University of Warsaw
Faculty of Physics

Kacper Chorzela
Record book number: 450224

Investigating Site Effects in Multi-Center EEG Data for Neuroscreening: A Machine Learning Approach

Bachelor's thesis
in the field of physics within the MISMaP

The thesis was writen under the supervision of
dr hab. Jarosław Żygierewicz, prof. UW
Biomedical Physics Division
Institute of Experimental Physics

Warsaw, September 2025

Abstract

The site effect—systematic, non-biological variance between medical centers—is a significant challenge when combining multi-center EEG data for machine learning. This thesis aimed to quantify, interpret, and reduce the site effect in the large ELM19 dataset and evaluate its impact on a clinical neuroscreening task. A CatBoost classifier was trained to identify the source hospital from EEG features, and SHAP analysis was used to interpret the model. The classifier identified the source hospital with high accuracy (MCC = 0.865). Surprisingly, harmonization using ComBat and site-wise standardization reduced feature-level differences but appeared to make the source hospitals even easier to classify (MCC > 0.98). Despite this, harmonization offered a slight improvement in the clinical task. To sum up, the site effect is a complex, multivariate challenge that standard methods cannot fully resolve. This highlights the need for more advanced approaches to develop truly generalizable neuroscreening models.

Keywords

Site Effect, EEG, Machine Learning (ML), SHAP (Model Explainability), Data Harmonization

Title of the thesis in Polish language

Analiza wpływu ośrodka badawczego na wieloośrodkowe dane EEG w neuroscreeningu: podejście oparte na uczeniu maszynowym

Contents

Chapter 1

Introduction

1.1. Background of the study

There is a great potential for improving and extending current diagnostic methods in the field of neurology using machine learning. An example of such an application is the analysis of data from electroencephalography (EEG), a technique used to record the brain's electrical activity. EEG is valued in both research and clinical practice because it offers high temporal resolution, is non-invasive, and is widely accessible. However, since the signal is highly complex to analyze, building models capable of capturing complex patterns, such as those indicating abnormalities, requires a large amount of data. This need can be met by combining data from multiple hospitals to create an extensive, multi-center EEG database.

Unfortunately, there is a significant challenge when working with multi-center datasets—the site effect. This term refers to systematic, non-biological differences in the data that come from inter-center variability. These differences may arise from a variety of sources, such as differences in EEG equipment, software, recording procedures, or patient demographics. The problem is that machine learning models can learn these patterns from the data, which leads to a significant generalization issue. As a result, when these models face data from institutions they weren't trained on, their performance can drop.

That's why it is important to be aware of these challenges when working with multi-center data. First, the problem should be analyzed to identify what causes the differences. Second, different methods can be applied to reduce their impact.

1.2. Thesis aims and research questions

The previously mentioned issue of generalization was observed within the GBE model, a gradient-boosted ensemble of 30 CatBoost classifiers, on the ELM19 dataset, a large database of EEG recordings [Poziomska et al., 2024]. The main aims of this thesis are related to an investigation of the site effect present in the ELM19 dataset and are guided by the following questions:

1. **Quantification:** Can machine learning models detect patterns caused by the site effect?
2. **Interpretation:** Which EEG features contribute most to the observed site effect?
3. **Reduction:** What methods can reduce the site effect from data?
4. **Validation:** How does the reduction of site effect impact model performance and generalization on a clinical prediction task?

1.3. Scope of the study

This study is focused on analyzing the ELM19 dataset provided by Elmiko Biosignals sp. z o.o.[1], containing approximately 55,000 EEG recordings from 30 different medical institutions. To explore the dataset more deeply, a CatBoost[2] classifier was used to identify the source of the recordings from a set of features extracted from each recording. The tool used for interpreting the model's decisions was SHAP[3]. An attempt to reduce the site effect was made by performing data harmonization methods, including site-wise standardization and neuroCombat[4], implementation of the ComBat algorithm.

[1] https://www.elmiko.pl
[2] https://catboost.ai/
[3] https://shap.readthedocs.io/
[4] https://github.com/Jfortin1/neuroCombat

Chapter 2

Theoretical background

2.1. Electroencephalography (EEG)

Electroencephalography (EEG) is a method for recording the electrical activity of the brain. It is performed by placing electrodes on the scalp, which makes it a portable and cost-effective tool compared to other techniques such as magnetoencephalography (MEG). EEG is also very safe because it is a non-invasive procedure. What's more, unlike imaging methods that use ionizing radiation, EEG does not bring risks with repeated measurements. Its high temporal resolution, on the order of milliseconds, allows EEG to capture rapid changes in brain activity. All of these characteristics have made the EEG a crucial tool for various neurological diagnoses [Lopes da Silva and Schomer, 2018].

EEG measures the summed electrical activity from millions of neurons, mainly the ones oriented perpendicularly to the cortical surface, called pyramidal neurons. These activities originate from the postsynaptic potentials of large populations of these neurons, which synchronously fire, generating an extracellular field detectable at the scalp [Lopes da Silva and Schomer, 2018].

2.1.1. Spectral analysis

The standard way to analyze EEG is in the frequency domain, where brain activity is examined corresponding to the frequency of its oscillations. The most commonly used approach is to divide the EEG activity into bands such as delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz). Each band is associated with a different mental and cognitive state, such as delta with deep sleep, theta with memory and drowsiness, alpha with relaxed wakefulness, and beta with active concentration [Teplan, 2002].

While these bands correspond to specific frequency ranges, this is mainly a convention rather than a strict rule. These frequencies could vary significantly with factors such as age, mental state, or cognitive engagement [Doppelmayr et al., 1998]. Because of that, the EEG signal in this work was analyzed using a custom set of overlapping frequency bands: [0.5–2], [1–3], [2–4], [3–6], [4–8], [6–10], [8–13], [10–15], [13–18], [15–21], [18–24], [21–27], [24–30], [27–40] Hz.

2.1.2. Normal and pathological EEG

The interpretation of the EGG often starts with an assessment of the background. A healthy awake EEG signal is characterized by a low-amplitude mix of different frequency background rhythms. The key rhythm observed with eyes closed is the posterior dominant rhythm (PDR),

which is a steady alpha wave originating from the occipital region of the brain. One of the key features of a normal EEG is that it is highly symmetrical between hemispheres in both frequency and amplitude, and any significant differences between the two are typically considered abnormal. Another essential component of a normal EEG is the beta activity, which is usually observed over the frontal and central areas of the brain [St. Louis et al., 2016].

On top of that, another characteristic of background that might show brain dysfunctions is the presence of background slowing, characterized by excessively slow waves in the theta or delta range. This slowing can be either focal (in one specific area) or generalized (across the whole brain). While some degree of slowing is a normal observation, for example, in children or during sleep, consistently seen focal slowing over one head region, or monotonous generalized slow wave activity in an awake person, should be considered pathological [St. Louis et al., 2016].

However, the diagnosis should not be based only on the background EEG activity. For example, in epilepsy, the background EEG is often normal without any abnormalities [Salinsky et al., 1987]. The main indicators of epilepsy risk typically appear as short, abnormal discharges called interictal epileptiform discharges (IEDs), such as spikes and sharp waves. These, along with other EEG features, are described using standardized terminology based on the SCORE system [Beniczky et al., 2017].

2.1.3. Sources of signal artifact

Unfortunately, the EEG signals are often affected by artifacts—undesired signals that appear in the recording. This noise comes from factors such as eye movements, blinks, muscle activity, cardiac activity, breathing, or other external and environmental factors, like powerline interference. In real-world EEG recordings, artifacts are difficult to avoid; however, their impact can be significantly reduced through artifact prevention strategies. These may include automatic removal with software, proper patient preparation, and the skill and technique of the person collecting the signal [Zhang et al., 2022, Croft et al., 2005]. However, different medical institutions may employ various methods to handle artifacts—some more effective than others—which can lead to differences in the resulting EEG signals.

2.2. The site effect in multi-center research

The site effect refers to systematic, non-biological variance in data that arises from differences in acquisition and processing protocols across hospitals or laboratories. This issue is well known across many scientific fields and is generally referred to as a batch effect—a term first widely recognized in genomics [Johnson et al., 2007]. When present in data, site effects can negatively impact the generalization and performance of tools such as machine learning models [Tobón Quintero et al., 2022]. That is why this source of variability must be carefully investigated and addressed to avoid undesirable effects and to create reliable models.

2.2.1. Sources of site-related variance

In EEG, site effect can arise from a combination of various inter-center differences. The most significant sources include:

- **Hardware:** Different EEG equipment, like amplifiers, electrodes, and caps, that vary in technical specifications, exists across institutions, which can result in differences in the recorded data [Tobón Quintero et al., 2022].

- **Software:** The software used for data acquisition and processing, along with its settings, can also differ between institutions. For example, the choice of filters applied during recording can significantly influence the final shape of the EEG signal [Karpiel et al., 2021].

- **Recording protocols:** Although recording protocols are often standardized across medical institutions, minor differences may still occur, resulting in the site effect. This could include variations in procedure duration or differences in instructions given to patients.

- **Patient populations:** The characteristics of the patients can differ systematically between sites. This can be due to the specialization of each hospital, which may vary significantly—for example, one hospital might be a specialized epilepsy clinic with a high proportion of pathology cases, while another might be a neurology department in a general hospital where most patients are relatively healthy. Centers may also have different patient demographics; for example, a pediatric hospital may have a much younger patient population. The age differences, along with the gender, are the major sources of variation between sites.

- **Environmental factors:** The recording environment itself can create site-specific differences. These may include varying levels of power-line noise, the presence of other electronic devices generating the noise, or differences in how technicians prepare patients and conduct procedures.

2.3. Machine learning for EEG data analysis

Analyzing EEG data is a significant challenge for traditional analysis due to its complexity. Because of that, machine learning (ML) has become an important tool, as it can learn the patterns and anomalies within such high-dimensional data. With the increasing availability of EEG data in recent years, more powerful models are being created. Applications that include tasks such as neuroscreening for abnormalities, developing brain-computer interfaces, or creating tools to support clinical decision-making all rely on that automatic processing and decoding of brain activity [Kübler et al., 2001, Poziomska et al., 2024].

Typically, machine learning models applied to EEG data perform either classification or regression tasks. In clinical studies, classification is the most widely used approach, where models predict discrete values, for example, identifying different brain disorders or different cognitive and emotional states. On the other hand, regression models are used to predict a continuous target, such as attention, drowsiness, or relaxation levels [Saeidi et al., 2021].

2.3.1. Supervised learning and classification

A supervised model requires a dataset for its learning, called a training set, in which each input is paired with its correct output. The model uses a specific learning algorithm and adjusts its input–output relationship by comparing its predictions with the actual answer. This process is known as learning by example, and the goal is to obtain a model with updated parameters that can be later used to predict results when applied to unseen, new data with unknown labels [Hastie et al., 2016].

Classification is a specific type of supervised learning. The goal of a model is to predict a discrete category, which can be either a binary class (like a normal vs. pathological task) or

a multi-class, predicting one of more than two categories (like a site-classification task where the model predicts one of many hospitals).

