost models of individual EEG feature sets will be combined at a later stage by taking their ensemble.

6.1.3. Addition of single clinical or report features

When incorporating clinical or report features into XGBoost, the features were defined as 'categorical' to prevent the model from interpreting unjust ordinal relationships. While this concern is mitigated when using one-hot encoding, defining features explicitly as categorical ensures consistency and avoids potential misinterpretation of the data structure by the algorithm. XGBoost also allows determining the maximum number of feature columns that can be viewed as being 1-hot encoded. After some testing, for a reasonable range of the 'max_cat_to_onehot' parameter, the model performance did not change. Since it does increase the dynamic complexity of the XGBoost algorithm, and because we deal with a varying amount of categories within each clinical or report feature, this parameter was not used.

There are 5 types of clinical features that can be added to the EEG feature sets; age, sex, vigilance, medical history, and medication. The medication use was encoded twice, with different levels of detail. Additionally, the report features consisted of 2 variant for EEG interpretation; the overall EEG background, and the individual EEG rhythms. In total, this yields 8 clinical and report features. Each clinical or report feature was appended to the separate EEG feature sets, whereafter XGBoost was trained using LOSO cross validation. The training of XGBoost models was bootstraped 10 times, using a different random state each time, to account for the stochasticity of training models. Hereafter the mean AUC (+- SD) could be calculated for feature combination.

6.1.4. Addition of multiple clinical and report features

From the perspective of XGBoost settings, the addition of multiple clinical and report features was equivalent to adding a single one. The optimal set of clinical and report features was found by performing a two-fold grid search over all possible combinations of clinical and report features for each EEG feature set. The highest mean AUC over this two-fold search, was deemed the optimal set of clinical and report features. The XGBoost model of the resulting combination of EEG feature set and optimal clinical and report features was then bootstrapped 10 times to calculate a mean AUC (+-SD).

6.1.5. Explainability and feature importance

Explainability of machine learning refers to the extent to which the inner workings of a machine learning algorithm can be made visible.[65] In the XGBoost Python package, one of such features is the model's 'feature importance'. The feature importance refers to the individual contribution of features on the final classification algorithm; the higher the feature importance, the more influence that feature has on outcomes. In XGBoost this function allows you to assess feature importance through one of the following metrics: gain, weight or cover.[66] 'Gain' refers accuracy gain in a branch after a node has been split based on this feature. The 'cover' refers to the relative amount of observations of this feature across leaf nodes. Finally, the 'weight' refers to the relative amount of observations of a feature across node splits. The 'gain' parameter has been reported to be the most important in assessing individual feature contributions, since it is the only one that takes into account the actual performance gain the feature brings about.[67] Since XGBoost only allows you to assess one feature importance metric per model, the 'gain' parameter was selected.

A second method that was used for enhancing interpretability is SHapley Additive exPlanations (SHAP). SHAP has its mathematical origins in game theory, where it is used to calculate players' contributions to the final game outcome. This player's contribution has since then been translated to machine learning for assessing feature contributions. The additive nature of SHAP makes them intuitive to use, since the sum of SHAP-calculated feature contributions is equal to the final model prediction probability.[68, 69]

The explainability of the XGBoost models with significant increases in AUC will be looked into. The main objective of using explainability metrics is to verify whether the performance gain can be traced back to the addition of clinical or report features.

6.2. Performance metrics and statistical comparisons

The Area Under the ROC (Receiver-Operating-Characteristic) Curve (AUC), was used as main metric of assessing model performance. For both the XGBoost models with single and multiple clinical and report features, the average AUC (+-SD) over 10 bootstraps was viewed as the overall model performance. The performance of the model incorporating clinical and report features was tested against the benchmark, which consisted of the XGBoost model trained on the same EEG feature set without clinical or report features. The resulting model performances were statistically compared using a non-paired t-test with unequal variance (Welch's t-test).[70] Due to the stochastic nature of the bootstrapping process, it is difficult to justify an exact one-to-one pairing between iterations, as required for a paired t-test.[71] Most likely the variances of the model AUCs are relatively similar between the benchmark and the test. In case of equal variance, Welch's t-test produces equal results to the regular independent t-test (that assumes equal variance). Welch's t-test is calculated by Equation 6.1. An overall significance level of $\alpha = 0.05$ was used. Additionally, because there are 10 different EEG benchmarks to which the clinical and report features are added for evaluation, the Bonferroni correction for multiple testing was used. With 10 testing instances for each clinical or report feature, this results in an adjusted $\alpha = 0.05/10 = 0.005$ for individual tests.

$$t = \frac{\bar{X}_b - \bar{X}_t}{\sqrt{\frac{s_b^2}{n_b} + \frac{s_t^2}{n_t}}} \tag{6.1}$$

where;

$\bar{X}_b$ and $\bar{X}_t$: mean of benchmark (b) AUCs and test (t) AUCs

s_b and s_t : standard deviation (SD) of benchmark (b) AUCs and test (t) AUCs

n_b and n_t : the sample size of benchmark (b) AUCs and test (t) AUCs

6.3. Ensemble models

The ensemble models are a combination of XGBoost models to promote generalizability and performance. For each EEG feature set, the best performing model was included. That could be a model with or without clinical and report features. Model performance was based on the highest mean AUC, even if that did not prove a significant increase over the benchmark model. An ensemble was created by taking the mean of the prediction probabilities of the separate models for each subject in the dataset.[72] This was done for the power set (=all possible combinations) of included XGBoost models, to find the best performing ensemble. Since all included individual models were bootstrapped 10 times in previous analysis, the ensembles could be calculated for 10 iterations as well. The final ensemble performance was represented by the mean AUC (+- SD) over these iterations.

7
Results

7.1. Clinical and report feature characteristics

7.1.1. Relative count of clinical and report feature categories in healthy and epileptic groups

The EEG background rhythms show that there is a large portion of subject EEGs that were annotated with a form of focal sharpness or slowing in both the healthy and epileptic subset, as shown in Figure 7.1. The incidence of focal sharp and focal slow activity was not significantly higher for the epileptic group than in the healthy group, as calculated with the χ^2-test.

But when zooming in on the individual EEG rhythms, there are 2 rhythms that occurred significantly more often in the epileptic group: delta activity (ft_delt) and sharp-and-slow wave complexes (sch_tr_golf), as can be seen in Appendix B.1

All other clinical feature categories were not significantly more abundant in one of the subject groups.

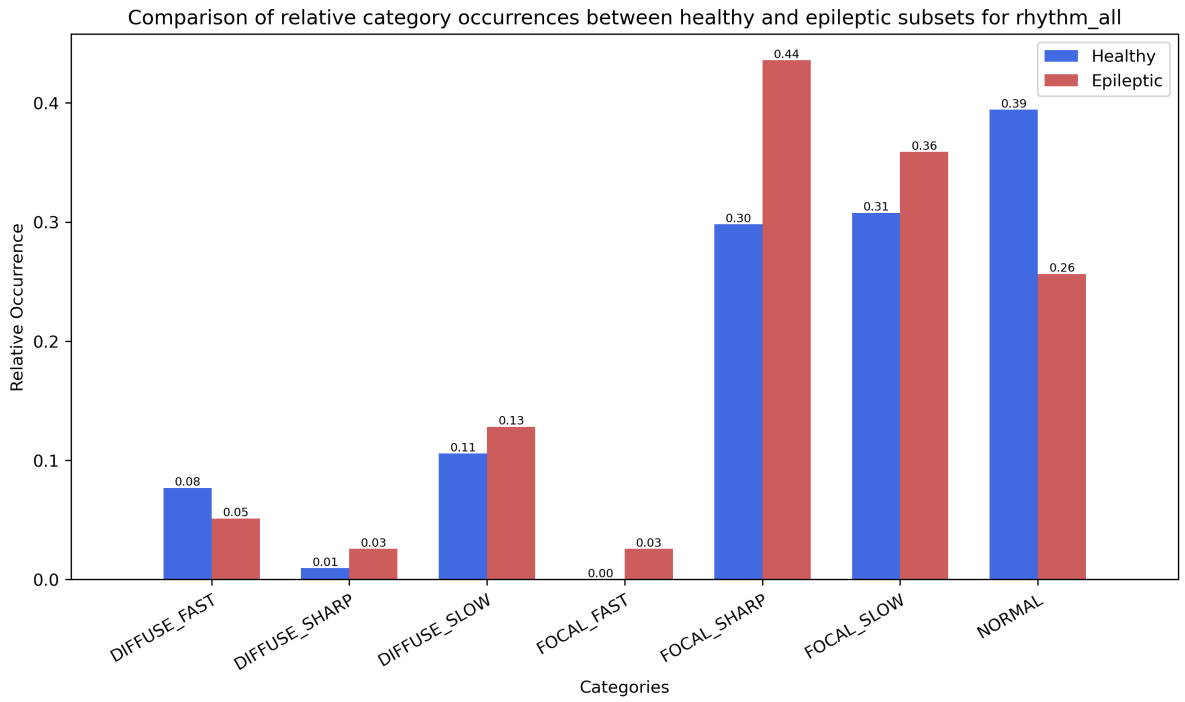


Figure 7.1: Relative category counts of EEG background for healthy and epileptic groups. There were no significance differences for the EEG background features.
[EEG = ElectroEncephaloGram]

7.1.2. Co-occurrence of clinical and report feature categories

The co-occurrence of EEG background rhythms, Figure 7.2, shows that a portion of the reported focal sharpness and focal slowing occurs in the same subjects. A similar trend could not be observed in the co-occurrence of individual EEG rhythms because these individual rhythms were sparser. Since the relative count of the focal sharpness and focal slowing was also notable, these might be feature categories that contain valuable information for training a machine learning model.

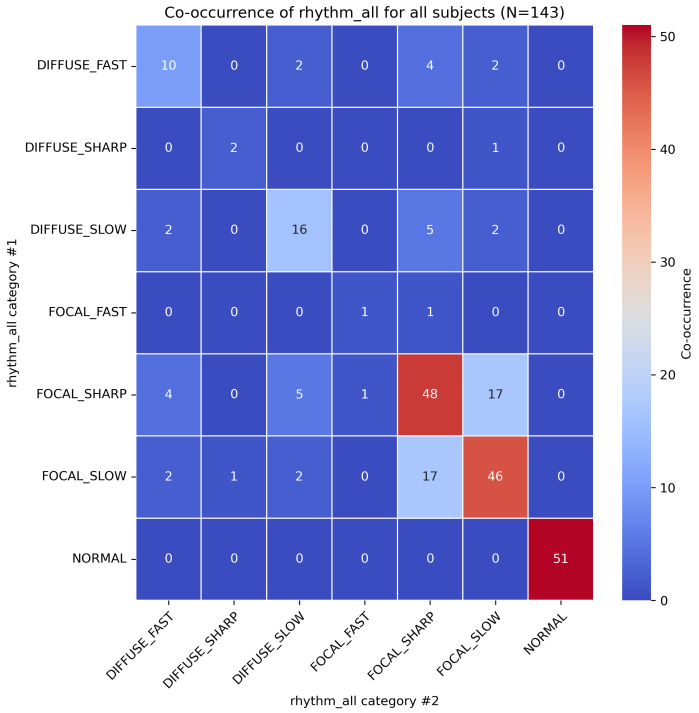


Figure 7.2: Co-occurrence of categories within EEG background. The figure shows the amount of subjects that present with both the EEG background categories of the y-axis as well as the x-axis.
[EEG = ElectroEncephaloGram]

There was some grouping of medications across ATC codes A (alimentary tract and metabolism), B (Blood and blood forming organs), C (Cardiovascular system), and N (Nervous system). The first three all contain quite common types of medication; for example, vitamin supplements and drugs for constipation are found in group A, antithrombotic agents are found in group B, and antihypertensives and diuretics are found in group C. Furthermore, manual inspection of medication data shows that the high count of nervous system medication can be attributed to emergency seizure medication that was administered during the first seizure or was prescribed for possible future seizure clusters. This predates the EEG registration date and was thus included. The co-occurrence of medication categories based on 1st order ATC codes can be found in Figure B.1 in Appendix B.

The co-occurrence of medication categories based on 2nd and 3rd order ATC codes did not show similar grouping, because of the sparsity of included medication categories. The co-occurrence of individual EEG rhythms and medical history did not show clear groupings either.

7.2. Addition of single clinical or report features to EEG featureset for training XGBoost models

The addition of a single clinical or report feature to EEG feature sets, did not significantly improve the XGBoost model performance in all but one case. An overview of test results for all clinical and report features can be found in Appendix C. Although results were not significant, the addition of a single clinical or report feature did increase the mean AUC in certain instances.

When adding the EEG background to XGBoost models, the GPLV feature set shows a significant increase in AUC ($p < 0.001$), even after the Bonferroni correction for multiple testing which changes the significance level to $\alpha = \frac{0.05}{10} = 0.005$. See Figure 7.3.

