

Motif Discovery Framework for Psychiatric EEG Data Classification

Melanija Kraljevska* Kateřina Hlaváčková-Schindler* Lukas Miklautz*
Claudia Plant *

Abstract

In current medical practice, patients undergoing depression treatment must wait four to six weeks before a clinician can assess medication response due to the delayed noticeable effects of antidepressants. Identification of a treatment response at any earlier stage is of great importance, since it can reduce the emotional and economic burden connected with the treatment. We approach the prediction of a patient response to a treatment as a classification problem, by utilizing the dynamic properties of EEG recordings on the 7th day of the treatment. We present a novel framework that applies motif discovery to extract meaningful features from EEG data distinguishing between depression treatment responders and non-responders. We applied our framework also to classification tasks in other psychiatric EEG datasets, namely to patients with symptoms of schizophrenia, pediatric patients with intractable seizures, and Alzheimer disease and dementia. We achieved high classification precision in all data sets. The results demonstrate that the dynamic properties of the EEGs may support clinicians in decision making both in diagnosis and in the prediction depression treatment response as early as on the 7th day of the treatment. To our best knowledge, our work is the first one using motifs in the depression diagnostics in general.

1 Introduction

Depression represents a common mental disorder that affects people globally. When diagnosing depression, medical practitioners check for the presence of the debilitating disease called Major Depressive Disorder (MDD). It is estimated that 3.8% of the population suffers from depression [1]. The treatments of depression include psychological treatment and antidepressant medication in more serious cases. However, the response rate of antidepressants in average, i.e., the percentage of cases of an improvement of depression symptoms, is in the range of 42-53 % [2]. A patient undergoing treatment needs to wait 4 to 6 weeks before getting checked by a clinician whether they are responding to the medication, as the antidepressants take time to produce noticeable effects of alleviated depressive symptoms [3]. To assess the effectiveness of the antidepressant treatment and monitor changes over

*Research Group Data Mining and Machine Learning, Faculty of Computer Science, University of Vienna, Währingerstrasse 29, 1090 Vienna, Austria and ds:UniVie, University of Vienna, Austria

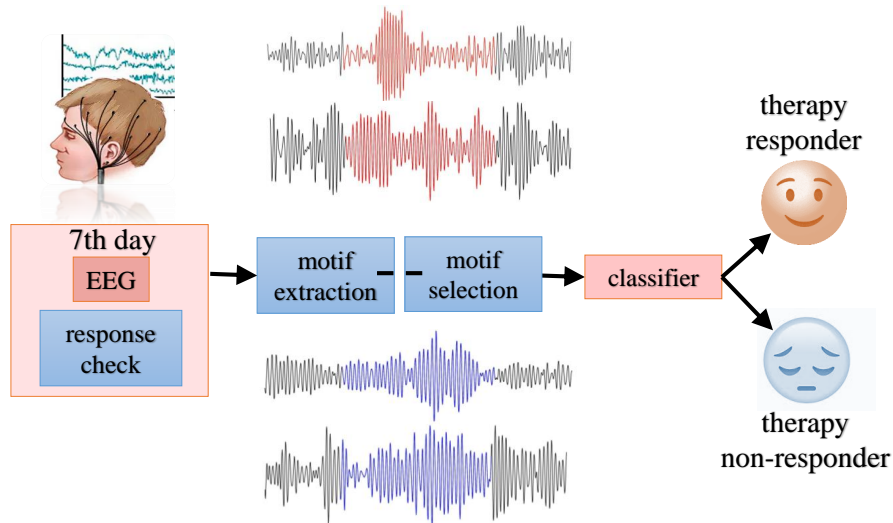


Figure 1: Overview on our treatment prediction framework

time, the Montgomery-Åsberg Depression Rating Scale (MADRS) [4] is used in practice. The MADRS questionnaire measures the severity of depression in the individual; a higher score indicates worse symptoms. If the MADRS score obtained after 4–6 weeks of treatment shows no improvement in the patient’s symptoms, then the treatment needs to be changed or adjusted. Due to the latency of the drug effect, in case of non-responsiveness to the medication, the patient can endure a considerable amount of distress and waste time on ineffective treatments. Electroencephalography (EEG) is a cost-effective clinical tool. To reduce the emotional and economic burden of a patient, in this work we propose a motif-based framework (see a schematic overview in Figure 1) which based on the EEG of the 7th day of treatment predicts whether a patient will respond to the medication or not.