2.3.2. Ensemble methods and gradient boosting

In machine learning, an ensemble predictor combines the outputs of multiple base predictors—simple models whose performance is limited when used alone. However, when multiple predictors like these are combined, they can create a much stronger and more accurate model. Boosting is a type of ensemble method where models are built sequentially, with each new model trying to correct the errors of the previous one. Gradient boosting goes further by performing a gradient descent for each new base predictor to reduce the model's errors. The process starts with a single, basic predictor whose initial prediction is often close to random guessing. Each next model is explicitly trained to reduce the errors, also known as residuals, that the previous model made. The process repeats until a final, stronger model is built [Friedman, 2001].

Decision trees are commonly used as base predictors in ensemble methods. They work by recursively splitting the feature space into disjoint regions. The process starts at the root node with the entire dataset, then each step performs a binary split based on feature values. Each split creates new branches, dividing the data into two groups. This continues until the terminal nodes, leaves, are reached. A final prediction is then assigned to all observations that fall into each of these leaves [Rokach and Maimon, 2005].

CatBoost [Prokhorenkova et al., 2018] is an open-source library that improves the standard gradient boosting algorithm by preventing the overfitting issue, known as target leakage. It also provides an original algorithm for effectively processing categorical features. CatBoost uses the previously mentioned decision trees as its base predictors.

2.4. Feature importance in machine learning

2.4.1. The concept of model explainability

Although machine learning applications can be effective across many tasks, in some fields, such as healthcare, performance alone is not sufficient. Understanding why a model makes a prediction is equally important. Yet, for a model to capture hidden and complex patterns in data, it often needs to be highly complex itself, with a large number of parameters and interactions among them. As a result, such models cannot be directly interpreted and are often treated as black-box methods. This lack of understanding creates problems of trust, accountability, and practical implementation in sensitive domains [Ortigossa et al., 2024].

Such models may not be directly interpretable, but specialized tools can help explain their decisions. This field, known as explainable artificial intelligence (XAI), aims to increase transparency in how AI systems make decisions. One commonly used technique is feature importance analysis, which assigns a score to each input feature based on its contribution to the model's output. This analysis can be performed at the local level, offering insights into a single prediction, or at the global level, explaining the model's overall behavior [Ortigossa et al., 2024].

2.4.2. SHAP (SHapley Additive exPlanations)

The SHAP (SHapley Additive exPlanations) method provides a way to measure of feature importance using classic game-theoretic Shapley values [Shapley, 1953]. For any class, the

model's prediction can be split into two parts. The first is the base value, representing the average prediction for that class across the entire dataset. The second is the sum of feature effects, showing how each feature influences the prediction relative to the base value. This is formally expressed as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2.1}$$

where $g(z')$ is the explanation model that approximates the original model's prediction, ϕ_0 is the base value (the average prediction over the training data), M is the number of features, and ϕ_i is the SHAP value for feature i. A single SHAP value provides a local explanation for a single prediction, quantifying how much a given feature has affected the model's prediction score for a certain class.

The SHAP TreeExplainer [Lundberg et al., 2020] was developed specifically for tree-based models. This method offers a polynomial-time computation for SHAP values, which is a significant optimization over model-agnostic alternatives, such as KernelSHAP [Lundberg and Lee, 2017].

2.5. Data harmonization techniques

The issue of site effect was first observed and analyzed in fields such as molecular biology and medical imaging [Johnson et al., 2007, Fortin et al., 2018]. These and other studies led to the development of statistical methods known as data harmonization techniques, which are explicitly designed to remove unwanted site-related differences. Data harmonization has become a standard practice in many fields when working with multi-center datasets [Hu et al., 2023].

In this section, i indexes subjects, j indexes features, and $s = b[i]$ indicates the site of subject i.

2.5.1. Site-wise standardization

A simple way to harmonize data is site-wise standardization. The idea is to transform the features within each site so they follow the same baseline distribution, with a mean of 0 and a standard deviation of 1. This is achieved by applying Z-score normalization independently for each feature at each site:

$$z_{ij} = \frac{x_{ij} - \mu_{sj}}{\sigma_{sj}}, \qquad \text{where } s = b[i]. \tag{2.2}$$

where:

- x_{ij} is the observed value of feature j for subject i,
- μ_{sj} is the mean of feature j calculated using only subjects from site s,
- σ_{sj} is the standard deviation of feature j within site s,
- $b[i]$ maps subject i to its site s.

However, because the main strength of this method is its simplicity, it also has limitations. The transformation only removes basic differences in the mean and variance between sites, may not fully correct for more complex site effects, which can arise, for example, from higher-order distributional shifts.

2.5.2. ComBat harmonization

To have a more robust method, it is possible to fit a linear regression model, treating the site as one of its covariates. It is also highly recommended to include biological variables in the model. Without this, the model assumes there are no differences related to any biological factor. If this assumption is not satisfied, the regression could accidentally remove the wanted, biological variations. The model can be expressed mathematically as follows:

$$y_{ij} = \alpha_j + X_i\beta_j + \gamma_{sj} + \epsilon_{ij}, \qquad \text{where } s = b[i]. \tag{2.3}$$

where:

- y_{ij} is the observed value of feature j for subject i,
- α_j is the intercept for feature j,
- X_i is the vector of biological covariates for subject i and β_j are the corresponding coefficients for feature j,
- γ_{sj} is the site effect for the site to which subject i belongs,
- ϵ_{ij} is random error.

Now, with the estimated site effect, the next step is to directly subtract this variability, resulting in y^*_{ij} that represents the data after removing the estimated site effect, ideally reflecting only biological variability:

$$y^*_{ij} = y_{ij} - \hat{\gamma}_{sj}, \tag{2.4}$$

However, even with biological covariates included, the regression approach is still too weak for complex site effects, as it only adjusts the mean, while in most cases the data also differ in the variance. As a solution to these limitations, the combining batches (ComBat) method was proposed [Johnson et al., 2007]. Unlike the regression model, it models the site effect as additive (shift in the mean) and multiplicative (scaling of the variance), which makes the feature distributions more comparable. The model for a given feature j and subject i can be written as:

$$y_{ij} = \alpha_j + X_i\beta_j + \gamma_{sj} + \delta_{sj}\epsilon_{ij}, \tag{2.5}$$

Here, δ_{sj} represents the multiplicative (variance-scaling) site effect, in addition to the additive effect γ_{sj} already introduced in the regression model.

For estimation, ComBat uses empirical Bayes, which is a significant advantage when working with small site sample sizes. It assumes that the site parameters follow a prior distribution:

$$\gamma_{sj} \sim N(\gamma_j, \tau_j^2), \quad \delta^2_{sj} \sim \text{Inverse-Gamma}(\lambda_j, \theta_j). \tag{2.6}$$

After estimating these parameters, the harmonized feature value y^*_{ij} is obtained by subtracting the site-specific impacts while preserving the contributions of biological covariates:

$$y^*_{ij} = \frac{y_{ij} - \alpha_j - X_i\beta_j - \hat{\gamma}_{sj}}{\hat{\delta}_{sj}} + \alpha_j + X_i\beta_j, \tag{2.7}$$

ComBat has proven effective at harmonizing a wide range of features in imaging ([Voß et al., 2022, Pomponio et al., 2020]), and researchers are now beginning to apply it to EEG data [Jaramillo-Jimenez et al., 2024, Henao Isaza, 2023].

2.5.3. Other ComBat variants

Different extensions of the original ComBat method have been developed to deal with challenges across various domains. For example:

- *NeuroHarmonize* [Pomponio et al., 2020]—a variant designed for scenarios where biological covariates have nonlinear relationships with features. It extends the ComBat by using Generalized Additive Models to preserve these nonlinear relationships.
- *OPNComBat-GMM (Optimized Nested ComBat with Gaussian Mixture Model)* [Horng et al., 2022]—is an extension of ComBat designed for scenarios where multiple site-related technical or biological factors contribute to site effects. It applies an iterative process for site combinations, and the Gaussian Mixture Model handles multi-modal features, preserving variation and reducing hidden site effects.
- *CovBat* [Chen et al., 2021]—extends ComBat by adjusting not only for mean and variance differences, but also for covariance across sites. Studies show CovBat outperforms other methods in both statistical tests and machine learning tasks across multiple neuroimaging datasets.

2.5.4. Other harmonization approaches

While ComBat and its variants have proven effective in harmonizing data, several other approaches have also been proposed. These include deep learning–based methods. A more detailed overview of other methods can be found in [Hu et al., 2023].

2.6. Quantifying effect size with Cohen's d

To measure the magnitude of a difference between two groups, a standardized measure of effect size Cohen's d [Cohen, 1988] is often used. Cohen's d quantifies the difference between two means in terms of their pooled standard deviation. The formula for Cohen's d is given by:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{\text{pooled}}} \tag{2.8}$$

where μ_1 and μ_2 are the means of the two groups being compared, and σ_{pooled} is the pooled standard deviation, calculated as:

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2.9}$$

with n_1, n_2 being the sample sizes and s_1, s_2 being the standard deviations of the two groups. A common practice is to interpret Cohen's d values around 0.2 as small, 0.5 as medium, and 0.8 or higher as large differences.

Chapter 3

Materials and methods

3.1. Data

3.1.1. ELM19 dataset

The dataset used in this work is the dataset presented in the study [Poziomska et al., 2024], where issues with the generalization ability of the GBE model were observed.

The dataset was collected by Elmiko Biosignals sp. z o.o.[1] and includes 55,787 clinical EEG recordings from 39 medical institutions across Poland. These recordings were originally gathered during clinical diagnostics. The data cover a wide range of patient cases from standard exams to assessments for various clinical conditions. Each recording is labeled as normal or pathological, based on the accompanying medical description, as detailed in [Poziomska et al., 2024].

Site selection

Because of the particular needs of this study, suitable preprocessing steps were applied to the dataset. Institutions with fewer than 50 normal recordings were excluded in order to guarantee a sufficient number of normal cases from each source for accurate classification. The distribution of normal cases across institutions was used to determine this threshold. Figure 3.1a shows the number of recordings classified as normal for each institution, while Figure 3.1b shows the total number of recordings (normal and pathological). Throughout this work, the resulting dataset will be referred to as ELM_s.

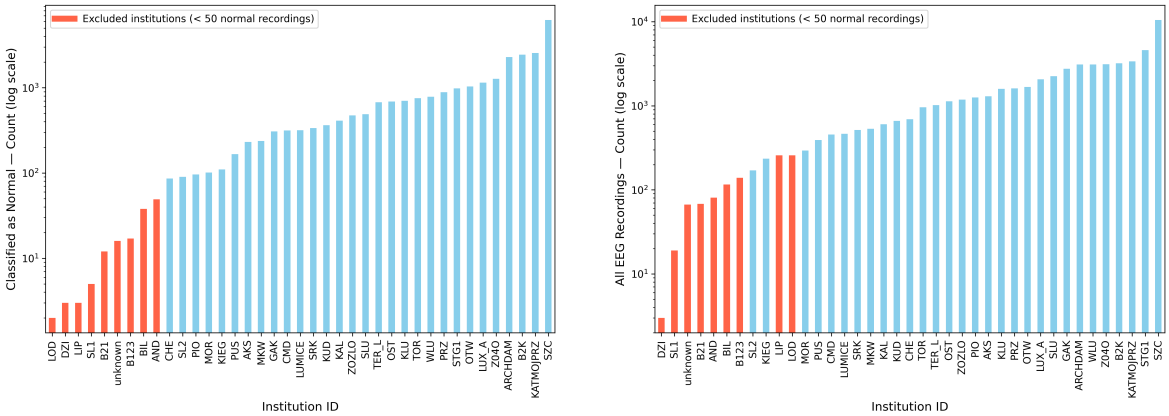
Characteristics of the final ELM_s

After filtering, the final ELM_s dataset contains 54,779 EEG recordings from the remaining 30 medical institutions. The patients' ages range from 16 to 70. There are 28,960 females (52.9%) in the gender distribution. The dataset is well balanced, with 26,568 (48.5%) recordings classified as normal and 28,211 (51.5%) classified as pathological.