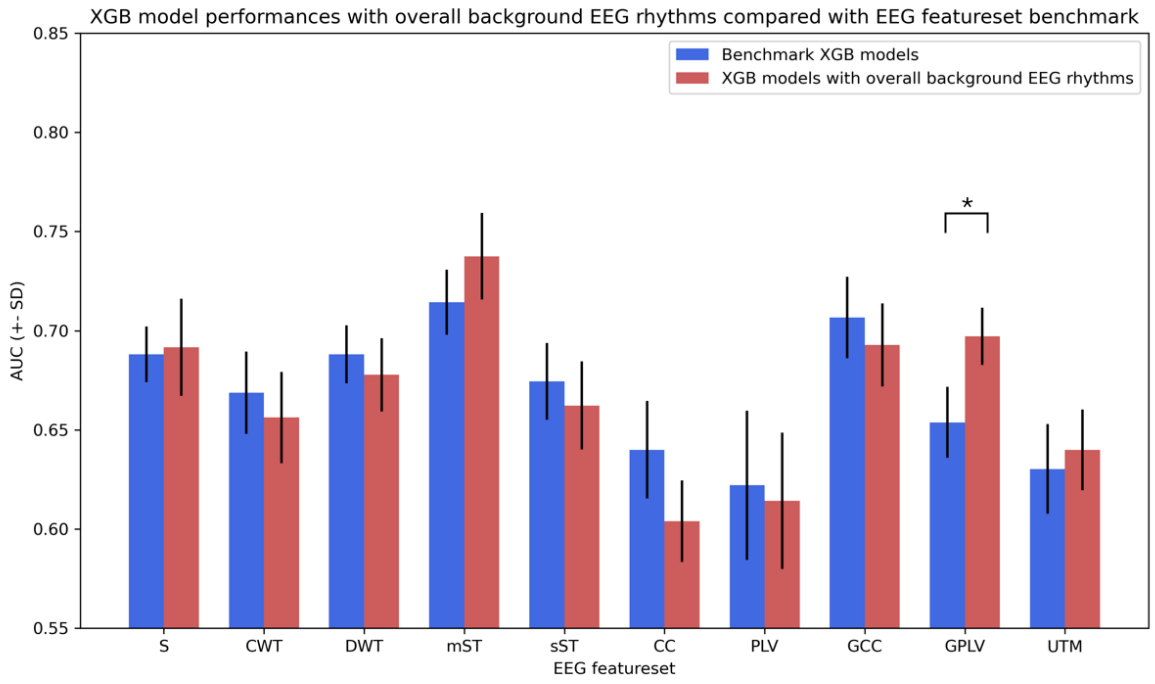


Figure 7.3: AUCs of XGBoost models from EEG feature sets including background EEG, compared to their respective EEG feature set benchmark. Significant differences are indicated by an asterisk.
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

7.3. Optimal set of clinical and report features of each EEG feature set for training XGBoost models

The grid search for optimal combination of clinical and report features for each EEG feature set also allowed the addition of a single clinical or report feature; it could be observed that in many cases the addition of a single clinical or report feature outperformed the addition of multiple clinical or report features. Table 7.1 gives an overview of the clinical features used for each EEG feature set. The mean AUC shows some increase in most EEG feature sets after adding the optimal clinical and report features, as seen in Table 7.1. Again, only GPLV shows a significant difference, but this refers to the same XGBoost model as mentioned in Chapter 7.2.

Table 7.1: An overview of optimal clinical and report features for each EEG feature set and their performance compared to the benchmark. Significant differences are printed in bold font and marked with an asterisk.

EEG feature set	Optimal clinical and report features	Bench. (AUC±SD)	+ Cli. (AUC±SD)
CC	Age	0.640 ± 0.025	0.648 ± 0.025
CWT	Age, Vigilance state, Medical history, Medication (ATC-1), Individual EEG rhythms	0.669 ± 0.021	0.661 ± 0.029
DWT	Age, Sex, Vigilance state, Medical history, Medication (ATC-1), EEG background	0.688 ± 0.015	0.701 ± 0.023
GCC	Medical History	0.707 ± 0.020	0.710 ± 0.024
GPLV	EEG background	0.654 ± 0.018	**0.697 ± 0.014***
mST	EEG background	0.714 ± 0.016	0.738 ± 0.022
PLV	Individual EEG rhthyms	0.622 ± 0.038	0.632 ± 0.041
S	Medical history, EEG background	0.688 ± 0.014	0.709 ± 0.020
sST	Vigilance state, Medication (ATC-1), EEG background	0.675 ± 0.019	0.689 ± 0.018
UTM	Age	0.630 ± 0.023	0.649 ± 0.030

ATC = Anatomical Therapeutical Chemical classification system, EEG = ElectroEncephaloGram, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures

Figure 7.4 shows the ROC curves of the XGBoost models of EEG feature sets including the clinical and report features mentioned in Table 7.1. The XGBoost model based on the mST EEG feature set has the highest mean AUC of 0.738. Its ROC curve stands out from other models around a FPR of 0.2-0.4. Interestingly, ROC curves of XGBoost models from sST and S feature sets also stand out, but at FPRs of 0-0.1 and 0.6-0.9, respectively. This difference in ROC-curve morphology might indicate that (a portion of) their predictive power is based on different population characteristics.

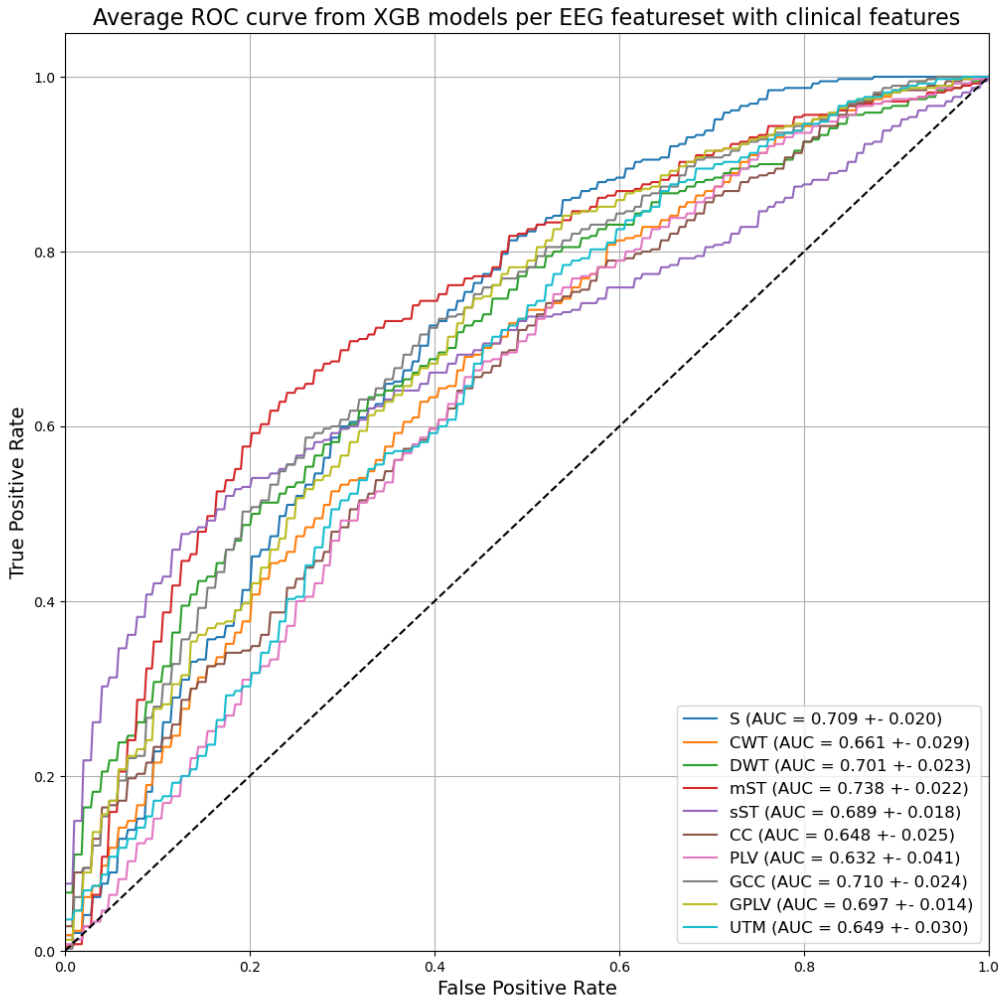


Figure 7.4: ROC curves of XGBoost models from EEG feature sets with their optimal set of clinical and report features. The respective benchmarks for each type of EEG feature set can be observed in Figure 4.2
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

7.4. SHAP and XGB importance

The SHAP values and XGB importance were analyzed for the GPLV feature set including EEG background, since it showed a significant increase in AUC after a clinical or report feature was added.

Figure 7.5 shows a SHAP summary plot for the top 12 contributing features of the GPLV + EEG background XGBoost model. The higher the feature is notated in the figure, the higher its contribution. A positive SHAP value indicates the feature shifts model prediction for that subject towards outcome label '1'/'epilepsy'. A negative SHAP value indicates the feature shifts model prediction towards outcome label '0'/'healthy'. In this plot, the one-hot encoded feature categories were kept separate, so that the individual contribution of each EEG background category could be assessed; red corresponds to 1 (EEG background category is present in subject), blue corresponds to 0 (EEG background category is

not present in subject). Focal sharp activity was among the top contributing features. There is a very clear distinction in SHAP values between the presence and absence of focal sharp activity, with the presence of focal sharp activity shifting model predictions towards epilepsy. Other categories within the EEG background were not among the top contributing features.

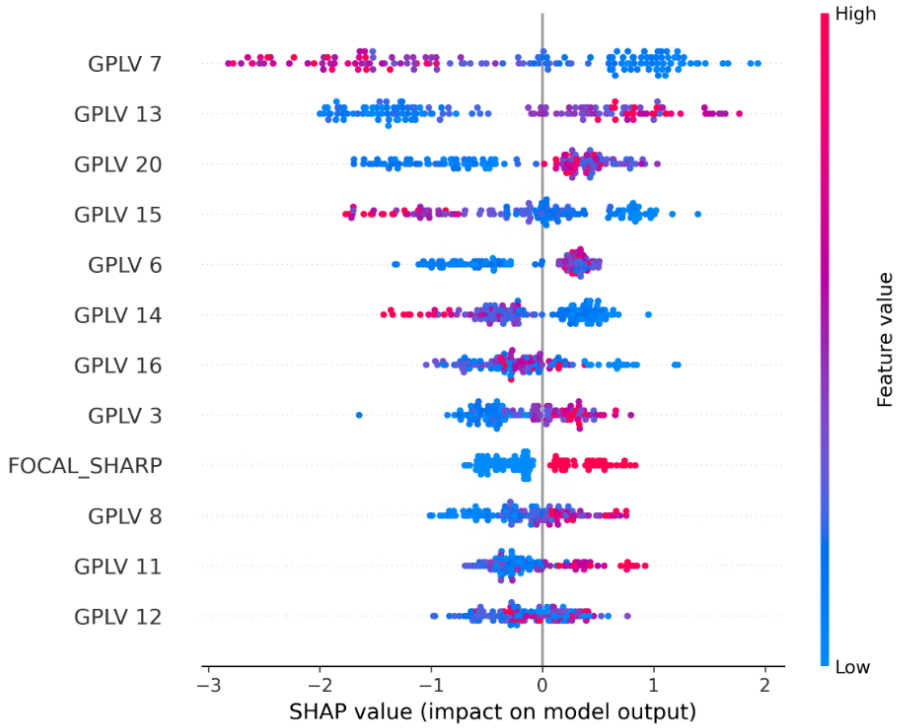
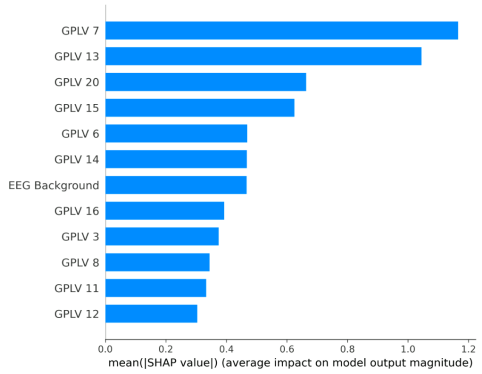


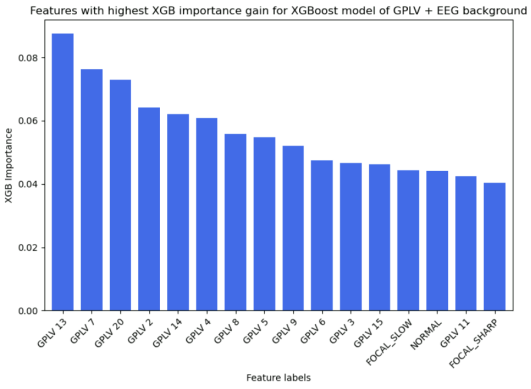
Figure 7.5: SHAP summary plot for the top 12 performing features in the XGBoost model of GPLV feature set including EEG background. The top contributing feature is displayed at the top, with contribution decreasing as the features are lower. Each dot represents a subject in the test set, with the color representing the height of feature value in that subject. The EEG background was split in its categories for evaluating the SHAP values; red corresponds to 1 (category is present in subject), blue corresponds to 0 (category is not present in subject). A positive SHAP value indicates the feature shifts model prediction for that subject towards outcome label '1'/'epilepsy'. A negative SHAP value indicates the feature shifts model prediction towards outcome label '0'/'healthy'.
[EEG = ElectroEncephaloGram, GPLV = Graph measures of Phase Lock Values, SHAP = SHapley Additive exPlanations, XGB(oost) = eXtreme Gradient Boosting]

To evaluate the overall impact of EEG background on the model performance, the SHAP values of EEG background categories could be summed, because of the additive nature of SHAP values. In the summation and evaluation of the overall SHAP values, the absolute value of SHAP values is used to prevent cancelling out positive and negative SHAP values. Figure 7.6a shows the EEG background is among the top contributing features, as was expected from the individual contribution of focal sharp activity. The overall impact of EEG background is higher ranked than focal sharp activity alone, indicating other EEG background categories did contribute to the eventual model predictions.