We approach the prediction of patient response to treatment as a classification problem. In order to build a classifier for earlier prediction of responsiveness of treatment, features extracted from the EEGs of the 7th day of the treatment are used. For training, we use as class labels the MADRS score of the patient in the 28th day of the treatment. We perform the so-called motif discovery in EEGs of depressive patients and utilize the identified motifs in the feature engineering step, with the aim of differentiating between responsive and non-responsive patients.

The goal of motif discovery is to identify frequent, unknown patterns in a time series, without the pre-existing knowledge about their location and shape [5]. Motifs are defined in the literature as short time series that represent reoccurring patterns, frequent trends, or approximately repeated sequences [6][7][8]. In the case of EEG, motifs can correspond to

specific brain activities or states, e.g., different stages of being asleep can be differentiated using motifs detected in EEG sleep data [10]. There are several challenges when discovering motifs; as motifs represent patterns that are similar to each other, the similarity measure needs to account for possible noise and different scales, amplitudes, and variability of the patterns throughout the signal. Moreover, detecting motifs with unknown lengths requires the algorithm to be flexible enough to handle a wide range of possible lengths, which increases computational complexity in the case of high-dimensional or large-scale data [5]. We need to develop a reliable and interpretable classifier to distinguish treatment responders and non-responders, with as low-dimensional as possible representation of the input space. Another desired property of the classifier should be its scalability, meaning it can handle large, high-dimensional datasets and can be easily applied to new patients. This requires taking the computational efficiency of the motif discovery algorithms into account, as well as ensuring that the classifier can be easily integrated into existing psychiatric workflows and systems.

Our main contributions can be summarized as follows:

- We propose a novel framework including extracting discriminative motifs from EEG data and their usage in classification problem.
- We analyze the EEG recordings within alpha, beta, and theta frequency bands in separate experiments.
- To obtain the final set of meaningful motifs, we propose a motif selection criteria to rank the motifs based on their higher likelihood of distinguishing between the classes.
- We applied our framework to classification of depression treatment responders and non-responders, as well as to classification in other psychiatric EEG datasets, namely with patients with schizophrenia symptoms, patients with intractable seizures, Alzheimer disease and dementia. Figure 2 illustrates the examples of motifs pairs selected by our framework for each class label from each dataset.
- In all data sets we achieved high classification precision.
- To our best knowledge, our method is the first one using motifs to predict the treatment response of depression patients and in depression diagnostics in general.

2 Related Work

2.1 Classification Using EEG Data in MDD

Recently, researchers have been investigating EEG signals to discover characteristics or patterns related to a certain psychiatric disease. Such findings are often referred to as

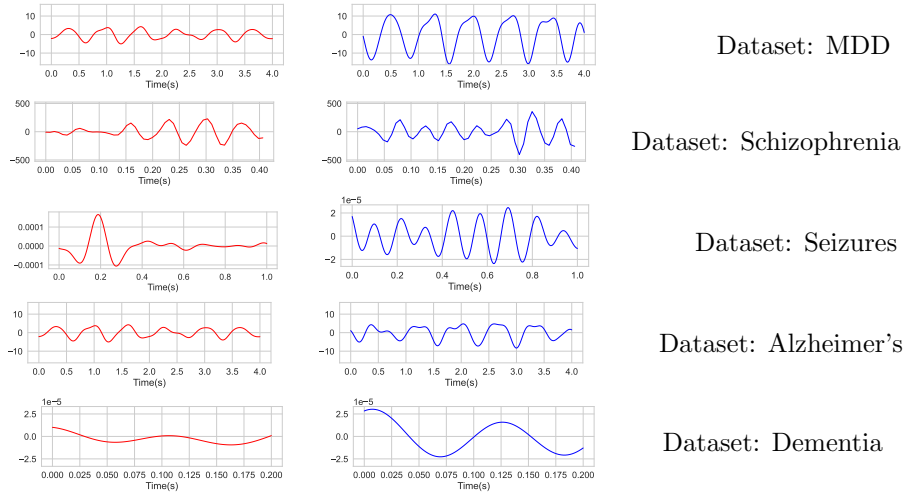


Figure 2: Example of motif pairs for each label (label 0, in red - non-responders for MDD and healthy subjects for other sets, and label 1 in blue - responders for MDD and pathology for other sets) from each dataset.