3.1.2. Analysis using a normal data subset

To specifically isolate the site effect from any variance caused by clinical conditions, a key decision was made for this study. Unless otherwise specified, all analyses—including exploratory

[1] `https://www.elmiko.pl`



(a) Number of recordings classified as normal per institution.

(b) Total number of recordings (normal and pathological) per institution.

Figure 3.1: Distribution of recordings across medical institutions.

data analysis, site classification, and the evaluation of harmonization techniques—were performed only on the subset of recordings labeled as normal. This subset will be referred to as ELM_n. Thanks to that, the analysis more clearly targeted the baseline differences between sites. The only time when the whole dataset ELM_s was used was in the final clinical neuroscreening task, where the pathology classification was made.

3.1.3. EEG acquisition parameters, data format, and recording protocols

The EEG signals were acquired using the standard 10-20 system with 19 channels allocated as follows: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2. The data were recorded with original sampling frequencies ranging from 200 to 500 Hz and stored in the European Data Format (EDF) [Kemp et al., 1992]. The recordings originate from routine clinical diagnostic sessions, and the exact procedures could vary between institutions and patients, though they often include standard tasks such as photostimulation, hyperventilation, and periods of eye-opening and closing.

3.2. EEG data preprocessing

To prepare the ELM_s data for further feature extraction, a standardized preprocessing pipeline was applied. The aim of this step was to remove noise and artifacts that were present in the data. All preprocessing steps were implemented using custom scripts in Python, using the MNE-Python library [Gramfort, 2013].

The EEG signals were first filtered using Butterworth filters to eliminate frequencies commonly affected by muscle artifacts and slow drift. Specifically, a high-pass filter with a cutoff at 0.1 Hz and less than 1 dB passband ripple above 0.5 Hz was applied, along with a low-pass filter with a cutoff at 40 Hz, which decreased frequencies by at least 20 dB above 50 Hz while keeping a passband ripple below 1 dB. Next, to eliminate power-line noise, a notch filter with a frequency of 50 Hz and a quality factor $Q = 5$ was used. Afterward, all signals were re-referenced to a common average reference and resampled to 100 Hz to ensure a uniform sampling rate across all institutions. Subsequently, frames of adjacent 6-second segments were extracted from the preprocessed signals. Frames with voltages exceeding 800 μV or containing flat-line channels were rejected as artifacts.

3.3. Feature extraction

For the extraction of features, this study used the methodology described in prior work [Poziomska et al., 2024]. Several types of features were extracted from each 6-second EEG segment and later aggregated using the median across all frames in each recording.

For frequency-domain features, 14 overlapping frequency bands (f_b) were defined: $[0.5, 2]$, $[1, 3]$, $[2, 4]$, $[3, 6]$, $[4, 8]$, $[6, 10]$, $[8, 13]$, $[10, 15]$, $[13, 18]$, $[15, 21]$, $[18, 24]$, $[21, 27]$, $[24, 30]$, and $[27, 40]$ Hz. These bands cover the standard EEG frequency ranges: delta (δ), theta (θ), alpha (α), and low and high beta (β).

Three groups of features were extracted: time-domain covariance matrices, frequency-domain coherence features, and frequency-band power features. While the coherence and power features depended on the selected frequency bands, the covariance features were calculated directly in the time domain.

3.3.1. Time-domain covariance matrices

Time-domain covariance matrices between EEG channels lie on a Riemannian manifold [Congedo et al., 2013, Tibermacine et al., 2024]. So, since standard Euclidean operations are insufficient to capture their structure fully, a representative matrix that captures the spatial relationships between EEG channels was produced using the Riemannian metric [Moakher, 2005]. The covariance matrices were then vectorized by extracting the coefficients of the lower triangular part. This procedure creates 190 features per recording that efficiently capture the spatial covariance structure of the EEG signal. The pyRiemann[2] library was used to compute these features from beginning to end.

3.3.2. Frequency-domain power features

The power spectral densities (PSD) corresponding to the selected frequency bands, represented as $S_{xx}(f_b)$, where x is the channel index, are the most fundamental frequency-domain features. The multitaper method [Thomson, 1982] is used to estimate these PSDs. After estimation, the PSDs are normalized such that their sum across all channels and frequency bands equals one within each frame. This procedure results in 266 power features per recording extracted from the EEG signals using the MNE-Python library [Gramfort, 2013].

3.3.3. Frequency-domain coherence features

Band-wise coherence features are derived from the cross-spectral densities, $S_{xy}(f_b)$, according to Eq. 3.1:

$$C_{xy}(f_b) = \frac{|S_{xy}(f_b)|}{\sqrt{S_{xx}(f_b) \cdot S_{yy}(f_b)}} \tag{3.1}$$

where $S_{xx}(fb)$ and $S_{yy}(fb)$ are the power spectral densities of channels x and y, respectively. Only the sub-diagonal elements are kept, as the coherence matrix is symmetric with ones on the diagonal. This results in 2,394 coherence features per recording.

3.3.4. Final feature set

A total of 2,850 features are included in the final feature set, which was designed to capture multiple aspects of the EEG signal, including its spatial structure in the time domain, its

[2] `https://pyriemann.readthedocs.io/`

spectral power characteristics across a wide frequency range, and the connectivity between all electrode pairs.

3.4. Exploratory data analysis

One potential source of site effect, as discussed in section 2.2, is variation in patient populations and recording protocols. With the additional information available for the ELM_n dataset, it was possible to analyze these characteristics by comparing the distributions of patient age, gender, and EEG recording duration across all 30 institutions.

3.5. Investigating the site effect using machine learning

The site effect can result from complex combinations of factors that can be undetectable by simple observation. Machine learning models can capture such patterns even in high-dimensional data. To quantify the site effect, a supervised machine learning approach was used with a specific task: to classify the source hospital of an EEG recording based on its features. For this multi-class classification, the institution ID served as a target variable, and the 2,850 covariance, coherence, and power features (described in section 3.3) were used as the model's input.

3.5.1. Classifier selection and hyperparameter tuning

For this task, the CatBoost[3] classifier [Prokhorenkova et al., 2018] was selected. Its main strengths are the ability to handle high-dimensional data, efficiency in multi-class classification, and robustness to overfitting. Its hyperparameters were specifically adjusted for this task using the Optuna framework[4] [Akiba et al., 2019].

3.5.2. Performance evaluation

To evaluate model performance, the 5-fold stratified cross-validation was applied. Because of the class imbalance, stratification by hospital ID kept the proportion of recordings from each institution constant across folds. The Matthews Correlation Coefficient (MCC) was used as the evaluation metric. The one-vs-rest approach was used to compute the individual MCC scores for each hospital class. Specifically, the MCC for a given hospital was calculated by treating that hospital as the positive class and all other hospitals as the negative class (Eq. 3.2).

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{3.2}$$

where tp, tn, fp, and fn are respectively the number of true positives, true negatives, false positives, and false negatives.

Additionally, an overall MCC value was reported, calculated using the multiclass generalization of MCC. For a K-class problem, let $\mathbf{C} \in \mathbb{N}^{K \times K}$ be the confusion matrix, where C_{ij} denotes the number of samples known to be in class i (true class) and predicted to be in class j (predicted class). To define the multiclass MCC, the following intermediate variables are used:

[3] `https://catboost.ai/`

[4] Available at `https://optuna.org/`

- $t_k = \sum_{j=1}^{K} C_{kj}$ — the total number of times class k truly occurred (i.e., the sum of row k of $\mathbf{C}$).
- $p_k = \sum_{i=1}^{K} C_{ik}$ — the total number of times class k was predicted (i.e., the sum of column k of $\mathbf{C}$).
- $c = \sum_{k=1}^{K} C_{kk}$ — the total number of samples correctly predicted (i.e., the sum of the main diagonal of $\mathbf{C}$).
- $s = \sum_{i=1}^{K} \sum_{j=1}^{K} C_{ij}$ — the total number of samples.

The multiclass MCC is then given by (Eq. 3.3):

$$\text{MCC} = \frac{cs - \sum_{k=1}^{K} p_k t_k}{\sqrt{\left(s^2 - \sum_{k=1}^{K} p_k^2\right)\left(s^2 - \sum_{k=1}^{K} t_k^2\right)}} \tag{3.3}$$

3.5.3. Feature importance analysis

One of the primary objectives of this thesis is to interpret the model's predictions for the source institution and to identify the features that were most significantly influenced by the site effect. To quantify the feature contribution, as introduced in the theoretical background (section 2.4.2), the SHAP[5] approach was applied.

Feature contributions for the trained CatBoost model were computed using the SHAP TreeExplainer. Since this is a multi-class problem, SHAP values were calculated for each feature for every output class. To create a global measure of feature importance that summarizes the overall impact of each feature, the mean of the absolute SHAP values was calculated across all predictions and all classes.

3.5.4. Ablation study of feature groups

To understand how different types of features contributed to the site-classification task, an ablation study was performed. Features were grouped based on their type: time-domain covariance, power spectral densities, and band-wise coherences. Firstly, three separate models were trained, each using only features, to check the performance that could be achieved with each group individually. Next, three additional models were trained, each leaving out one group, to see how the model performs without that particular group of features.

3.6. Data harmonization

3.6.1. Selection of harmonization techniques

Two harmonization techniques, which are described in the section 2.5, were selected for this study: site-wise standardization as a straightforward baseline method, and the neuroCombat, Python implementation of ComBat [Fortin et al., 2018], as a more robust method.

[5] `https://shap.readthedocs.io/`

3.6.2. Application of harmonization

Both harmonization methods were used on the full feature set. For site-wise standardization, the Z-score transformation was applied to each feature independently within each hospital group based on the hospital ID. For ComBat, the hospital ID was used as the batch variable, and it also included patient age and gender as biological covariates.

3.6.3. Evaluating harmonization effectiveness

Firstly, to evaluate the impact of harmonization on the individual feature distributions, Cohen's d (as described in section 2.6) was used. The effect size was calculated for each feature for every pair of institutions and then averaged across all features. This produced an all-vs-all comparison matrix, from which a summary score for each hospital was obtained by averaging across rows.

Secondly, to assess the overall impact on the site effect, the CatBoost site-classification model was retrained and evaluated on the data after site-wise standardization and ComBat harmonization, using the same 5-fold cross-validation and MCC metric as the initial analysis.

3.7. Clinical neuroscreening task

The final task was to evaluate the practical impact of data harmonization on a real-world neuroscreening task.

3.7.1. Task definition and data

The task was to classify EEG recordings as either "normal" or "pathological". For this reason, this analysis utilized the full dataset ELM_s, which includes recordings labeled both as normal and pathological.