The features with highest XGB importance are shown in Figure 7.6b. XGB importance does not explicitly state additivity and therefore the categories of EEG background could not be aggregated to show overall XGB importance for the EEG background feature. Instead, the separate categories of EEG background were evaluated. XGB importance shows some overlap with the highest model impact features according to SHAP analysis, as the top 3 contributing features are similar in both analyses. But there are also notable differences, as the features GPLV 2/5/9 rank high up in XGB importance, but are not found in the top 12 contributing features in SHAP analysis. With respect to the XGB importance of EEG background categories, focal slow activity and normal background rank higher up than focal sharp activity. This contrasts SHAP analysis, where almost all EEG background model impact could be attributed to focal sharp activity.



(a) SHAP bar plot for the XGBoost model of GPLV feature set including EEG background. The mean absolute value shows the average impact of a feature on the model output. The absolute SHAP value was taken for EEG background overall instead of its individual categories.



(b) Features with highest XGB importance gain for XGBoost model of GPLV feature set including EEG background. The XGBoost importance gain of EEG background is split in its categories.

[EEG = ElectroEncephaloGram, GPLV = Graph measures of Phase Lock Values, SHAP = SHapley Additive exPlanations, XGB(oost) = eXtreme Gradient Boosting]

Since the optimal clinical and report features for the DWT feature set used all different types of clinical and report features, its SHAP values were also looked into. Focal sharpness from EEG background was again the main factor of model impact. Similar to the GPLV model, the presence of focal sharpness shifted model predictions towards epilepsy, as can be seen in Figure D.1 in Appendix D.

In the SHAP values of the DWT model, 'age' was also among the top contributing features. The SHAP value of age shows that younger age groups shifted the model output towards 'healthy', although this effect was only minor. Figure D.2 demonstrates that all other clinical or report features had very small contributions to model output. Even when summing the SHAP values from all categories within a clinical or report feature, none were among the most contributing features.

7.5. Ensembles of EEG feature sets including their optimal clinical and report features

These results show ensembles that combine 2, 3, 4, 5, 6, and 7 individual EEG feature set models. For each EEG feature set, the XGBoost model with clinical and report features from Table 7.1 Incorporating more than seven EEG feature sets did not yield additional performance gains. Figure 7.7 shows the top-performing ensemble for each combination size, as measured by the highest AUC. The ensembles demonstrated significant improvements in model performance compared to the individual XGBoost models.

The XGBoost models from sST and mST form the foundation of all top performing ensemble models. Their high performance at lower to average FPR values, referring back to Figure 7.4, seem to translate into good performing ensembles. Notably, as the size of the ensemble increases, the best-performing ensemble consistently includes the previous combination of XGBoost models. This is likely because the addition of a new model only slightly alters the overall prediction probabilities generated by the previous ensemble. This is further illustrated by the diminishing ensemble performance gain as a single model is added to larger ensembles.

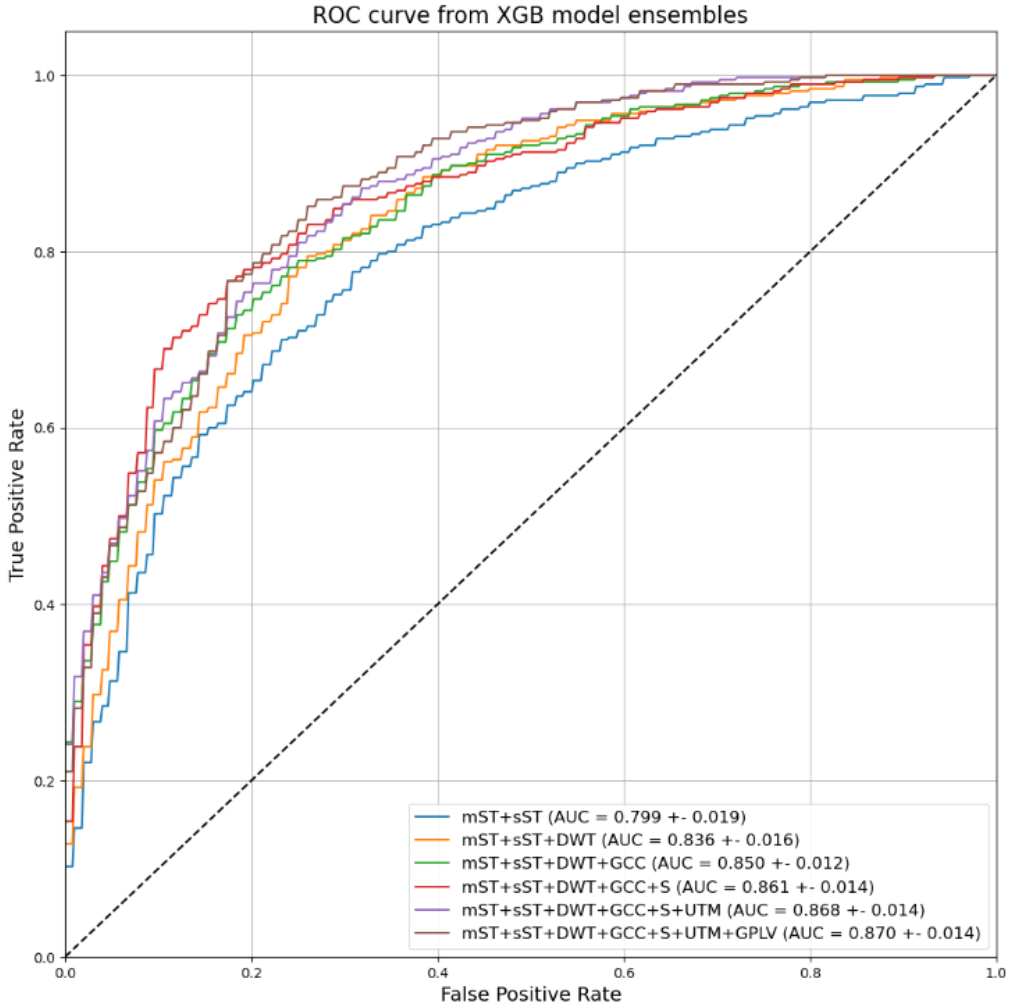


Figure 7.7: ROC curves corresponding to the XGBoost model ensembles. EEG featuresets that were used in the ensembles included their optimal clinical and report features.
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

8

Clinical Implementation

8.1. Bayesian statistics

Bayesian statistics, as first described by Thomas Bayes in the 1700s, are a type of statistic to update the probability of a hypothesis based on (new) observations. The application of Bayesian statistics use Bayes' Theorem (Equation 8.1) as the backbone. This theorem hinges around the prior (or initial) probability of a hypothesis, which is given by $P(A)$, where $P()$ is the notation of a probability, and A denotes the positive outcome to be evaluated.[73] The prior probability is the likelihood that an individual within the population satisfies event 'A', if no other observations or characteristics of an individual are known. Bayes Theorem (Equation 8.1) then describes how an observation 'B' changes the probability of a subject to satisfy event 'A'. This altered probability is called the posterior probability and is given by $P(A \mid B)$; which means the probability an individual satisfies event A, given observation B.[74]

Standard Bayes' Theorem; the posterior probability of A given B:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{8.1}$$

where;

$P(A \mid B)$: The posterior probability of A given B

$P(B \mid A)$: The likelihood of observing B given that A is true.

$P(A)$: The prior probability of A

$P(B)$: The total probability of B, irrespective of A

This total probability of B can be difficult to deduce from model performance metrics. Using the 'law of total probability', we can rewrite $P(B)$ into terms that are dependent on A.[75] The law of total probability decomposes the total probability of B into different scenarios of A, e.g. A_1, A_2, ... A_n. Here the sum of the probability of B for each scenario of A is equal to the total $P(B)$. Since this research employs 2 outcome labels, we express $P(B)$ only in 2 scenarios; A, and $\neg A$ (notation for not A)

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A) \tag{8.2}$$

where;

$P(B)$: The total probability of B, irrespective of A.

$P(B \mid A)$: The probability of B given A is true.

$P(A)$: The prior probability of A

$P(B \mid \neg A)$: The probability of B given that A is not true.

$P(\neg A)$: The prior probability that A is not true.

Substituting Equation 8.2 into Equation 8.1, yields;

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A)} \tag{8.3}$$

8.2. Translation to model performance metrics

We evaluate the models in this thesis mainly on the AUC of the ROC-curve, which itself is a depiction of a models' TPR, and FPR (1-specificity) when varying the prediction threshold. To re-iterate; the TPR (also known as sensitivity, or $sens$ for short) is defined as the fraction of positive subjects that were rightly identified by the model. The FPR (which equals '1 - specificty') is defined as the fraction of negative subjects that were unjustly identified as being positive. The specificity (or $spec$ for short), is also known as the true negative rate

If we let A denote the outcome of a subject with a positive case indicating epilepsy, and B the prediction (or observation) of our model, the contents of Equation 8.3 translate to;

- $P(A)$: The prior probability of a subject being a positive case.
- $P(B \mid A)$: The probability the model predicts a positive case, given the subject is positive, which equals the $sens$,
- $P(\neg A)$: The prior probability of a subject being a negative case. All subjects either belong to A or $\neg A$. Hence, $P(\neg A)$ is equal to $1 - P(A)$
- $P(B \mid \neg A)$: The probability the model predicts a positive case, while this is not true. In other words, the FPR (or $1 - spec$)
- $P(\neg A)$: The prior probability that A is not true.

Substituting the model performance metrics into Equation 8.3 where possible, gives us an equation for the posterior probability of a subject being a positive case, given the model prediction is positive (Equation 8.4);

$$P(A \mid B) = \frac{sens \cdot P(A)}{sens \cdot P(A) + (1 - spec) \cdot (1 - P(A))} \tag{8.4}$$

where;

- $P(A \mid B)$: The posterior probability of a subject being positive (A) given the model prediction is positive (B)
- $P(A)$: The prior probability of a subject being a positive case.
- $sens$: The sensitivity of the evaluated model
- $spec$: The specificity of the evaluated model

As can be deduced from Equation 8.4, changing the sensitivity and specificity influences the relation between the $P(prior)$ and $P(posterior)$. This is illustrated in Figure 8.1

8.3. Implementing the ILAE 2014 definition of epilepsy

The ILAE definition of epilepsy defines multiple ways through which a patient can be (clinically) diagnosed with epilepsy, as mentioned in Chapter 2. The second criterion states a risk of >60% of having a recurrent seizure is necessary to clinically diagnose someone with epilepsy. Thus, in order for a predictive model to be clinically relevant, the posterior probability of having a recurrent seizure should be >60% (or >0.6 in our notation).[3]

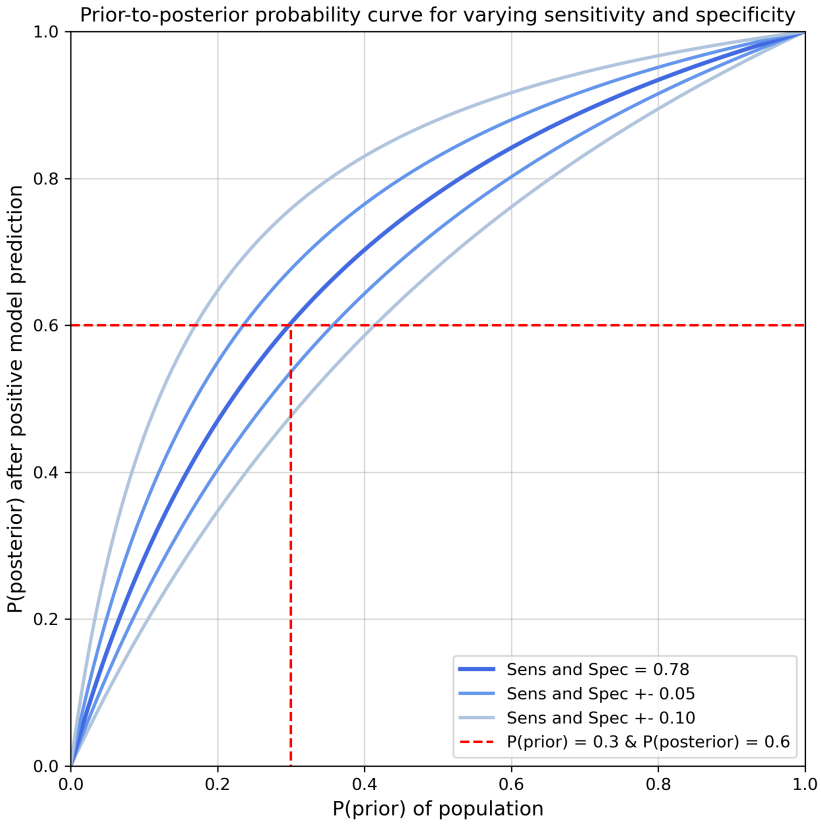


Figure 8.1: The relation between $P(prior)$ and $P(posterior)$ for varying sensitivity and specificity as calculated by Equation (8.4). The dashed line in red corresponds to the relevant values of $P(prior)$ and $P(posterior)$ for this research according to Chapter 8.3
[sens = sensitivity, spec = specificity]

Furthermore, the prior probability of a person having epilepsy can be deduced from our dataset. Since all patients were consecutively included from 2010 to 2022 we can interpret the ratio of epileptic vs. healthy subjects in our dataset as a representative prior probability for this specific population. Furthermore, the epileptic label within this study fits well with the ILAE definition of epilepsy, since it was defined based on clinical follow-up of recurrent seizures. The observed prior probability of recurrent seizures in our study population was approximately equal to 0.3.