3.7.2. Classification model

This study used the same gradient-boosted ensemble (GBE) model as in the prior work [Poziomska et al., 2024], to ensure the results are comparable. The GBE is composed of 30 individual CatBoost classifiers, which were trained independently and differ slightly from one another because of stochastic components in the learning algorithm. The output probabilities are then averaged to produce the final prediction. The choice of this model was mainly to make the results comparable across the two studies. The model was run using the same hyperparameters as the original study for consistency.

3.7.3. Performance evaluation

This study was performed with two strategies. Performance for both settings was measured using the Matthews Correlation Coefficient (MCC) and the Area Under the Curve (AUC). The description of them is as follows:

- **5-Fold cross-validation:** A 5-fold cross-validation was used to obtain a robust measure of the model's general performance. The stratification was performed based on the hospital ID, ensuring that each fold's training and test sets contain a representative sample of data from all 30 institutions.

- **Leave-one-site-out (LOSO):** To test the ability of generalization of a model, a Leave-One-Site-Out (LOSO) analysis was chosen. In each fold, the model was trained on data from all sites except one and then tested on that held-out site. The data harmonization methods used in this analysis are unable to transform the hospital's data, which was not observed during the fitting process. Because of that, a specific procedure was needed to perform such a task. A random calibration sample ($N = 30$) of recordings classified as normal from the test site was set aside. The remaining data from that site was then used as the final test set. For the harmonized evaluations, this calibration sample was used to fit the harmonization algorithm together with the training data from other institutions. For the unharmonized evaluation, the calibration sample was simply discarded to guarantee that all models were trained and tested on the exact same data.

Chapter 4

Results

4.1. Exploratory analysis of institutional characteristics

4.1.1. Age analysis

Analyzing the age distribution of patients itself revealed significant differences between hospitals, as shown in the boxplot in Figure 4.1. More detailed statistics are provided in Appendix A (see Table A.1). For instance, institutions like PIO, SL2, and TER_L have patients that are significantly younger, with median ages of approximately 17.6, 19.5, and 17.8 years, respectively. In contrast, institutions such as MOR and TOR served significantly older populations, with median ages of 52.1 and 50.8 years, respectively. The 30-year difference in the median age of patients suggests significant heterogeneity in the dataset.

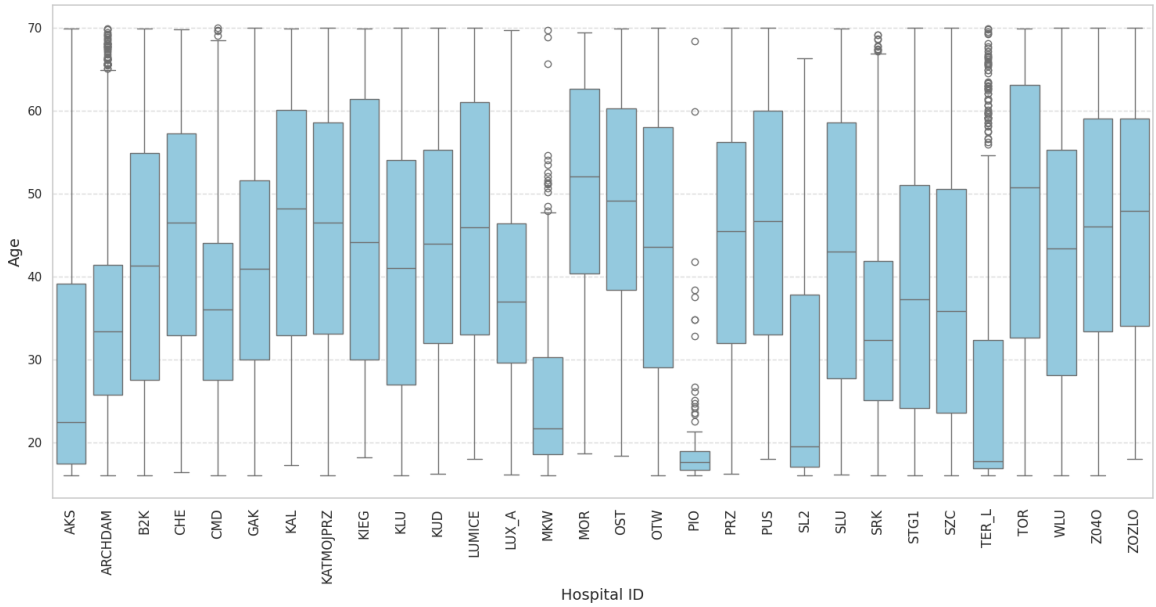


Figure 4.1: Age Distribution by Hospital

4.1.2. Gender analysis

Another analyzed data characteristic was a patient's gender distribution (Figure 4.2). While the gender distribution was generally balanced across most institutions, a few showed slight

imbalances. For instance, SRK has more female participants, while MKW has more male participants.

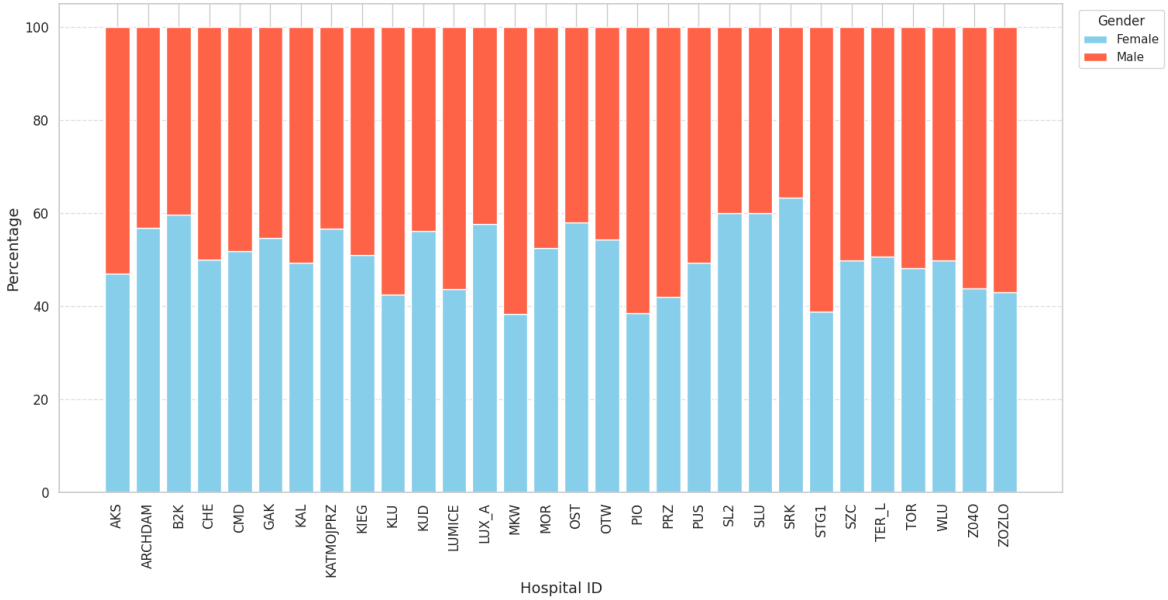


Figure 4.2: Gender Distribution by Hospital

4.1.3. Recording duration

Analysis of the recording durations revealed further heterogeneity between institutions, as shown in Figure 4.3. More detailed statistics are provided in Appendix A (see Table A.2). While the majority of centers followed protocols resulting in recordings with a typical duration of approximately 1,000 seconds, some significant deviations were noted. For instance, the institution KATMOJPRZ used much longer protocols, with some sessions lasting over 8,000 seconds. Furthermore, the institution ZOZLO, in addition to standard-length recordings, included a substantial number of sessions with a duration of approximately 4,000 seconds, suggesting that multiple recording protocols were used.

4.2. Quantifying and interpreting the site effect

4.2.1. Site classification performance

The CatBoost classifier, configured and validated as described in Chapter 3 (Section 3.5), showed a strong ability to classify the hospital of origin from extracted EEG features. The model achieved an overall MCC value of 0.865 ± 0.003 across the 30 ELM_s institutions in the 5-fold stratified cross-validation. The averaged normalized confusion matrices are presented in the appendix (see Figure A.1), while the averaged one-vs-rest MCC scores for each hospital are illustrated in Figure 4.4. Although only the average MCC is shown in the figure for visual clarity, the complete mean values along with their standard deviations across the 5 cross-validation folds are provided in the appendix (see Table A.3).

The model performed very well for the majority of hospitals, with 23 out of 30 achieving average MCC scores above 0.7. However, a small subset of institutions—AKS, CHE, KIEG, and SL2—performed poorly, with average MCC scores ranging from 0.39 to 0.52, making them

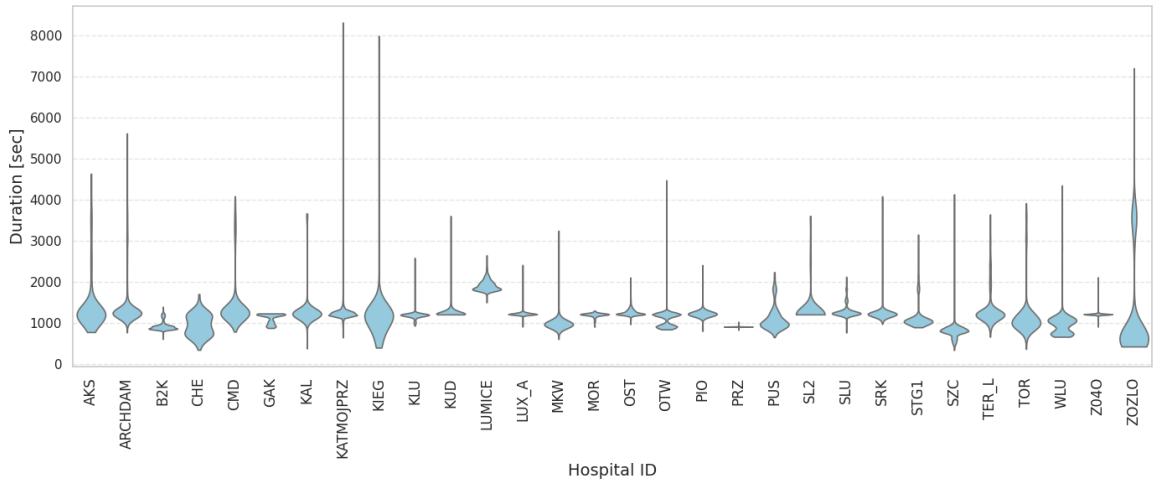


Figure 4.3: Distribution of Recording Durations by Hospital

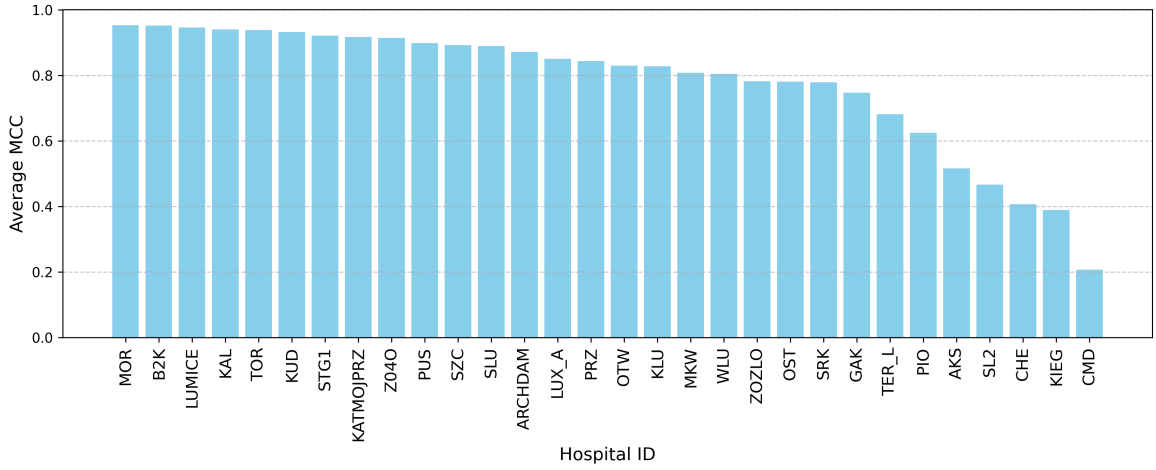


Figure 4.4: Average MCC for Each Class (5-Fold CV)

significantly more difficult to classify. The lowest average MCC score of 0.21 was achieved by the institution CMD.

4.2.2. Identifying key drivers with SHAP analysis

To identify what's behind the site differences, the global feature importance was analyzed using the average SHAP values, grouped by feature type and frequency band. This resulted in several plots, showing the overall importance of each feature type. A more detailed set of visualizations, showing all frequency sets, is provided in the appendix (Figure A.2).

Several key findings were observed. Among the coherence features, the connection between O1-O2 electrodes stood out, achieving the highest scores across the majority of frequency bands. Additionally, an increased importance of coherence in the beta band was observed, particularly in the temporal areas (T3, T4, T5, T6, F7, F8).

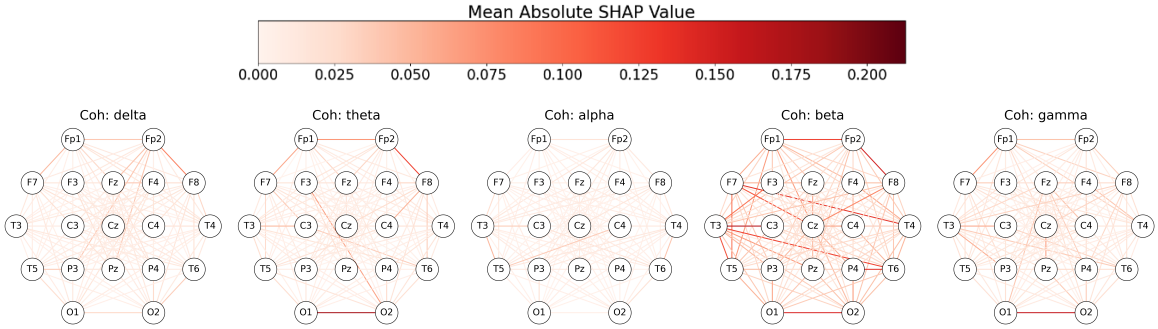


Figure 4.5: Average SHAP values for coherence features, grouped by EEG frequency bands.

While coherence features were the most dominant for the model, spectral power features also played a key role. The low-frequency delta band was particularly significant. In addition, the power of the Fz electrode stood out as important across multiple frequency bands.

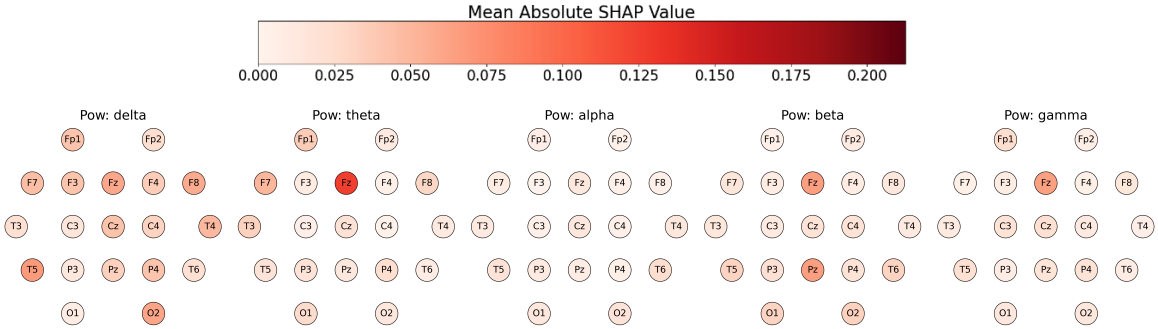


Figure 4.6: Average SHAP values for power features, grouped by EEG frequency bands.

In contrast, covariance features contributed little to the model's predictions and were found to be of much lower importance.

4.2.3. Ablation study of feature groups

The next step of the analysis was to evaluate the contributions of different feature types to the site effect. The model was trained separately on coherence, power, and covariance features using 5-fold stratified cross-validation. The overall MCC scores for each feature group are

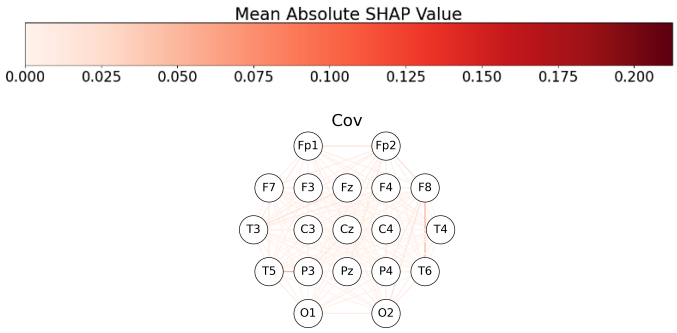


Figure 4.7: Average SHAP values for covariance features.

summarized in Table 4.1, and the averaged one-vs-rest MCC scores for each hospital are shown in Figure 4.8.

Table 4.1: Overall MCC scores and standard deviations for the site-classification task using individual feature groups.

Feature Type	**Overall MCC**	**Std. Dev.**
Coherence	0.850	0.003
Power	0.600	0.005
Covariance	0.752	0.009

While excluding a specific feature group, the overall MCC scores are summarized in Table 4.2 for models trained without coherence, power, and covariance features, respectively. Additionally, averaged one-vs-rest MCC scores for each hospital are shown in Figure 4.9.

Table 4.2: Overall MCC scores and standard deviations for the site-classification task when excluding a specific feature group.

Excluded Feature Type	**Overall MCC**	**Std. Dev.**
Coherence	0.785	0.007
Power	0.857	0.002
Covariance	0.863	0.004

The results from both Figure 4.8 and Figure 4.9 show that a model trained only on coherence features achieved site classification performance comparable to a model trained on all available features. In contrast, the model that was trained only on the power features performed significantly worse, indicating that this feature group carries less site-specific information.

4.3. Applying and evaluating data harmonization

After the site effect was quantified, the next step was to apply data harmonization techniques to reduce it. Two methods were evaluated: site-wise Z-score standardization and the ComBat. The effectiveness of these techniques was first evaluated at the feature level and then by

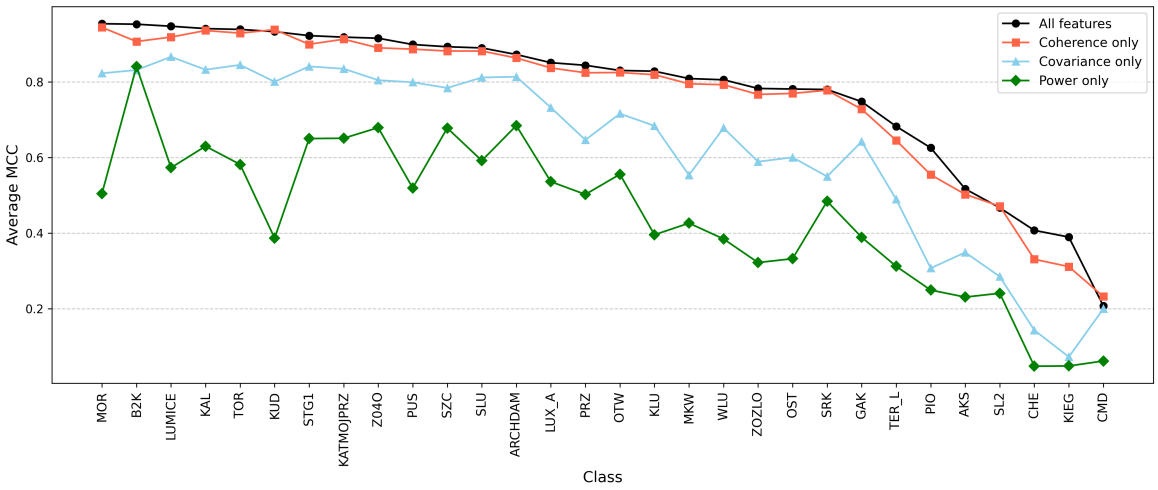


Figure 4.8: Average MCC per Hospital (5-Fold CV): Performance with Individual Feature Group

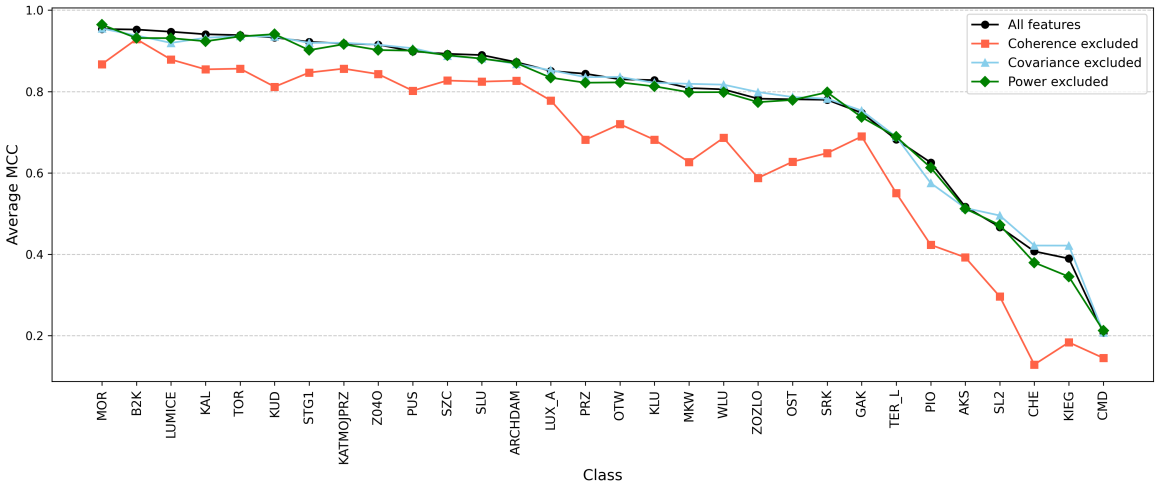


Figure 4.9: Average MCC per Hospital (5-Fold CV): Performance with Feature Group Exclusion

observing their impact on the site-classification model, directly comparing the results with those from the previous section.

4.3.1. Feature-level harmonization

The impact of harmonization was evaluated using Cohen's d, and the results of this analysis are presented in Figure 4.10. The results from the unharmonized data show that there were significant differences at the trait level, with some institutions, such as MOR, showing a large mean Cohen's d value of 0.79. After harmonization with both site-wise standardization and ComBat, the mean site difference scores were mostly reduced across all institutions. This reduction is an expected outcome of standardization and a confirmation that the methods successfully minimized inter-site differences at the feature level.