Since we are interested in the required model performance in terms of sensitivity and specificity, we first input the required posterior probability and the observed prior probability into Equation 8.4:

$$0.6 = \frac{0.3 \cdot sens}{0.3 \cdot sens + 0.7 \cdot (1 - spec)} \tag{8.5}$$

Cross multiplication gives;

$$0.6 \cdot (0.3 \cdot sens + 0.7 \cdot (1 - spec)) = 0.3 \cdot sens \tag{8.6}$$

Simplifying terms;

$$0.7 \cdot (1 - spec) = 0.2 \cdot sens \tag{8.7}$$

Rearranging terms to get an expression of $spec$ in terms of $sens$

$$spec = 1 - (\frac{2}{7}sens) \tag{8.8}$$

Equation 8.8 shows the conditional relation between the specificity and sensitivity of a predictive model to satisfy the recurrence risk of 60% after a positive model prediction. Plotting the specificity against the sensitivity results in Figure 8.2. Here the green dashed line depicts the combinations of specificity and sensitivity that exactly satisfy $P(posterior) = 0.6$. The shaded area in green is the result of all possible combinations of specificity and sensitivity that yield $P(posterior) \geq 0.6$ if the model prediction is positive (epilepsy). An immediate observation that can be seen from Figure 8.2 and Equation 8.8, is that irrespective of the sensitivity of a model, the specificity must at least be $\frac{5}{7}$ or $\simeq 0.71$ to produce a $P(posterior) \geq 0.6$.

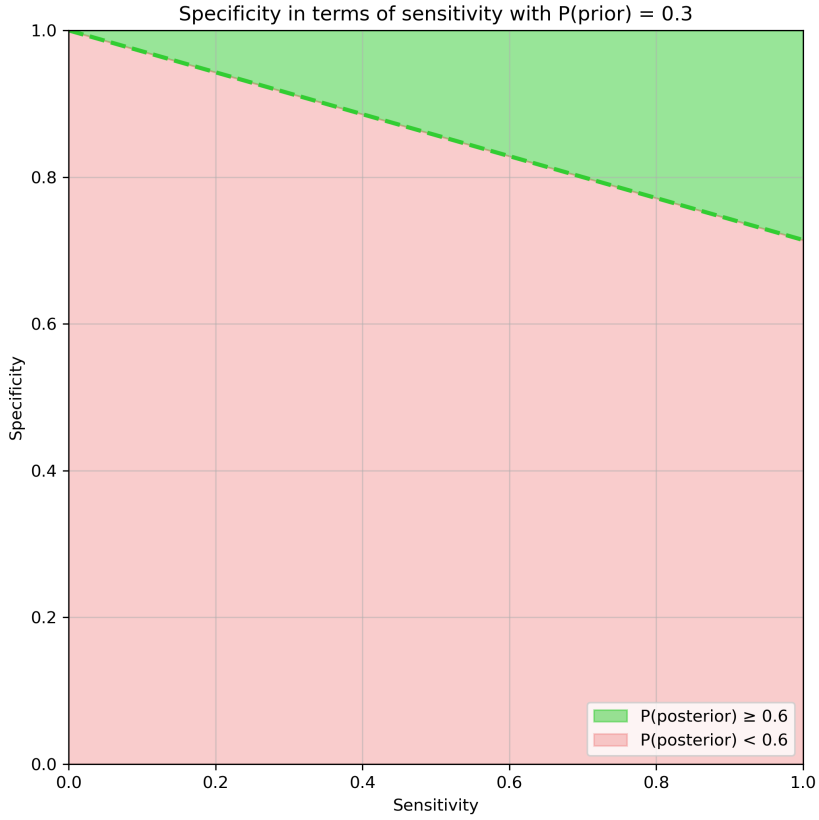


Figure 8.2: The relationship between the required specificity and sensitivity of a prediction model to satisfy $P(posterior) \geq 0.6$ given that $P(prior) = 0.3$, as expressed in Equation 8.8

8.4. Results in the ROC-domain

Expressing model performance in a single combination of specificity and sensitivity lacks insight into the influence of thresholding on the model performance. With the purpose of evaluating a model across thresholds, we turn to the ROC-domain. Here the y-axis corresponds to the sensitivity (or true positive rate) and the x-axis corresponds to the false positive rate (or (1 - specificity)).

Rewriting Equation 8.8 to express the sensitivity in terms of $(1 - specificity)$ gives;

$$sens = \frac{7}{2}(1 - spec) \tag{8.9}$$

To exemplify the implementation of Equation 8.9 we use the ROC-curve of a model with an AUC of $\simeq 0.8$. Figure 8.3 shows the ROC-domain with the example model performance curve in blue and Equation 8.9 represented by the green line. Here, the area shaded in green represent all combinations of specificity and sensitivity that satisfy the recurrent risk of seizure >60%. With darker green shading corresponding to the area enclosed by the model ROC, which would indicate the model performs above the threshold of $P(posterior) \geq 0.6$

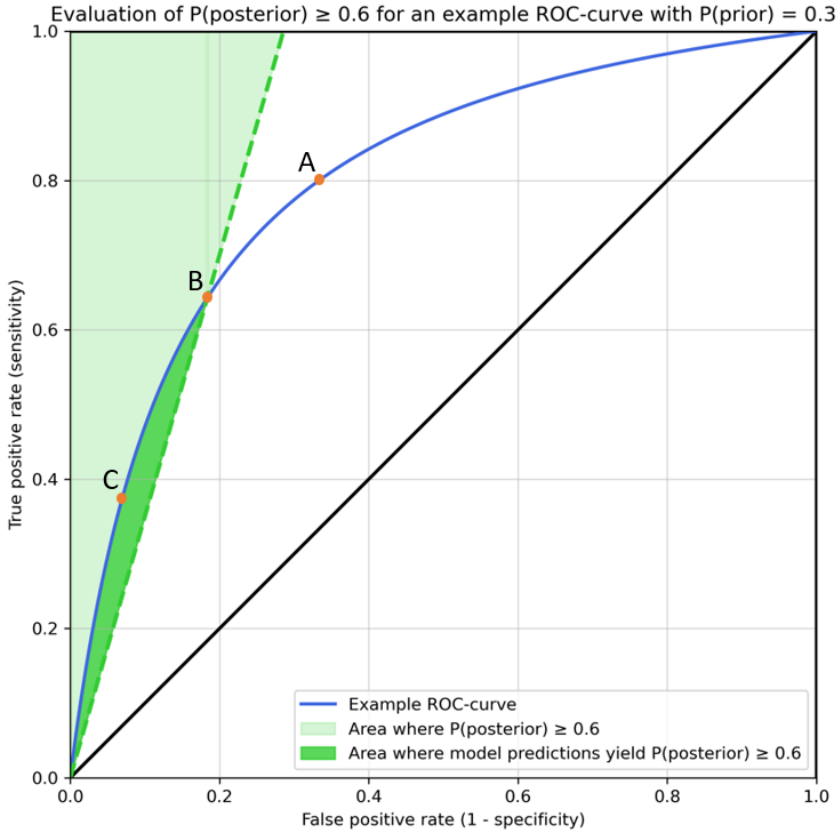


Figure 8.3: An example ROC-curve with an AUC $\simeq 0.8$ to exemplify model performance that falls within the area of $P(posterior) \geq 0.6$
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic]

Using an example population (N=100) that shares the same $P(prior)$ as the Erasmus MC dataset, we can show what the different thresholds mean for subject classification. Figure 8.3 shows three points on the example ROC curve; A, B, and C. Each point corresponds to a certain threshold of prediction probabilities of the XGBoost model. Each point has a corresponding sensitivity and specificity. Table 8.1 gives an overview of classifications in the population for the corresponding points on the ROC-curve. The $P(posterior)$ has been calculated according to Equation 8.4. For point A, outside the shaded area, the $P(posterior) = 0.51$, for both point B and C $P(posterior) \geq 0.6$. Interestingly, we can observe that the positive predictive value ($\frac{TP}{TP+FP}$) coincides with the $P(posterior)$. Since the sample size of this population is small, there is a difference due to rounding errors. From that we can conclude that in stating the minimum recurrence risk of seizure, the ILAE indirectly states the relative amount of false positives for clinical epilepsy diagnosis that is deemed acceptable.

Table 8.1: An overview of classifications in a population of 100 subjects for different points along the ROC curve of Figure 8.3, with a $P(prior) = 0.3$

	Point A	Point B	Point C
Total subjects (N)	100	100	100
Healthy	70	70	70
Epileptic	30	30	30
P(prior)	0.3	0.3	0.3
Sensitivity	0.80	0.64	0.37
Specificity	0.67	0.82	0.93
P(posterior)	0.51	0.60	0.71
True positive	24	19	11
False positive	23	13	5
True negative	47	57	65
False negative	6	11	19
Positive predictive value	0.510	0.594	0.680

ROC = Receiver-Operating Characteristic

8.5. Metrics for evaluating model performance based on XGBoost models from this research

Figure 8.4 shows some of the best performing models based on a single EEG feature set (including clinical and report feature(s)). The AUC has been used throughout this research as a measure of model performance. However, only looking at the AUC might not accurately describe model performance in the context of diagnosing epilepsy using machine learning. For example, we see that even though the XGBoost model trained on S or GCC have AUCs above 0.7, their ROC curve is never inside the $P(posterior) \geq 0.6$ area. Contstrastingly, the XGBoost model trained on sST, with a lower AUC, has a considerable amount of thresholds that result in model performance within the $P(posterior) \geq 0.6$ area. This shows that ROC-curves that are skewed towards a low false positive rate are favoured over those that perform consistently over the entire false positive rate range.

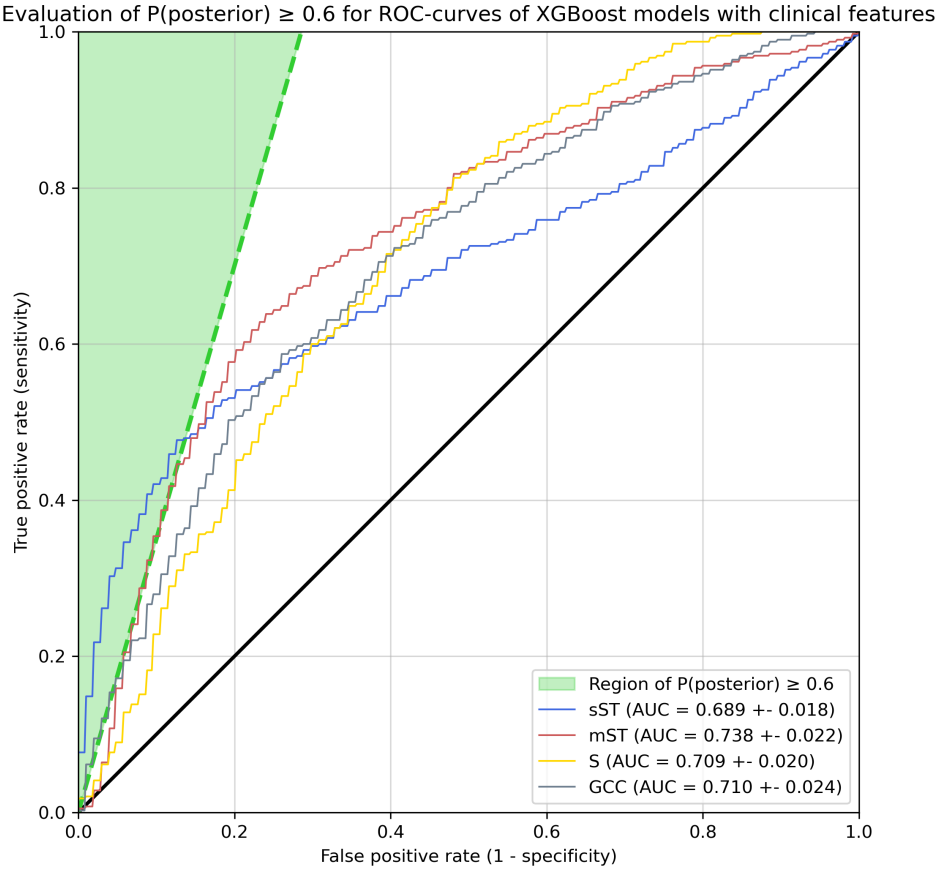


Figure 8.4: ROC-curves corresponding to XGBoost models trained on single EEG featuresets with their respective optimal clinical and report features. The ROC curves are compared to the green region defined by $P(posterior) \geq 0.6$
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, GCC = Graph measures of Cross-Correlation, mST = mean Stockwell Transform, sST = square root Stockwell Transform, S = Spectral, XGB(oost) = eXtreme Gradient Boosting]

Previous research also used the Balanced Accuracy (BAC) at 80% sensitivity as a metric for model performance.[17, 18] The balanced accuracy can be calculated as: $BAC = \frac{sens+spec}{2}$. The balanced accuracy is a useful metric, because it takes into account both the sensitivity and specificity of a model at a certain threshold. Preceding analysis of Bayes Theorem and the ILAE definition of epilepsy have clarified the need of a minimal $P(posterior)$ of 0.6, and it was illustrated that high specificity is necessary to attain high values of $P(posterior)$. Setting a rigid boundary of sensitivity at 80%, comes at the cost of specificity, and thus the $P(posterior)$.