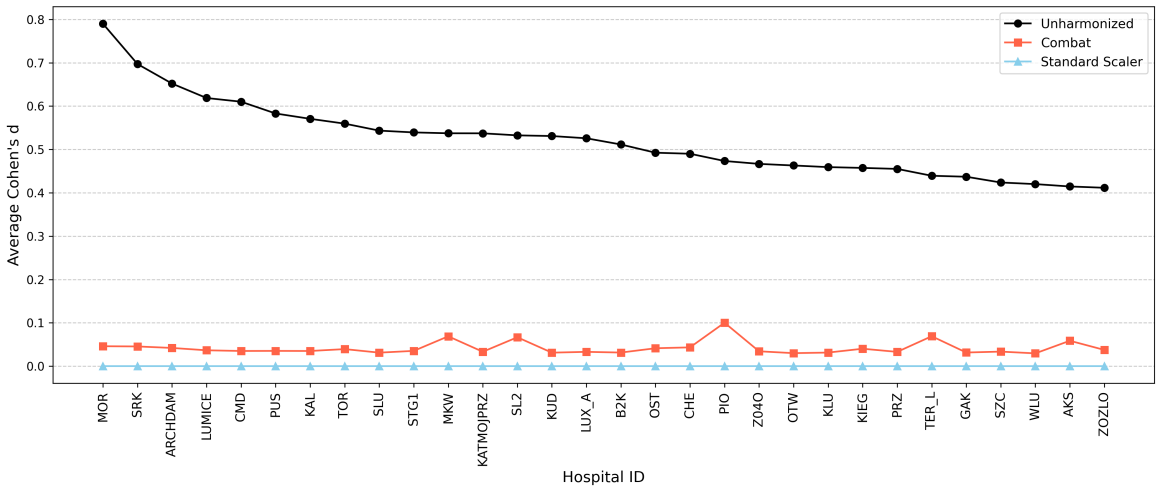


Figure 4.10: Effect of Harmonization on feature level (Cohen's d)

4.3.2. Paradoxical impact on site-classification

After confirmation that harmonization indeed reduced the inter-site differences in the feature distributions, the next step was to assess its effect on the site-classification model. The model was retrained and evaluated on the harmonized datasets. A counter-intuitive result was observed—rather than making the source hospital more difficult to identify, both harmonization methods led to a significant improvement in the model's classification performance, as shown in Figure 4.11.

A summary of the overall MCC scores is presented in Table 4.3.

Table 4.3: Overall MCC score and standard deviation for the site-classification task on the original data and after applying two harmonization methods

Harmonization Method	**Overall MCC**	**Std. Dev.**
Original Data (Unharmonized)	0.865	0.003
Site-wise Standardization	0.986	0.003
ComBat	0.983	0.002

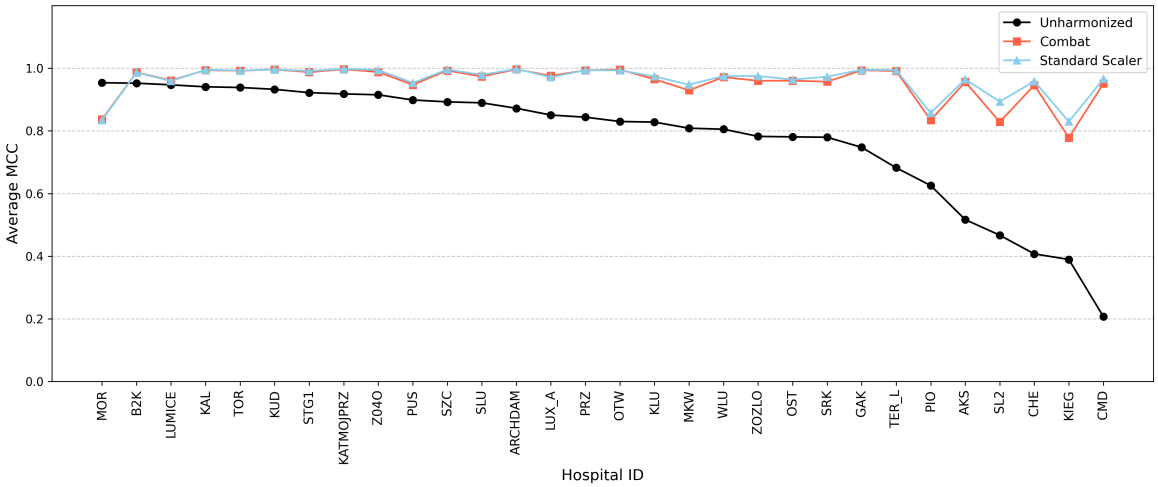


Figure 4.11: Effect of harmonization on hospital classification performance

4.4. Impact of harmonization on the clinical neuroscreening task

The harmonization methods were evaluated on the pathology classification task, a neuroscreening task. Two analyses were made, including a standard 5-fold cross-validation and Leave-One-Site-Out (LOSO) analysis.

4.4.1. Performance in standard cross-validation

The summary of the 5-fold cross-validation performance for each harmonization method is presented in Table 4.4. The model trained on the harmonized data performed slightly better than using the original data. There was no significant difference in score between ComBat and site-wise standardization.

Table 4.4: Summary of 5-fold cross-validation performance on the pathology classification task.

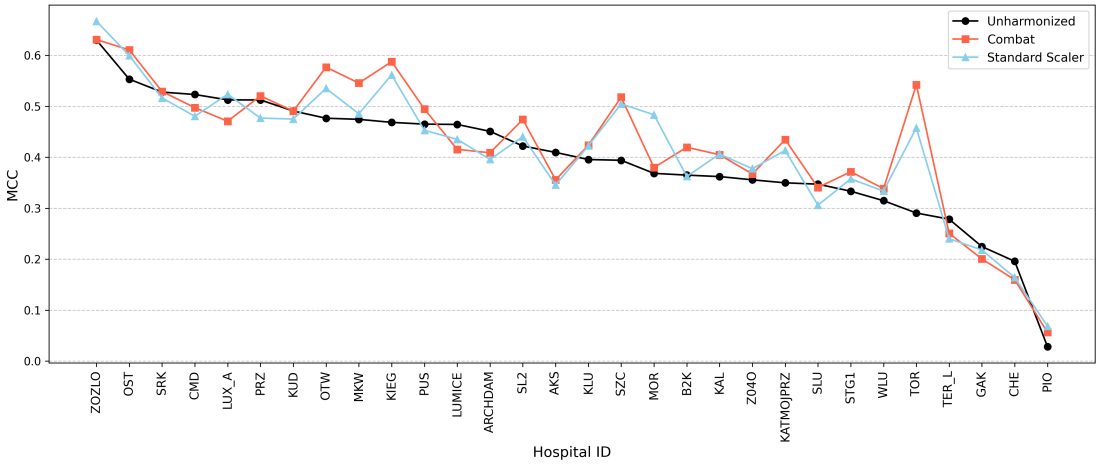
Harmonization Method	**MCC ± SD**	**AUC ± SD**
Original Data (Unharmonized)	0.560 ± 0.005	0.863 ± 0.001
Site-wise Standardization	0.592 ± 0.008	0.878 ± 0.003
ComBat	0.590 ± 0.011	0.877 ± 0.003

4.4.2. Generalization performance (leave-one-site-out analysis)

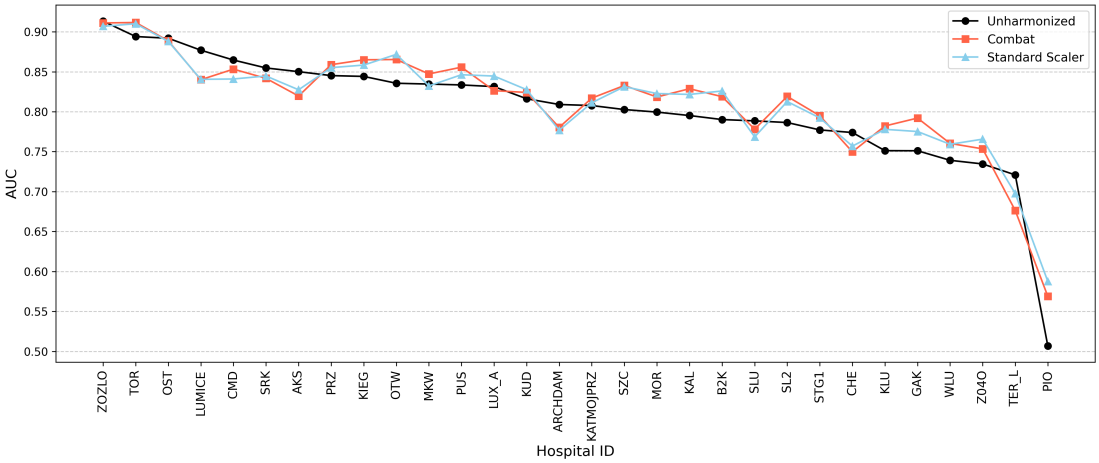
The results from the Leave-One-Site-Out analysis, which tests generalization to different hospitals, are summarized in Table 4.5. The detailed performance for each individual site is visualized in Figure 4.12. Models using harmonized data showed slightly better results than those using the original dataset. Again, ComBat and site-wise standardization performed similarly.

Table 4.5: Summary of LOSO Harmonization Performance for Pathology Classification (MCC)

Harmonization Method	**MCC ± SD**	**AUC ± SD**
Original Data (Unharmonized)	0.400 ± 0.005	0.863 ± 0.001
Site-wise Standardization	0.417 ± 0.008	0.878 ± 0.003
ComBat	0.427 ± 0.011	0.877 ± 0.003



(a) LOSO Harmonization Performance for Pathology Classification (MCC)



(b) LOSO Harmonization Performance for Pathology Classification (AUC)

Figure 4.12: LOSO Harmonization Performance for Pathology Classification. Results are shown for (a) MCC and (b) AUC.

Chapter 5

Discussion

5.1. Summary of key findings

The exploratory analysis revealed differences between medical institutions, especially in patient demographics and recording protocols. Other factors that could contribute, including hardware or software differences, were not analyzed. Together, these differences resulted in a strong site effect, confirmed by a CatBoost classifier, which could identify the source hospital from EEG features with a high overall MCC score (0.865 ± 0.003).

To better understand the factors underlying this effect, a feature importance analysis was performed. SHAP analysis indicated that beta-band coherence in the temporal regions, coherence between O1 and O2 electrodes, delta-band power, and the power at electrode Fz were the most important factors behind the model's predictions. An ablation study confirmed the importance of coherence features, as a model trained exclusively on them achieved comparable performance (MCC of 0.850 ± 0.003) to that of the full feature set.

The main findings from the analysis of harmonization methods, site-wise standardization, and the ComBat algorithm are:

1. **Feature level:** Both methods were successful at the feature level, reducing the differences in feature distributions as measured by Cohen's d.

2. **Model level:** Unexpectedly, when the model was trained on the harmonized data, both methods actually improved the site-classification model (increasing the MCC to over 0.98), making the source hospital easier to identify.

3. **Neuroscreening task:** In the clinical task, both methods provided a slight but consistent improvement compared to unharmonized features.