The 'maximum sensitivity at $P(posterior) = 0.6$' is proposed as a metric for evaluating model performance in future research. It aims to give insight into the model performance that falls within the $P(posterior) \geq 0.6$-zone, while still being intuitive to understand. It can be calculated while creating the ROC-curve of a model, using the FPR and TPR. By adjusting Equation 8.4, so it is based on TPR (=sensitivity) and FPR (= 1- specificity), we can calculate the $P(posterior)$ for each point on the ROC curve through Equation 8.10. Then the $P(posterior)$ can be set at 0.6, and finally a grid search for the maximum sensitivity among those point on the ROC will give the proposed model performance metric.

$$P(posterior) = \frac{TPR \cdot P(prior)}{TPR \cdot P(prior) + FPR \cdot (1 - P(prior))} \tag{8.10}$$

8.6. Performance of XGBoost model ensembles from this research

The optimal model ensembles discussed in Chapter 7.5 are presented together with the region defined by $P(\text{posterior}) \geq 0.6$ in Figure 8.5. All ensembles have a substantial portion of their ROC curves within the $P(\text{posterior}) \geq 0.6$ region. The highest sensitivity at $P(\text{posterior}) = 0.6$ was achieved by the ensemble combining seven XGBoost models, yielding a value of 0.81. This performance is closely followed by the ensemble combining five XGBoost models, with a sensitivity of 0.79 at $P(\text{posterior}) = 0.6$.

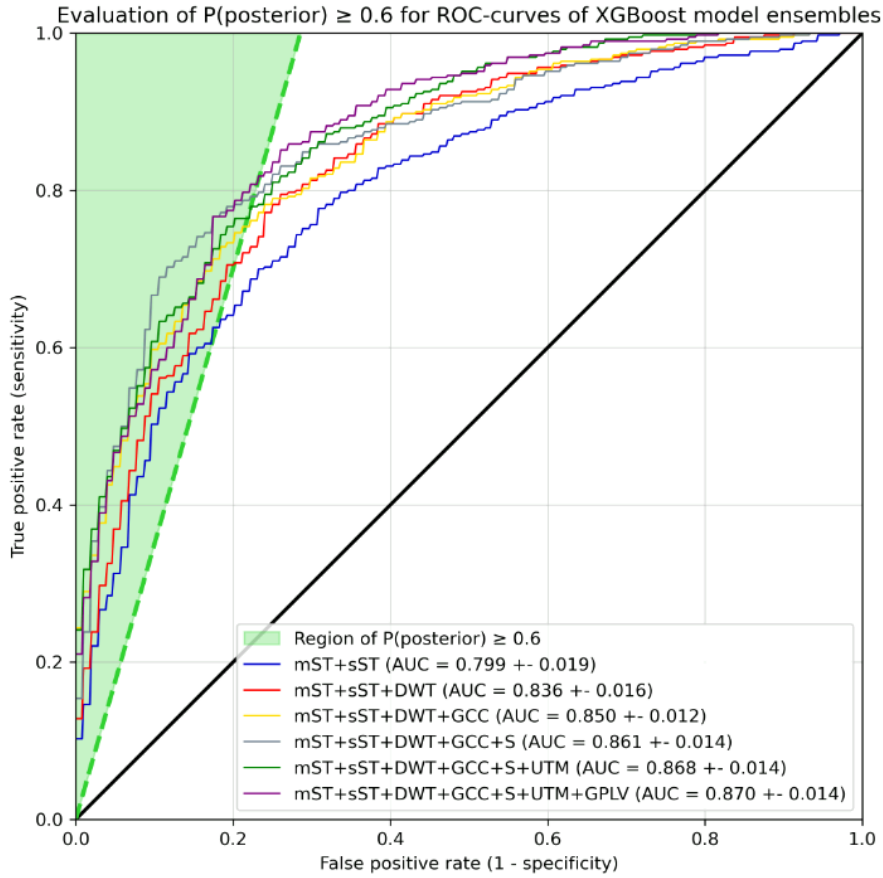


Figure 8.5: ROC-curves corresponding to ensembles of XGBoost models trained on single EEG feature sets with their respective optimal clinical and report features.
[AUC = Area Under the (ROC) Curve, ROC = Receiver-Operating Characteristic, EEG = ElectroEncephaloGram, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, XGB(oost) = eXtreme Gradient Boosting]

9 Discussion

Thangavel et al 2022 has indicated that IED-free EEGs across multiple datasets might contain valuable diagnostic information.[17] The work of Y. Mirwani showed that the implications are similar in the Erasmuc MC dataset.[18] The addition of clinical and report features in this study does not prove a drastic increase in model performance, while the ensembles of the XGBoost models from this research did show major improvement. More so, this research shows the way model performance should be evaluated. Where AUC and BAC (at sensitivity of 80%) were the main metrics that were used by Thangavel and Mirwani, this thesis shows the importance of model performance maintaining a $P(Posterior) \geq 0.6$. To that extent, an alternate metric of model performance has been suggested in Chapter 8; the maximum sensitivity at $P(posterior) \geq 0.6$. By assessing future test models according to this metric, model development will target clinical need more accurately.

The approach in this study assumes that the recurrence risk of seizures can be quantitatively modeled and used in a mathematical framework for calculating disease probabilities. There have been some online tools to calculate recurrence risk of seizure, such as described by van Diessen et al 2018.[76] They use patient characteristics to estimate recurrence risk of seizure to guide future treatment. Additionally, the recurrence risk associated with certain etiologies has also been described in order to make an estimation of seizure recurrence.[77] Similarly, Bonnet et al 2022 included available patient characteristics for predictive modeling and indicates its potential use in clinical practice. Beaulieu-Jones et al 2023 used more advanced modeling to predict seizure recurrence from routine clinical notes, following an initial seizure. They show that their large language models can significantly contribute to predicting seizure recurrence risk.[78] The common denominator of these studies is the fact that they only use clinical or interpreted features, and did not combine them with calculated EEG-level features. Even though not all of them employ machine learning techniques, they do indicate that predictive modeling for estimating recurrence risk of seizure is an accepted methodology.

A potential pitfall of using probability statistics on machine learning outcomes stems from the representativeness of training and test data. As shown in Chapter 8 for a perfectly representative population, the $P(posterior)$ should coincide with the positive predictive value. If the training or test data is skewed towards either side, calculating the $P(posterior)$ from the ROC curve gives an over- or underestimation of recurrence risk. This can also be observed from the prior-to-posterior probability curve from Figure 8.1. For equal model performance, relatively small changes in prior probability can have a big impact on the posterior probability. Since in the future we aim to base the clinical diagnosis of epilepsy partially on this value, its precision is of great importance. This study sought to ensure a representative population by consecutively enrolling patients. However, inevitable patient exclusions introduced a degree of selection bias, which may have influenced the prior probability. Prospectively defining inclusion and exclusion criteria for training, testing, and validation datasets can address this issue. It is crucial that the resulting model is applied to a population that adheres to the same predefined criteria to ensure reliability in clinical practice.

An approach to incorporate machine learning into the current workup for epilepsy diagnosis, is to integrate the predictions of the XGBoost model into a Bayesian Network. A Bayesian Network is a graph model in which variables are depicted as nodes, connected by edges that define probabilistic dependencies between them. This structure allows for the modeling of inter-variable relationships by specifying the conditional probabilities along the edges.[79] Variables can be chosen to represent patient characteristics or known risk factors for epilepsy. The output of a machine learning model can be represented by such a node as well, rather than being interpreted as a diagnosis in itself. Bayesian Networks have

been successfully applied in patients with epilepsy for the real-time predictions of seizures.[80] Additionally, the combination of Bayesian Network methodologies with deep learning neural networks has been used to analyze EEG data for tasks such as sleep stage classification, demonstrating the potential of such hybrid approaches in medical diagnostics.[81]

In clinical practice, the current standard of care advises against initiating epilepsy treatment unless the ILAE's criteria for epilepsy are met. This study aligns with that guideline by shifting the evaluation of machine learning models towards lower false positive rates. Since model performance has some inherent uncertainty when applied to unseen data, clinical implementation at the exact boundary of $P(posterior) = 0.6$ is not recommended. To obtain sufficient support in clinical practice for initial prospective studies, the required model performance should be discussed with the neurologists in question.

One of the critical limitations of this study lies in the availability and quality of clinical data, which poses challenges for the generalization and robustness of the machine learning models. The lack of available clinical data poses a risk in two ways: first, the medical history of a patient is often unavailable when treated in another hospital in the past, creating gaps in the dataset. Second, the dataset's small population size and the high cardinality and sparsity of some clinical and report features hinder the ability of machine learning models to identify and learn representative patterns effectively. Increasing the population size and focusing on the features that have less split-up categories might increase the contribution of clinical and interpretation features.

Additionally, the lack of consistent in-hospital reporting of patient medical history in EEG reports further limits the dataset. Medical history, obtained through anamnesis, often contained valuable information that could be used as medical history. A potential solution to this issue would involve employing text-mining techniques on the complete EPD. As medical history is often summarized within updates in the EPD, automated text-mining approaches could extract and structure this information for inclusion in the feature set. This approach would not only improve the completeness of the dataset but also offer a scalable way to integrate a patient's medical history into machine learning pipelines.

SHAP values and XGBoost feature importance were used in this study to gain insight into feature contributions. However, a complete agreement between the two approaches was not observed. Because XGBoost did not allow multiple feature importance metrics to be evaluated at once, only the 'gain' metric was used, where XGBoost also offers options for 'cover' and 'weight'. In contrast, SHAP values provide a representation of a feature's contribution to model predictions as a single metric.

Notably, focal sharp activity emerged as a potentially contributing feature in both its initial characteristics analysis and the SHAP analysis. This finding aligns with current diagnostic practices, where some sharp activity is denoted as an IED. It is important to note that the sharp activity that was observed in the dataset of this study, was never deemed an IED. Being able to use residual sharp activity in the EEG for additional classification performance would represent a meaningful advancement in using machine learning to support epilepsy diagnosis. An interesting observation is that in Figure 7.3, not all XGBoost models seem to share this improvement in model performance when adding the interpretation of EEG background as a feature. A possible explanation is that the added discriminatory value of residual sharp activity is only present after earlier splits in the decision tree based on specific features from the EEG feature sets.

From the perspective of the XGBoost algorithm, most of the identified optima were determined using grid search methods, with each search performed twice to account for stochastic variability. However, it is likely that local, rather than global, optima were identified for both the benchmark and clinical/report feature models. Ideally, an iterative optimization process would have been used that considers all possible hyperparameter configurations of both the EEG feature sets and the clinical/report features simultaneously. This approach would avoid appending clinical parameters to potentially suboptimal configurations of EEG feature sets. However, due to the exponential increase in computational complexity associated with exploring all such combinations, this approach was not feasible in this study.

It was aimed to promote generalizability in the XGBoost models by including stochastic elements during training. However, all performance metrics were derived using LOSO CV and were not evaluated on an unseen validation set. This limitation makes the assessment of the models' generalizability to

unseen data difficult. The strong performance demonstrated by the ensembles calls for thorough validation on an unseen validation set before applying these findings to subsequent research. Additionally, in these ensembles, some degree of overlap among the included features was possible. The best predictive models for each EEG feature set were used, which included their optimal clinical and report features. Across these individual models, clinical or report features could occur multiple times. This overrepresentation could bias the ensemble model toward these features, giving them disproportionate weight compared to others. However, this issue may not be limited to clinical and interpretation features alone. The EEG feature sets themselves might also exhibit some overlap in feature information. Although each EEG feature set uses a distinct mathematical approach to calculate features, this does not necessarily prohibit them from capturing the same underlying EEG characteristic in their feature information.

Since the entire dataset was utilized for both training and testing during the development of the XGBoost models and ensembles, an unused validation set was not readily available. A potential solution to this is the use of nested cross-validation, which splits the dataset into training and testing subsets, with an additional holdout set per iteration for performance evaluation. This process can be repeated across multiple folds, and the aggregated ROC curve across folds provides an estimate of the model's generalizability to unseen data. It is important to note that even with nested cross-validation, the validation is limited to patients within the same dataset. Therefore, the results may not be directly transferable to external datasets.

10 Conclusion and future research

10.1. Conclusion

The addition of clinical and report features is not guaranteed to improve results in an XGBoost model based on EEG features for the classification of IED-free EEGs. The variability in the optimal clinical and report features across different EEG feature sets suggests that the impact of clinical and report features varies depending on the specific EEG feature set used. The combination of individual XGBoost models into an ensemble appears to substantially enhance overall performance. Considering the International League Against Epilepsy (ILAE) definition of epilepsy and applying probabilistic statistical methods, it has been proposed that the maximum sensitivity at $P(posterior) \geq 0.6$ serves as a valuable metric for evaluating the performance of future machine learning models. The top-performing ensemble in this study achieved a sensitivity of 0.81 based on this evaluation metric. However, these findings have not been validated yet on an external validation data set.

10.2. Future Research

Future research should focus on validation to enhance the robustness and applicability of the findings. Initially, validation should be conducted using prospectively included new data, nested cross-validation, or an external, unseen dataset that describes a similar population with a representative ratio of epileptic cases.

Subsequently, it may be necessary to explore methodologies for reliably determining the prior probability specific to a patient cohort. This includes defining the appropriate inclusion and exclusion criteria for patient selection, ensuring that the model is tailored to clinically relevant populations.

Additionally, advanced techniques such as text mining and large language models could be investigated to extract clinical characteristics directly from electronic health records. This approach could provide clinical features more comprehensively, further clarifying the added value of clinical and EEG interpretation features in predictive models.