5.2. Interpretation of findings

5.2.1. The nature of the site effect

The feature importance analysis revealed that coherence features were the main contributors to the site classification model's performance. Both SHAP analysis and the ablation study confirmed this. In particular, coherence between O1 and O2 electrodes and coherence in the beta band, mainly in the temporal regions. The O1-O2 coherence importance might reflect differences in visual state, such as variations in eyes-open and eyes-closed protocols across sites. Beta band activity is in the range of muscle (EMG) signals, and the temporal regions are often affected by them. This suggests that hospital differences could be related

to procedural variations across sites, including how patients are prepared and instructed for recording, as well as natural differences in relaxation or alertness.

In addition, power features were also important in the model's predictions. Low-frequency delta power may be related to differences in patient sleepiness, but it could also arise from differences in electrode placement, such as skin cleaning procedures or the use of conductive gel. Similarly, the Fz electrode could reflect frontal muscle activity or differences in cognitive state, but it could also be influenced by the type of cap used at each site.

5.2.2. The harmonization paradox

The improvement in the site classification model's performance is counterintuitive. One reason why harmonization might not successfully remove the site effect is that both methods used in this study adjust the mean and variance of each feature individually. The problem is that site effects are not only present in single features, but they are also in the covariance relationships between them [Chen et al., 2021]. In a high-dimensional feature space, even small differences in correlation can accumulate, so adjusting only the mean and variance does not remove site-specific patterns.

At the same time, harmonization might actually make it easier to classify the site. By making the feature distributions more comparable and reducing the noise from simple shifts, the underlying site-specific structure could become more visible to the model.

5.2.3. The impact on the clinical neuroscreening task

The final result of this study is that both harmonization methods led to a small but consistent improvement in the clinical neuroscreening task. From previous results, harmonization is actually making the site classification easier, and because of that, these promising results should be taken with caution. For example, the model could benefit from site-specific information and learn that a particular hospital has a higher rate of pathology cases, and use that to make predictions. If this were the case, it could lead to problems with generalization.

However, results from the LOSO analysis suggest that this does not create major generalization issues, as the model's performance did not drop and even improved in certain hospitals. The harmonization appears to provide a benefit in the neuroscreening task, primarily due to the alignment of feature distributions, as ComBat did not offer any additional improvement over simple site-standardization.

At the same time, it is important to remember that site effects are still present in the data and could influence the model's predictions. While harmonization appears to be helpful, its impact is limited, and the challenge of site effects remains unresolved.

5.3. Study limitations and future work

The limitation of the harmonization methods used in this study is that they only adjust the mean and variance of individual features. Future research should focus on more powerful, multivariate harmonization techniques, such as CovBat (introduced in subsection 2.5.3), which is specifically designed to harmonize the covariance structure of the data. Additionally, advanced deep learning approaches could be explored to further reduce site effects.

The choice of features has a strong impact on the site effects. As shown in the ablation study, not all feature types carry the same amount of site-specific information, and some features appear to be more affected than others. Future work could explore why certain

features are particularly sensitive to site differences and whether alternative features exist that are less affected.

5.4. Overall conclusion

In conclusion, this study demonstrates that the site effect is present in the ELM_s dataset, representing a highly complex and multivariate challenge. The fundamental methods used in this study, ComBat and site-wise standardization, are among the most popular harmonization techniques. When applying them to the ELM_s dataset, the results show that, although they still offer benefits for clinical applications, they are insufficient to eliminate site effects completely. More advanced, multivariate harmonization approaches are needed to create truly generalizable neuroscreening models.

Bibliography

[Akiba et al., 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework.

[Beniczky et al., 2017] Beniczky, S., Aurlien, H., Brøgger, J. C., Fuglsang-Frederiksen, A., Martins da Silva, A., Trinka, E., Wiebe, S., and Tomson, T. (2017). Standardized computer-based organized reporting of eeg: Score—second version. *Clinical Neurophysiology*, 128(11):2334–2346.

[Chen et al., 2021] Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., and Shou, H. (2021). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human Brain Mapping*, 43(4):1179–1195.

[Cohen, 1988] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.

[Congedo et al., 2013] Congedo, M., Barachant, A., and Andreev, A. (2013). A new generation of brain-computer interface based on riemannian geometry. *CoRR*, abs/1310.8115.

[Croft et al., 2005] Croft, R. J., Chandler, J. S., Barry, R. J., Cooper, N. R., and Clarke, A. R. (2005). Eog correction: a comparison of four methods. *Psychophysiology*, 42(1):16–24.

[Doppelmayr et al., 1998] Doppelmayr, M., Klimesch, W., Pachinger, T., and Ripper, B. (1998). Individual differences in brain dynamics: important implications for the calculation of event-related band power. *Biological Cybernetics*, 79(1):49–57.

[Fortin et al., 2018] Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

[Gramfort, 2013] Gramfort, A. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7.

[Hastie et al., 2016] Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[Henao Isaza, 2023] Henao Isaza, V. (2023). Machine learning model for the classification of individuals at risk of dementia type alzheimer from multimodal databases of eeg and clinical information. Master's thesis, Universidad de Antioquia.

[Horng et al., 2022] Horng, H., Singh, A., Yousefi, B., Cohen, E. A., Haghighi, B., Katz, S., Noël, P. B., Kontos, D., and Shinohara, R. T. (2022). Improved generalized combat methods for harmonization of radiomic features. *Scientific Reports*, 12(1).

[Hu et al., 2023] Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T. D., Yu, M., et al. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage*, 274:120125.

[Jaramillo-Jimenez et al., 2024] Jaramillo-Jimenez, A., Tovar-Rios, D. A., Mantilla-Ramos, Y.-J., Ochoa-Gomez, J.-F., Bonanni, L., and Brønnick, K. (2024). Combat models for harmonization of resting-state eeg features in multisite studies. *Clinical Neurophysiology*, 167:241–253.

[Johnson et al., 2007] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

[Karpiel et al., 2021] Karpiel, I., Kurasz, Z., Kurasz, R., and Duch, K. (2021). The influence of filters on EEG-ERP testing: Analysis of motor cortex in healthy subjects. *Sensors (Basel)*, 21(22):7711.

[Kemp et al., 1992] Kemp, B., Värri, A., Rosa, A. C., Nielsen, K. D., and Gade, J. (1992). A simple format for exchange of digitized polygraphic recordings. *Electroencephalography and Clinical Neurophysiology*, 82(5):391–393.

[Kübler et al., 2001] Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., and Birbaumer, N. (2001). Brain–computer communication: Unlocking the locked in. *Psychological bulletin*, 127(3):358.

[Lopes da Silva and Schomer, 2018] Lopes da Silva, F. and Schomer, D. L. (2018). *Niedermeyer's electroencephalography : basic principles, clinical applications and related fields*. Oxford University Press, Oxford, seventh edition.

[Lundberg et al., 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4765–4774.

[Moakher, 2005] Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747.

[Ortigossa et al., 2024] Ortigossa, E. S., Gonçalves, T., and Nonato, L. G. (2024). Explainable artificial intelligence (xai)—from theory to methods and applications. *IEEE Access*, 12:80799–80846.

[Pomponio et al., 2020] Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., et al. (2020).

Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450.

[Poziomska et al., 2024] Poziomska, M., Dovgialo, M., Olbratowski, P., Niedbalski, P., Ogniewski, P., Zych, J., Rogala, J., and Żygierewicz, J. (2024). Quantity versus diversity: Influence of data on detecting eeg pathology with advanced ml models. *arXiv preprint arXiv:2411.17709*.

[Prokhorenkova et al., 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6639–6649.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487.

[Saeidi et al., 2021] Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., and Al-Juaid, A. (2021). Neural decoding of EEG signals with machine learning: A systematic review. *Brain Sci.*, 11(11):1525.

[Salinsky et al., 1987] Salinsky, M., Kanter, R., and Dasheiff, R. M. (1987). Effectiveness of multiple eegs in supporting the diagnosis of epilepsy: An operational curve. *Epilepsia*, 28(4):331–334.

[Shapley, 1953] Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pages 307–317. Princeton University Press.

[St. Louis et al., 2016] St. Louis, E. K., Frey, L. J., Britton, J. W., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., and Pestana-Knight, E. M. (2016). *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, Chicago, IL.

[Teplan, 2002] Teplan, M. (2002). Fundamentals of EEG measurement. *Measurement Science Review*, 2(1):1–11.

[Thomson, 1982] Thomson, D. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096.

[Tibermacine et al., 2024] Tibermacine, I. E., Russo, S., Tibermacine, A., Rabehi, A., Nail, B., Kadri, K., and Napoli, C. (2024). Riemannian geometry-based eeg approaches: A literature review.

[Tobón Quintero et al., 2022] Tobón Quintero, C. A., Ochoa Gómez, J. F., Li, M., Wang, Y., López Naranjo, C., Hu, S., García Reyes, R. C., Paz Linares, D., Areces González, A., Abd Hamid, A. I., et al. (2022). Harmonized-multinational qeeg norms (harmnqeeg). *NeuroImage*, 258:119352.

[Voß et al., 2022] Voß, H., Schlumbohm, S., Barwikowski, P., Wurlitzer, M., Dottermusch, M., Neumann, P., Schlüter, H., Neumann, J. E., and Krisp, C. (2022). Harmonizr enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nature Communications*, 13(1):3523.

[Zhang et al., 2022] Zhang, C., Sabor, N., Luo, J., Pu, Y., Wang, G., and Lian, Y. (2022). Automatic removal of multiple artifacts for single-channel electroencephalography. *J. Shanghai Jiatong Univ.*, 27(4):437–451.

Appendix A

Additional figures and tables

Table A.1: Statistics for age by hospital

Hospital	Mean	Median	Std	Min	Max
AKS	29.37	22.42	14.69	16.00	69.92
ARCHDAM	34.93	33.33	12.48	16.01	69.83
B2K	41.50	41.33	15.48	16.00	69.92
CHE	44.21	46.50	15.54	16.42	69.75
CMD	37.64	36.05	13.34	16.04	70.00
GAK	41.25	40.92	13.84	16.00	70.00
KAL	46.20	48.15	15.57	17.25	69.92
KATMOJPRZ	45.56	46.50	14.78	16.02	70.00
KIEG	44.70	44.08	16.63	18.17	69.92
KLU	41.03	41.00	15.95	16.00	70.00
KUD	43.53	43.92	14.23	16.25	70.00
LUMICE	46.06	45.92	15.44	18.00	70.00
LUX_A	39.04	36.92	12.60	16.08	69.67
MKW	26.12	21.67	10.75	16.01	69.67
MOR	48.62	52.05	15.39	18.67	69.42
OST	48.12	49.17	13.72	18.42	69.92
OTW	43.52	43.58	16.04	16.00	70.00
PIO	20.26	17.63	8.25	16.00	68.33
PRZ	44.19	45.42	14.54	16.25	70.00
PUS	45.64	46.63	15.63	18.00	70.00
SL2	28.42	19.54	15.07	16.00	66.25
SLU	42.79	42.96	16.58	16.07	69.92
SRK	34.97	32.29	13.35	16.04	69.08
STG1	38.16	37.25	14.95	16.05	70.00
SZC	37.71	35.79	15.42	16.00	70.00
TER_L	26.29	17.75	14.98	16.00	69.83
TOR	47.62	50.75	16.48	16.04	69.92
WLU	42.28	43.33	15.44	16.00	70.00
Z04O	45.47	46.00	15.11	16.00	70.00
ZOZLO	46.30	47.92	14.68	18.00	70.00

Table A.2: Statistics of EEG recording duration (in seconds) by hospital

Hospital	**Mean**	**Median**	**Std**	**Min**	**Max**
AKS	1423.16	1210.00	694.90	774.00	4635.23
ARCHDAM	1433.89	1235.00	599.39	765.00	5620.00
B2K	928.45	906.00	107.90	606.00	1392.00
CHE	944.44	928.50	267.94	342.00	1701.00
CMD	1434.49	1235.00	611.01	785.00	4085.00
GAK	1114.57	1206.00	124.17	875.00	1230.00
KAL	1299.85	1215.00	433.60	375.00	3665.00
KATMOJPRZ	1236.52	1212.06	169.56	642.03	8322.42
KIEG	1146.71	1206.06	706.37	396.02	7992.40
KLU	1184.02	1195.02	103.01	930.00	2585.04
KUD	1254.27	1218.00	138.11	1204.99	3606.00
LUMICE	1921.79	1890.00	138.58	1500.00	2646.00
LUX_A	1226.05	1212.00	71.78	906.00	2412.00
MKW	997.44	966.00	226.02	606.00	3246.00
MOR	1196.89	1205.00	52.64	906.00	1300.00
OST	1255.93	1225.00	81.57	966.00	2105.00
OTW	1076.59	1190.06	186.16	840.04	4475.22
PIO	1233.28	1206.06	153.60	804.04	2406.12
PRZ	908.59	906.00	11.75	828.00	1026.00
PUS	1109.12	980.00	293.35	650.00	2235.00
SL2	1328.17	1205.00	457.43	1205.00	3605.00
SLU	1292.20	1236.00	149.84	762.00	2124.00
SRK	1227.97	1206.00	200.25	976.00	4085.00
STG1	1142.95	1035.00	321.29	891.00	3153.00
SZC	806.40	790.00	197.36	336.00	4135.00
TER_L	1342.44	1206.00	436.30	660.03	3642.18
TOR	1160.52	1025.00	546.80	365.00	3918.00
WLU	966.69	995.00	305.34	660.00	4350.00
Z04O	1212.96	1206.00	48.39	906.00	2112.00
ZOZLO	1146.58	609.00	1180.99	423.00	7206.00

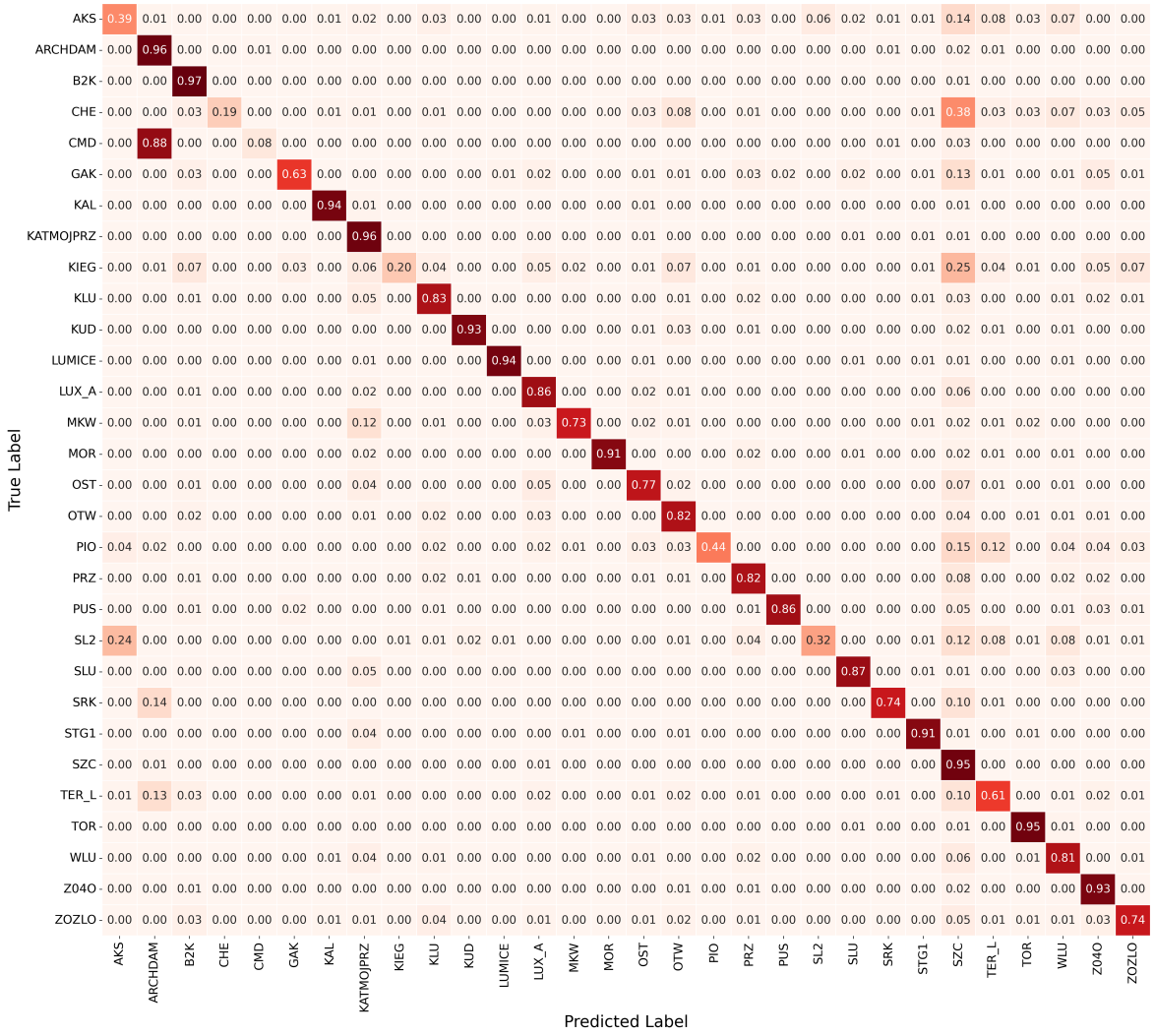


Figure A.1: Averaged normalized confusion matrix across the 30 ELM19 institutions in the 5-fold stratified cross-validation.

Table A.3: Mean MCC values and standard deviations for each hospital (5-fold cross-validation).

Hospital	**Mean MCC**	**Std. Dev.**
AKS	0.517	0.117
ARCHDAM	0.872	0.014
B2K	0.952	0.006
CHE	0.407	0.113
CMD	0.207	0.047
GAK	0.748	0.014
KAL	0.941	0.030
KATMOJPRZ	0.918	0.012
KIEG	0.390	0.079
KLU	0.828	0.015
KUD	0.933	0.017
LUMICE	0.947	0.034
LUX_A	0.851	0.016
MKW	0.809	0.034
MOR	0.954	0.035
OST	0.781	0.016
OTW	0.830	0.010
PIO	0.626	0.098
PRZ	0.844	0.016
PUS	0.899	0.048
SL2	0.467	0.112
SLU	0.890	0.026
SRK	0.780	0.036
STG1	0.922	0.010
SZC	0.893	0.005
TER_L	0.682	0.025
TOR	0.938	0.015
WLU	0.805	0.015
Z04O	0.915	0.012
ZOZLO	0.782	0.048

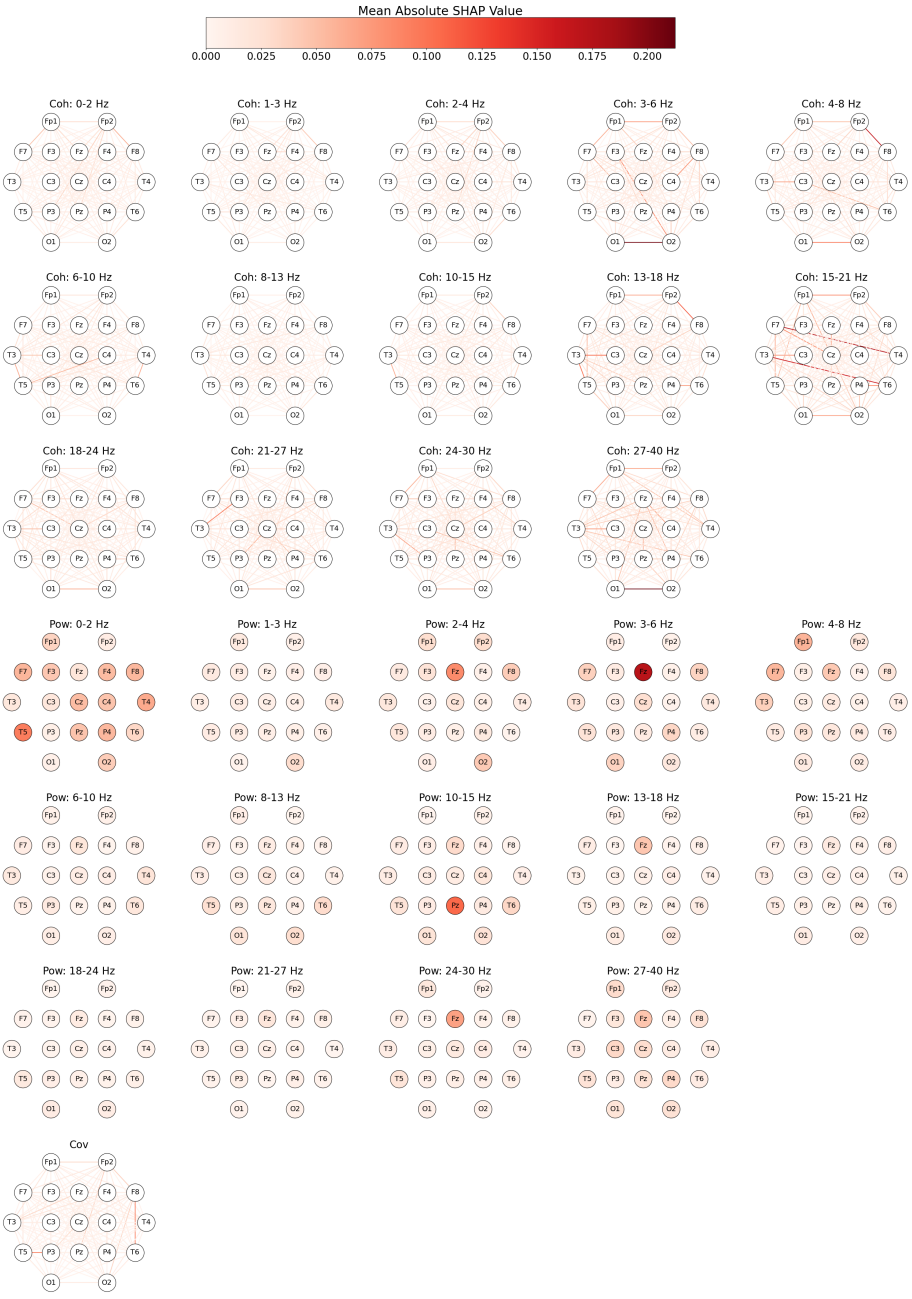


Figure A.2: Average absolute SHAP values for individual EEG features across all 30 ELM19 institutions.