Future studies should also explore the application of deep learning models as an alternative to XGBoost. Neural networks may uncover more complex and nuanced relationships between clinical, report, and EEG features, offering a potential advantage over decision tree-based methods.

Finally, all models developed in future research should be evaluated using the sensitivity at $P(posterior) \geq 0.6$ as the primary metric. This will ensure alignment with clinical guidelines and provide a standardized metric for assessing model performance.

References

1. Steinmetz, J. D. *et al.* Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Neurology* **23.** Publisher: Elsevier, 344–381. ISSN: 1474-4422, 1474-4465. https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(24)00038-3/fulltext (2024) (Apr. 1, 2024).
2. *Epilepsy* https://www.who.int/news-room/fact-sheets/detail/epilepsy (2024).
3. Fisher, R. S. *et al.* ILAE Official Report: A practical clinical definition of epilepsy. *Epilepsia* **55,** 475–482. ISSN: 0013-9580, 1528-1167. https://onlinelibrary.wiley.com/doi/10.1111/epi.12550 (2025) (Apr. 2014).
4. Rizvi, S., Ladino, L. D., Hernandez-Ronquillo, L. & Téllez-Zenteno, J. F. Epidemiology of early stages of epilepsy: Risk of seizure recurrence after a first seizure. *Seizure* **49,** 46–53. ISSN: 1059-1311. https://www.sciencedirect.com/science/article/pii/S1059131117301188 (2025) (July 1, 2017).
5. Berg, A. T. Risk of recurrence after a first unprovoked seizure. *Epilepsia* **49.** _eprint: https://onlinelibrary.wiley.com/do 1167.2008.01444.x, 13–18. ISSN: 1528-1167. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1528-1167.2008.01444.x (2025) (s1 2008).
6. Noachtar, S. & Rémi, J. The role of EEG in epilepsy: A critical review. *Epilepsy & Behavior. Management of Epilepsy: Hope and Hurdles* **15,** 22–33. ISSN: 1525-5050. https://www.sciencedirect.com/science/article/pii/S1525505009000924 (2025) (May 1, 2009).
7. Askamp, J. & van Putten, M. J. A. M. Diagnostic decision-making after a first and recurrent seizure in adults. *Seizure* **22,** 507–511. ISSN: 1532-2688 (Sept. 2013).
8. *What is EEG (Electroencephalography) and How Does it Work? - iMotions* Section: Academia. https://imotions.com/blog/learning/research-fundamentals/what-is-eeg/ (2025).
9. Epileptiform Discharges: Overview, Distinction From Normal or Nonspecific Sharp Transients, Localization and Clinical Significance of IEDs. Publication: Medscape - eMedicine. https://emedicine.medscape.com/article/1138880-overview?form=fpf (2025) (Jan. 16, 2024).
10. Kural, M. A. *et al.* Criteria for defining interictal epileptiform discharges in EEG. *Neurology* **94.** Publisher: Wolters Kluwer, e2139–e2147. https://www.neurology.org/doi/10.1212/WNL.0000000000009439 (2025) (May 19, 2020).
11. Renzel, R., Baumann, C. R. & Poryazova, R. EEG after sleep deprivation is a sensitive tool in the first diagnosis of idiopathic generalized but not focal epilepsy. *Clinical Neurophysiology* **127,** 209–213. ISSN: 1388-2457. https://www.sciencedirect.com/science/article/pii/S1388245715006410 (2025) (Jan. 1, 2016).
12. Giorgi, F. S. *et al.* Controversial Issues on EEG after Sleep Deprivation for the Diagnosis of Epilepsy. *Epilepsy Research and Treatment* **2013,** 614685. ISSN: 2090-1348. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3694384/ (2025) (2013).
13. Harris, L. & Angus-Leppan, H. Epilepsy: diagnosis, classification and management. *Medicine* **48,** 522–528. ISSN: 1357-3039. https://www.sciencedirect.com/science/article/pii/S13573039203011 6X (2025) (Aug. 1, 2020).
14. Beach, R. & Reading, R. The importance of acknowledging clinical uncertainty in the diagnosis of epilepsy and non-epileptic events. *Archives of Disease in Childhood* **90,** 1219–1222. ISSN: 1468-2044 (Dec. 2005).
15. Hauser, W. A., Rich, S. S., Annegers, J. F. & Anderson, V. E. Seizure recurrence after a 1st unprovoked seizure. *Neurology* **40.** Publisher: Wolters Kluwer, 1163–1163. https://www.neurology.org/doi/10.1212/WNL.40.8.1163 (2025) (Aug. 1990).

16. Neligan, A. *et al.* Prognosis of adults and children following a first unprovoked seizure - Neligan, A - 2023 | Cochrane Library. ISSN: 1465-1858. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD013847.pub2/full (2025).
17. Thangavel, P. *et al.* Improving automated diagnosis of epilepsy from EEGs beyond IEDs. *Journal of neural engineering* **19,** 10.1088/1741–2552/ac9c93. ISSN: 1741-2560. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11549972/ (2025) (Nov. 24, 2022).
18. Mirwani, Y. M. Automated Epilepsy Diagnosis beyond IEDs by Multimodal Features and Deep Learning. https://repository.tudelft.nl/record/uuid:c829feac-3482-47a3-9c3e-2e27e89056c0 (2025) (2024).
19. Fisher, R. S. *et al.* Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **58,** 522–530. ISSN: 0013-9580, 1528-1167. https://onlinelibrary.wiley.com/doi/10.1111/epi.13670 (2025) (Apr. 2017).
20. Baker, G. A., Gagnon, D. & McNulty, P. The relationship between seizure frequency, seizure type and quality of life: Findings from three European countries. *Epilepsy Research* **30,** 231–240. ISSN: 0920-1211. https://www.sciencedirect.com/science/article/pii/S0920121198000102 (2025) (May 1, 1998).
21. *Tonic-clonic (grand mal) seizure - Symptoms and causes* Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/grand-mal-seizure/symptoms-causes/syc-20363458 (2025).
22. Ghulaxe, Y. *et al.* Understanding Focal Seizures in Adults: A Comprehensive Review. *Cureus* **15,** e48173. ISSN: 2168-8184. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10693312/ (2025).
23. Ighodaro, E. T., Maini, K., Arya, K. & Sharma, S. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024). http://www.ncbi.nlm.nih.gov/books/NBK500005/ (2025).
24. Moosa, A. N. V. & Wyllie, E. Focal epileptogenic lesions. *Handbook of Clinical Neurology* **111,** 493–510. ISSN: 0072-9752 (2013).
25. Van Diessen, E., Diederen, S. J. H., Braun, K. P. J., Jansen, F. E. & Stam, C. J. Functional and structural brain networks in epilepsy: What have we learned? *Epilepsia* **54.** _eprint: https://onlinelibrary.wiley.com/do 1855–1865. ISSN: 1528-1167. https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.12350 (2025) (2013).
26. Chang, B. S. Cortical Hyperexcitability: A New Biomarker in Generalized Epilepsy Syndromes. *Epilepsy Currents* **13,** 287–288. ISSN: 1535-7597. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3854747/ (2025) (2013).
27. Burman, R. J. & Parrish, R. R. The Widespread Network Effects of Focal Epilepsy. *Journal of Neuroscience* **38.** Publisher: Society for Neuroscience Section: Journal Club, 8107–8109. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/38/38/8107 (2025) (Sept. 19, 2018).
28. *Prevention* https://www.who.int/teams/mental-health-and-substance-use/treatment-care/mental-health-gap-action-programme/evidence-centre/epilepsy-and-seizures/antiseizure-medicines-for-management-of-epilepsy-in-adults-and-children (2025).
29. Fishman, J. *et al.* Patient emotions and perceptions of antiepileptic drug changes and titration during treatment for epilepsy. *Epilepsy & Behavior* **69.** Publisher: Elsevier, 44–52. ISSN: 1525-5050, 1525-5069. https://www.epilepsybehavior.com/article/S1525-5050%2816%2930677-1/fulltext?utm_source=chatgpt.com (2025) (Apr. 1, 2017).
30. Löscher, W. & Klein, P. The Pharmacology and Clinical Efficacy of Antiseizure Medications: From Bromide Salts to Cenobamate and Beyond. *CNS Drugs* **35,** 935–963. ISSN: 1172-7047. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8408078/ (2025) (2021).
31. *The 10-20 System for EEG* https://info.tmsi.com/blog/the-10-20-system-for-eeg (2025).
32. Hatton, S. L., Rathore, S., Vilinsky, I. & Stowasser, A. Quantitative and Qualitative Representation of Introductory and Advanced EEG Concepts: An Exploration of Different EEG Setups. *Journal of Undergraduate Neuroscience Education* **21,** A142–A150. ISSN: 1544-2896. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10426816/ (2025) (May 19, 2023).

33. Jin, J. E. *et al.* Soft, adhesive and conductive composite for electroencephalogram signal quality improvement. *Biomedical Engineering Letters* **13,** 495–504. ISSN: 2093-9868. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10382389/ (2025) (Apr. 13, 2023).
34. Melnik, A. *et al.* Systems, Subjects, Sessions: To What Extent Do These Factors Influence EEG Data? *Frontiers in Human Neuroscience* **11,** 150. ISSN: 1662-5161. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5371608/ (2025) (Mar. 30, 2017).
35. Thio, B. J., Aberra, A. S., Dessert, G. E. & Grill, W. M. Ideal current dipoles are appropriate source representations for simulating neurons for intracranial recordings. *Clinical Neurophysiology* **145,** 26–35. ISSN: 1388-2457. https://www.sciencedirect.com/science/article/pii/S1388245722009294 (2025) (Jan. 1, 2023).
36. Michel, C. M. & Brunet, D. EEG Source Imaging: A Practical Review of the Analysis Steps. *Frontiers in Neurology* **10.** Publisher: Frontiers. ISSN: 1664-2295. https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2019.00325/full (2025) (Apr. 4, 2019).
37. Eom, T.-H. Electroencephalography source localization. *Clinical and Experimental Pediatrics* **66,** 201–209. ISSN: 2713-4148. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10167408/ (2025) (Dec. 29, 2022).
38. Daly, I. *et al.* Electroencephalography reflects the activity of sub-cortical brain regions during approach-withdrawal behaviour while listening to music. *Scientific Reports* **9.** Publisher: Nature Publishing Group, 9415. ISSN: 2045-2322. https://www.nature.com/articles/s41598-019-45105-2 (2025) (July 1, 2019).
39. Van Beest, E. H. *et al.* The direct and indirect pathways of the basal ganglia antagonistically influence cortical activity and perceptual decisions. *iScience* **27,** 110753. ISSN: 2589-0042. https://www.sciencedirect.com/science/article/pii/S2589004224019783 (2025) (Sept. 20, 2024).
40. Puspita, J. W., Soemarno, G., Jaya, A. I. & Soewono, E. Interictal Epileptiform Discharges (IEDs) classification in EEG data of epilepsy patients. *Journal of Physics: Conference Series* **943.** Publisher: IOP Publishing, 012030. ISSN: 1742-6596. https://dx.doi.org/10.1088/1742-6596/943/1/012030 (2025) (Dec. 2017).
41. Geut, I., Weenink, S., Knottnerus, I. L. H. & Putten, M. J. A. M. v. Detecting interictal discharges in first seizure patients: ambulatory EEG or EEG after sleep deprivation? *Seizure - European Journal of Epilepsy* **51.** Publisher: Elsevier, 52–54. ISSN: 1059-1311, 1532-2688. https://www.seizure-journal.com/article/S1059-1311%2817%2930489-2/fulltext?utm_source=chatgpt.com (2025) (Oct. 1, 2017).
42. *Spike slow wave complex - EEGpedia* http://www.eegpedia.org/index.php?title=Spike_slow_wave_complex (2025).
43. Abdulrahman, A. *Introduction to Machine learning for beginners(PART I)* Analytics Vidhya. https://medium.com/analytics-vidhya/introduction-to-machine-learning-for-beginners-part-i-1df752d461cf (2025).
44. Brownlee, J. *4 Types of Classification Tasks in Machine Learning* MachineLearningMastery.com. https://www.machinelearningmastery.com/types-of-classification-in-machine-learning/ (2025).
45. Vasques, X. in *Machine Learning Theory and Applications: Hands-on Use Cases with Python on Classical and Quantum Machines* Conference Name: Machine Learning Theory and Applications: Hands-on Use Cases with Python on Classical and Quantum Machines, 35–174 (Wiley, 2024). ISBN: 978-1-394-22063-2. https://ieeexplore.ieee.org/document/10444109 (2025).
46. DelSole, M. *What is One Hot Encoding and How to Do It* Medium. https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179 (2025).
47. *Encoding Categorical Data with One-hot Encoding* Paperspace by DigitalOcean Blog. https://blog.paperspace.com/encoding-categorical-data-with-one-hot-encoding/ (2025).
48. *XGBoost Documentation — xgboost 2.1.1 documentation* https://xgboost.readthedocs.io/en/stable/ (2025).

49. Anshul. *What is Decision Tree? [A Step-by-Step Guide]* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/ (2025).
50. Sonawane, P. *XGBoost — How does this work* Medium. https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb (2025).
51. *Difference Between Random Forest and XGBoost - GeeksforGeeks* https://www.geeksforgeeks.org/difference-between-random-forest-vs-xgboost/ (2025).
52. Narkhede, S. *Understanding AUC - ROC Curve* Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (2025).
53. Brownlee, J. *LOOCV for Evaluating Machine Learning Algorithms* MachineLearningMastery.com. https://www.machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/ (2025).
54. *Our story* anDREa-cloud. https://andrea-cloud.com/our-story/ (2025).
55. Kotsopoulos, I. A. W., van Merode, T., Kessels, F. G. H., de Krom, M. C. T. F. M. & Knottnerus, J. A. Systematic review and meta-analysis of incidence studies of epilepsy and unprovoked seizures. *Epilepsia* **43,** 1402–1409. ISSN: 0013-9580 (Nov. 2002).
56. Beghi, E. & Giussani, G. Aging and the Epidemiology of Epilepsy. *Neuroepidemiology* **51,** 216–223. ISSN: 1423-0208 (2018).
57. Nayak, C. S. & Anilkumar, A. C. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024). http://www.ncbi.nlm.nih.gov/books/NBK537023/ (2025).
58. *dbc-uitgelegd* https://www.zilverenkruis.nl/consumenten/vergoedingen/dbc-uitgelegd (2025).
59. *ATCDDD - ATC/DDD Index* https://atcddd.fhi.no/atc_ddd_index/ (2025).
60. *chi2_contingency — SciPy v1.15.0 Manual* https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html (2025).
61. *Chi-Square Test of Independence* Statistics Solutions. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/chi-square/ (2025).
62. Yates, F. Contingency Tables Involving Small Numbers and the χ2 Test. *Supplement to the Journal of the Royal Statistical Society* **1.** Publisher: [Oxford University Press, Royal Statistical Society], 217–235. ISSN: 1466-6162. https://www.jstor.org/stable/2983604 (2025) (1934).
63. Patel, V. *The Impact of Local Geometry and Batch Size on Stochastic Gradient Descent for Non-convex Problems* May 5, 2018. arXiv: 1709.04718[math]. http://arxiv.org/abs/1709.04718 (2025).
64. Sabuncu, M. R. *Intelligence plays dice: Stochasticity is essential for machine learning* Aug. 17, 2020. arXiv: 2008.07496[cs]. http://arxiv.org/abs/2008.07496 (2025).
65. Kirmer, S. *The Meaning of Explainability for AI* Medium. https://towardsdatascience.com/the-meaning-of-explainability-for-ai-d8ae809c97fa (2025).
66. Marsh, E. K. *XGBoost Feature Importance* Medium. https://medium.com/@emilykmarsh/xgboost-feature-importance-233ee27c33a4 (2025).
67. Abu-Rmileh, A. *Be careful when interpreting your features importance in XGBoost!* Medium. https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7 (2025).
68. *An introduction to explainable AI with Shapley values — SHAP latest documentation* https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html (2025).

69. *An Introduction to SHAP Values and Machine Learning Interpretability* https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720821&utm_adgroupid=157156374951&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=684592139651&utm_targetid=dsa-2218886984380&utm_loc_interest_ms=&utm_loc_physical_ms=9064564&utm_content=&utm_campaign=230119_1-sea~dsa~tofu_2-b2c_3-row-p1_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na&gad_source=1&gclid=CjwKCAiA-Oi7BhA1EiwA2rIu23MSAvy8c5OKnCwqlZhihCeBXJRjN9icZTbogAHsZsnPmNdBpUaljBoCaVQQAvD_BwE (2025).

70. Bobbitt, Z. *Welch's t-test: When to Use it + Examples* Statology. https://www.statology.org/welchs-t-test/ (2025).

71. Sakai, T. in *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power* (ed Sakai, T.) 27–41 (Springer, Singapore, 2018). ISBN: 9789811311994. https://doi.org/10.1007/978-981-13-1199-4_2 (2025).

72. Brownlee, J. *A Gentle Introduction to Ensemble Learning Algorithms* MachineLearningMastery.com. https://www.machinelearningmastery.com/tour-of-ensemble-learning-algorithms/ (2025).

73. Schulman, P. Bayes' Theorem—A Review. *Cardiology Clinics* **2.** Publisher: Elsevier, 319–328. ISSN: 0733-8651, 1558-2264. https://www.cardiology.theclinics.com/article/S0733-8651(18)30726-4/abstract?utm_source=chatgpt.com (2025) (Aug. 1, 1984).

74. Clyde, M. *et al. Chapter 1 The Basics of Bayesian Statistics | An Introduction to Bayesian Thinking* https://statswithr.github.io/book/the-basics-of-bayesian-statistics.html (2025) ().

75. *Law of Total Probability | Partitions | Formulas* https://www.probabilitycourse.com/chapter1/1_4_2_total_probability.php (2025).

76. Van Diessen, E. *et al.* A Prediction Model to Determine Childhood Epilepsy After 1 or More Paroxysmal Events. *Pediatrics* **142,** e20180931. ISSN: 0031-4005, 1098-4275. https://publications.aap.org/pediatrics/article/142/6/e20180931/37514/A-Prediction-Model-to-Determine-Childhood-Epilepsy (2025) (Dec. 1, 2018).

77. Vergara López, S. *et al.* Epilepsy diagnosis based on one unprovoked seizure and ≥60% risk. A systematic review of the etiologies. *Epilepsy & Behavior* **125,** 108392. ISSN: 1525-5050. https://www.sciencedirect.com/science/article/pii/S1525505021006533 (2025) (Dec. 1, 2021).

78. Beaulieu-Jones, B. K. *et al.* Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. *The Lancet Digital Health* **5.** Publisher: Elsevier, e882–e894. ISSN: 2589-7500. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00179-6/fulltext (2025) (Dec. 1, 2023).

79. Polotskaya, K. *et al.* Bayesian Networks for the Diagnosis and Prognosis of Diseases: A Scoping Review. *Machine Learning and Knowledge Extraction* **6.** Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 1243–1262. ISSN: 2504-4990. https://www.mdpi.com/2504-4990/6/2/58 (2025) (June 2024).

80. Han, L. & Zhang, Y. *Epileptic seizure prediction based on dynamic Bayesian networks* in *Second International Conference on Biomedical and Intelligent Systems (IC-BIS 2023)* Second International Conference on Biomedical and Intelligent Systems (IC-BIS 2023). **12724** (SPIE, Aug. 28, 2023), 256–260. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12724/1272413/Epileptic-seizure-prediction-based-on-dynamic-Bayesian-networks/10.1117/12.2687813.full (2025).

81. Wang, J., Ning, X., Shi, W. & Lin, Y. *A Bayesian Graph Neural Network for EEG Classification — A Win-Win on Performance and Interpretability* in *2023 IEEE 39th International Conference on Data Engineering (ICDE)* 2023 IEEE 39th International Conference on Data Engineering (ICDE). ISSN: 2375-026X (Apr. 2023), 2126–2139. https://ieeexplore.ieee.org/document/10184782 (2025).

A

Appendix A: Feature set details

A.1. Hyperparameters of EEG benchmark feature sets

Table A.1: Overview of EEG feature set hyperparameters from previous research, and after patient exclusions and fixing UTM featureset

(a) Overview of initial EEG feature set hyperparameters from previous research

EEG feature set	Montage	Combiner	Epoch segment length (s)
CC	BipolarDB	Skewness	300
CWT	BipolarDB	Median	60
DWT	Cz	Median	60
GCC	BipolarDB	Median	2
GPLV	Cz	Skewness	10
mST	Laplacian	Median	2
PLV	BipolarDB	Skewness	300
sST	CAR	Mean	20
S	CAR	Median	20
UTM	CAR	Mean	20

(b) Overview of EEG feature set hyperparameters after patient exclusions and fixing UTM featureset

EEG feature set	Montage	Combiner	Epoch segment length (s)
CC	Laplacian	Mean	2
CWT	BipolarDB	Skewness	60
DWT	BipolarDB	Mean	10
GCC	Cz	Kurtosis	10
GPLV	Laplacian	Kurtosis	120
mST	Cz	Median	120
PLV	Laplacian	Mean	2
sST	CAR	Median	300
S	Cz	Skewness	5
UTM	CAR	Skewness	10

BipolarDB = Bipolar Double Banana, CAR = Common Average Reference, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, Cz = Cz referenced montage, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measure of Phase Lock Values, mST = mean Stockwell Transform, PLV = Phase Lock Values, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures

A.2. Medical history encoding categories

Table A.2: Overview of medical history feature categories

Abbreviation	Category explanation
CAR	Cardiovascular disease
FAM	Family history of epilepsy
MDL	Gastro-intestinal or liver disease
MID	Substance abuse
NEU	Neurological disorders
ONT	Developmental disorder
PSY	Psychological history
RESP	Respiratory disease
SEI	Earlier seizure or epilepsy, but no longer relevant according to ILAE definition
STRU	Structural brain abnormalities
TRAU	Traumatic injury

ILAE = International League Against Epilepsy

A.3. Medication encoding groups

Table A.3: Overview of medication feature groups in this research. The first medication encoding consisted of 1st order ATC codes. Their corresponding anatomical groups are shown. The second medication encoding consisted 2nd and 3rd order ATC codes (therapeutic and pharmacological subgroups resp.). An overview of included medication subgroups are shown in the table, their specific subgroup can be referenced at ATC/DDD Index

ATC 1st order	Medication anatomical group	ATC 2nd-3rd order
A	Alimentary tract and metabolism	01, 02, 03, 04, 06, 07, 10, 11, 12
B	Blood and blood forming organs	01, 02, 05
C	Cardiovascular system	01, 02, 03, 07, 08, 09, 10
D	Dermatologicals	01, 02, 06, 07
G	Genito urinary system and sex hormones	02, 03, 04
H	Systemic hormonal preparations, excl. sex hormones and insulins	01, 02, 03, 05
J	Antiinfectives for systemic use	01, 05
L	Antineoplastic and immunomodulating agents	04
M	Musculo-skeletal system	01, 03, 04
N	Nervous sytem	01A, 01B, 02A, 02B, 02C, 03A, 05A, 05B, 05C, 06A, 07B
P	Antiparasistic products, insecticides and repellants	01
R	Respiratory system	01, 03, 05
S	Sensory organs	01, 02
V	Various	03, 06

ATC/DDD = Anatomical Therapeutic Chemical classification system with Defined Daily Dose

A.4. Individual EEG rhythm encoding

Table A.4: An overview of included individual EEG rhythms and electrographical findings. Rhythms or electrographical findings that were not found in the subject population of this research are not shown.

Abbreviation	EEG rhythm or electrographical finding
PDR	Posterior dominant rhythm
alfa	Alfa activity (not PDR) $\simeq 8-13\,Hz$
beta	Beta activity $\simeq 13-30\,Hz$
centr	Central rhythm
ft_FIRDA	Frontal Intermittent Rhythmic Delta Activity
ft_delt	Focal delta activity $\leq 4\,Hz$
ft_trag	Focal slow activity
lamb	Lambda waves
mu	Mu rhythm
per_delt	Periodic delta activity $\leq 4\,Hz$
per_trag	Periodic slow activity
sch_golf	Sharp wave
sch_tr_golf	Sharp and slow wave complex
sch_trans	Sharp transient
thet	Theta activity $\simeq 4-8\,Hz$
trag	(Other) slow activity

EEG = ElectroEncephaloGram

A.5. Overall EEG background encoding

Table A.5: An overview of background EEG encoding groups. Subjects could have multiple groups present, but if diffuse or focal aberrances were found, the EEG was not deemed 'Normal'.

EEG Background category	Explanation
Normal	No EEG abnormalities documented
Diffuse_fast	Diffuse excess of fast activity $> 13\,Hz$
Diffuse_sharp	Sharp transients/waves occurring diffusely
Diffuse_slow	Diffuse excess slow activity $< 8\,Hz$
Focal_fast	Focal excess of fast activity $> 13\,Hz$
Focal_sharp	Sharp transients/waves with clear focal origin
Focal_slow	Focal excess of slow activity $< 8\,Hz$

EEG = ElectroEncephaloGram

B

Appendix B

B.1. Relative category occurrence of individual EEG rhythms in healthy and epileptic groups

Table B.1: Relative occurence of individual EEG rhythms in healthy and epileptic group, with χ^2 results indicated with Yates-p values. Rows with a significant difference between healthy and epileptic groups are in bold font and have an asterisk at the correponding p-value.

Category	Healthy+ (%)	Epilepsy+ (%)	Yates-p
PDR	104 (100%)	39 (100%)	1.0
alfa	1 (0.96%)	1 (2.56%)	1.0
beta	44 (42.31%)	17 (43.59%)	1.0
centr	9 (8.65%)	1 (2.56%)	0.3662
ft_FIRDA	1 (0.96%)	0 (0%)	1.0
ft_delt	**9 (8.65%)**	**10 (25.64%)**	**0.0170***
ft_trag	9 (8.65%)	5 (12.82%)	0.6666
lamb	16 (15.38%)	8 (20.51%)	0.6315
mu	50 (48.08%)	19 (48.72%)	1.0
per_delt	0 (0%)	1 (2.56%)	0.6086
per_trag	3 (2.88%)	0 (0%)	0.6768
sch_golf	16 (15.38%)	11 (28.21%)	0.1324
sch_tr_golf	**1 (0.96%)**	**4 (10.26%)**	**0.0290***
sch_trans	9 (8.65%)	5 (12.82%)	0.6666
thet	7 (6.73%)	3 (7.69%)	1.0
trag	2 (1.92%)	0 (0%)	0.9421

PDR = Posterior dominant rhythm, alfa = alfa activity, beta = beta activity, centr = central rhythm, ft_FIRDA = Frontal Intermittent Rhythmic Delta Activity, ft_delt = focal delta activity, ft_trag = focal slow activity, lamb = lambda waves, mu = mu rhythm, per_delt = periodic delta activity, per_trag = Periodic slow activity, sch_golf = sharp wave, sch_tr_golf = sharp and slow wave complex, sch_trans = sharp transient, thet = theta activity, trag = (other) slow activity

B.2. Co-occurrence of medication grouped on 1st order ATC codes

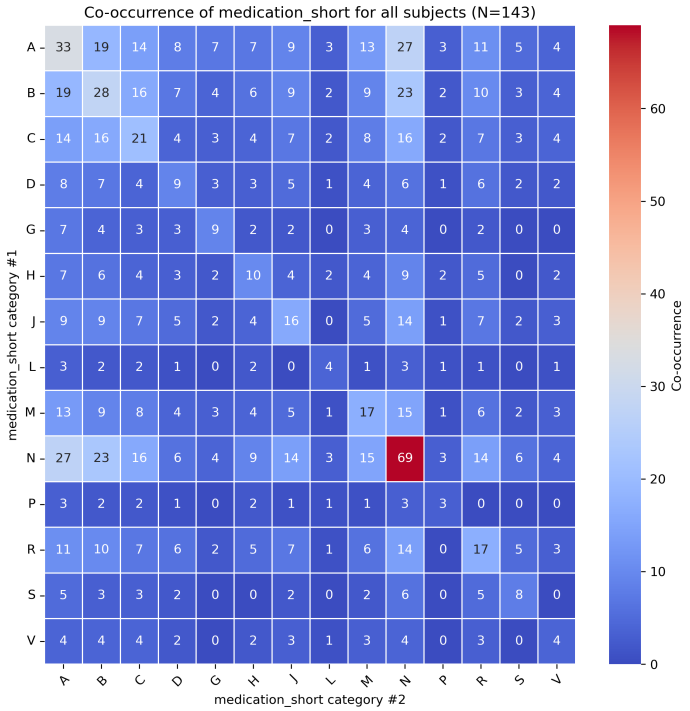


Figure B.1: Co-occurrence of categories within medication grouped on 1st order ATC codes. The figure shows the amount of subjects that present with both the medication group on the y-axis as well as the x-axis.
[ATC = Anatomical Therapeutic Chemical classification system, A = Alimentary tract and metabolism, B = Blood and blood forming organs, C = Cardiovascular system, D = Dermatologicals, G = Genito urinary system and sex hormones, H = Systemic hormonal preparations, excl. sex hormones and insulins, J = Antiinfectives for systemic use, L = Antineoplastic and immunomodulating agents, M = Musculo-skeletal system, N = Nervous sytem, P = Antiparasistic products, insecticides and repellants, R = Respiratory system, S = Sensory organs, V = Various]

C
Appendix C

C.1. AUCs of XGBoost models from EEG feature sets after adding a single clinical or report features

Full page table, see next page.

Table C.1: An overview over XGBoost model performances when adding a single clinical or report feature to the different EEG benchmark feature sets. Significant differences are indicated in bold font and include an asterisk.

EEG featureset	Bench.	+ Age	+ Sex	+ Vig.	+ His.	+ Med.1	+ Med.2/3	+ R.ind.	R.all.
CC	0.640 ± 0.025	0.648 ± 0.025	0.638 ± 0.031	0.624 ± 0.025	0.622 ± 0.021	0.630 ± 0.025	0.635 ± 0.017	0.623 ± 0.023	0.604 ± 0.021
CWT	0.669 ± 0.021	0.658 ± 0.017	0.662 ± 0.022	0.664 ± 0.032	0.660 ± 0.030	0.656 ± 0.028	0.653 ± 0.020	0.662 ± 0.020	0.656 ± 0.023
DWT	0.688 ± 0.015	0.681 ± 0.015	0.687 ± 0.011	0.680 ± 0.016	0.679 ± 0.023	0.672 ± 0.015	0.687 ± 0.022	0.669 ± 0.021	0.678 ± 0.018
GCC	0.707 ± 0.020	0.706 ± 0.022	0.702 ± 0.019	0.710 ± 0.027	0.710 ± 0.024	0.699 ± 0.021	0.706 ± 0.017	0.658 ± 0.023	0.693 ± 0.021
GPLV	0.654 ± 0.018	0.635 ± 0.016	0.645 ± 0.017	0.669 ± 0.012	0.662 ± 0.020	0.653 ± 0.024	0.659 ± 0.012	0.649 ± 0.014	**0.697 ± 0.014***
mST	0.714 ± 0.016	0.712 ± 0.009	0.717 ± 0.015	0.717 ± 0.029	0.724 ± 0.024	0.704 ± 0.018	0.693 ± 0.020	0.712 ± 0.023	0.738 ± 0.022
PLV	0.622 ± 0.038	0.609 ± 0.026	0.611 ± 0.033	0.619 ± 0.018	0.631 ± 0.032	0.608 ± 0.030	0.628 ± 0.023	0.632 ± 0.041	0.614 ± 0.034
S	0.688 ± 0.014	0.679 ± 0.020	0.682 ± 0.019	0.678 ± 0.018	0.675 ± 0.025	0.671 ± 0.020	0.687 ± 0.015	0.689 ± 0.021	0.692 ± 0.024
sST	0.675 ± 0.019	0.681 ± 0.025	0.681 ± 0.022	0.673 ± 0.020	0.667 ± 0.014	0.679 ± 0.019	0.671 ± 0.012	0.671 ± 0.019	0.662 ± 0.022
UTM	0.630 ± 0.023	0.649 ± 0.030	0.647 ± 0.025	0.640 ± 0.029	0.644 ± 0.030	0.623 ± 0.022	0.636 ± 0.021	0.647 ± 0.017	0.640 ± 0.020

***bold font and** * indicatie significant increase compared to benchmark, CC = Cross-Correlation, CWT = Continuous Wavelet Transform, DWT = Discrete Wavelet Transform, GCC = Graph measures of Cross-Correlation, GPLV = Graph measures of Phase Lock Values, His. = Medical history, Med.1 = Medication use/history based on 1st order ATC codes, Med.2/3 = Medication use/history based on 2nd and 3rd order ATC codes, mST = mean Stockwell Transform, PLV = Phase Lock Values, R.all = EEG background, R.ind. = Individual EEG rhythms and electrographical findings, sST = square root Stockwell Transform, S = Spectral, UTM = Univariate Time Measures, Vig. = Vigilance State*

D

Appendix D

D.1. SHAP plots for XGBoost model of DWT feature set including clinical and report features

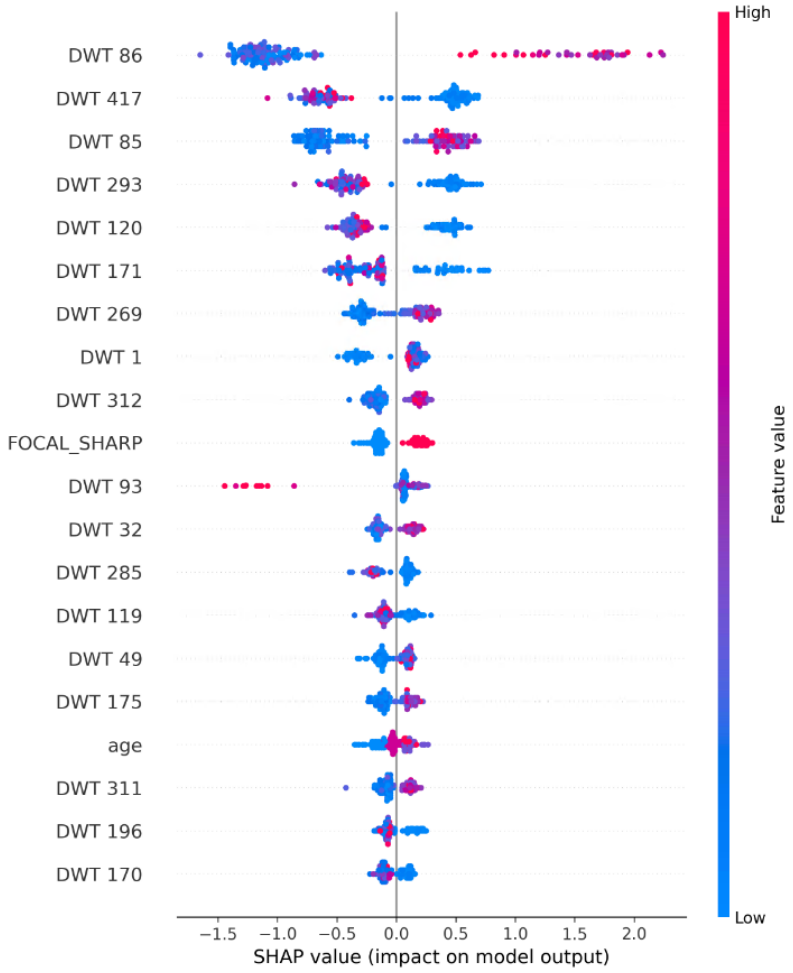


Figure D.1: SHAP summary plot for the top performing features in the XGBoost model of DWT feature set including clinical and report features. Each dot represents a subject in the test set, with its colour representing the height of feature value in that subject. The EEG background was split in its categories for evaluating the SHAP values; red corresponds to 1 (category is present in subject), blue corresponds to 0 (category is not present in subject). A positive SHAP value indicates the feature shifts model prediction for that subject towards outcome label '1'/'epilepsy'. A negative SHAP value indicates the feature shifts model prediction towards outcome label '0'/'healthy'.
[DWT = Discrete Wavelet Transform, EEG = ElectroEncephaloGram, SHAP = SHapley Additive eXplanations, XGB(oost) = eXtreme Gradient Boosting]

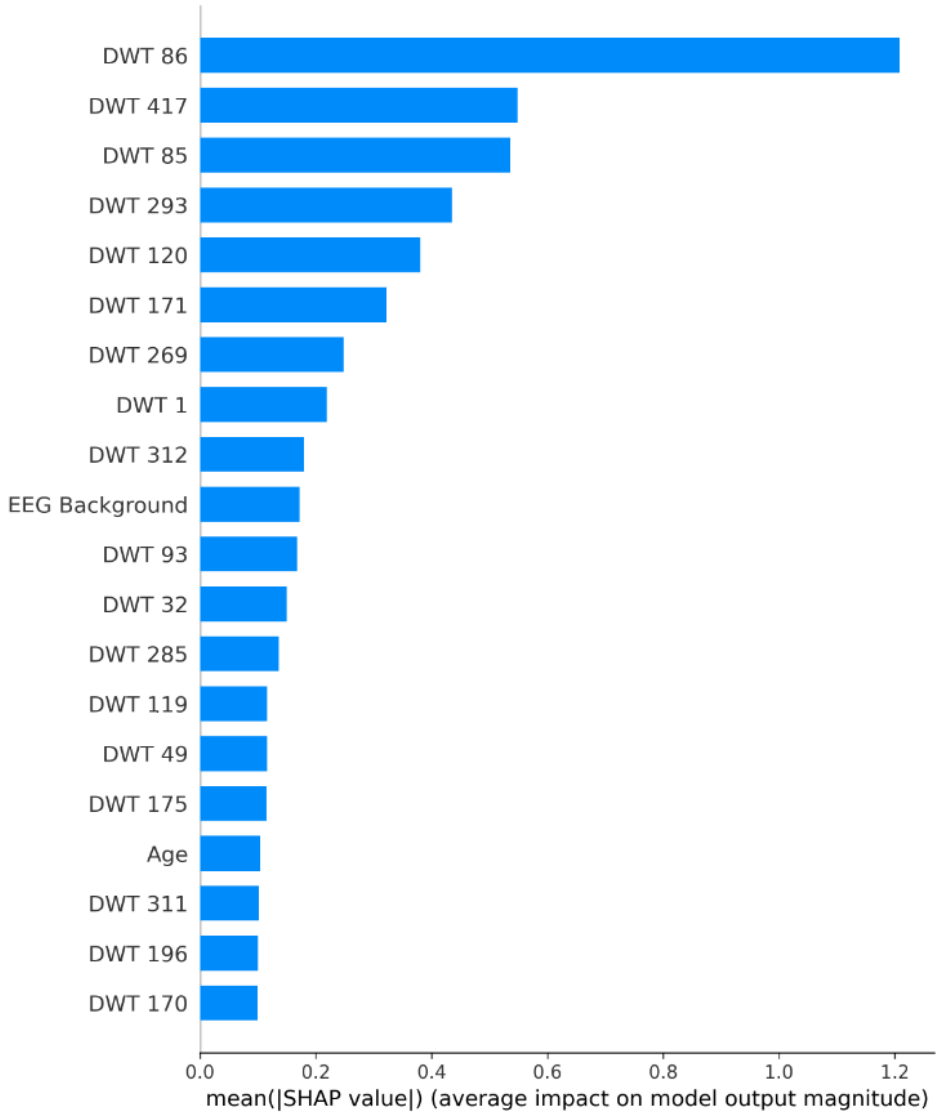


Figure D.2: SHAP bar plot for the XGBoost model of DWT feature set including EEG background. The mean absolute value shows the average impact of a feature on the model output. The absolute SHAP value was taken for EEG background overall instead of its individual categories.
[DWT = Discrete Wavelet Transform, EEG = ElectroEncephaloGram, SHAP = SHapley Additive eXplanations, XGB(oost) = eXtreme Gradient Boosting]