biomarkers, which can be useful in identifying the presence of a disease, in discovering pathophysiological mechanisms, and predicting outcomes of treatment. These classification problems require a suitable feature extraction technique that can discover informative features from the EEG signals. According to a review [12] that investigated EEG features for detecting MDD, EEG signals have been shown to have effective discriminative power over differentiating MDD patients and healthy subjects, In clinical research, EEG signals were initially utilized primarily for visual analysis of their spatial and temporal properties. More recent research focuses on investigating these properties within specific frequency bands, with the computation of more complex features. Ćukić et al. [13] present the effectiveness of Higuchi’s fractal dimension and Sample Entropy and explore multiple classification models, reaching an accuracy in the range of 90.24% to 97.56% on data set of 23 patients. Bučková et al. [14] used the same dataset of EEGs that we use but their objective was to classify biological sex from the EEG recordings. A review [15] of deep learning approaches for diagnosing MDD and bipolar disorder using EEG signals emphasizes that EEG-based methods have substantial potential for assessing and monitoring both diseases.

2.2 Predicting Response to MDD Treatment

Working with EEG signals often implies dealing with high- dimensional data, due to the number of electrodes, duration of recordings, and sampling rate. Althian et al. [16] investigated the impact of dataset size on the performance of common machine learning classifiers in the medical domain. They observed that the overall performance of classifiers depends

more on how much the extracted features represent the data rather than the data set size. [17] and [18] computed Granger-causal networks in different frequency bands on the same MDD patient data set we use. The classifier with the highest evaluation score was the decision tree classifier, with an F1 score 0.61. [19] used entropy-based measures on the same MDD set with classification accuracy 0.7 and an F1 0.65. Mueen et al. [25] proposed to use motifs as features for classification in disk-resident data. To our best knowledge, motifs as building blocks of features from EEGs have neither been used for predicting the diagnosis of depression, nor in predicting the outcome of the anti-depressive treatment.

2.3 Classification Using EEG Data for Other Psychiatric Diseases

There is a large number of publication on epileptic seizure detection by machine learning classifiers, see e.g. the recent review [20]. [21] used wavelet transform and deep network to classify in 16-channel EEG schizophrenia data basis on 84 patients from [45] (45 with schizophrenia and 39 healthy). [24] classifies 23 patients with fronto-temporal dementia, and 29 healthy patients and uses six common classifiers with the accuracy of 78.5-86.3% of correct detection. To our best knowledge, we are not aware of any publication using motif discovery on the data bases with these diseases. We used these data sets and tested our method on them. The results can be found in Section 5.5.

2.4 Motif Discovery Algorithms

Motif discovery algorithms vary based on the specific application. Some of the algorithms can identify exact or approximate motifs, motifs with fixed lengths, or variable lengths; Some algorithms find motifs in univariate and some in multivariate time series. Here we make a short overview of recent motif discovery algorithms and their applicability to our classification problem. Exact motif discovery refers to the process of identifying recurring patterns or motifs in time series data, where the discovered motifs are identical or nearly identical matches. Yeh et al. introduced an algorithm called STAMP [29] which utilizes a fast similarity search algorithm to find exact fixed-length motifs in a time series of length l . The authors further introduce STOMP [30] with reduced complexity. The algorithm SCRIMP++ [31] combines the features of both STOMP and STAMP. Schäfer et al. in [27] proposes an algorithm to find motifs, referred to as *Motiflets* - the set of exactly k occurrences of a motif of length l with minimal maximum pairwise distance. The authors argue that setting k is more intuitive and easier than setting the motif length l or the distance threshold between the motif occurrences r . The paper proposed two extensions to learn the input parameters k and l from the data and applied it in the experiment with EEG sleep signals. Two largest motif sets were found, which correspond to the well-known motifs in sleep EEG data (K-Complex and sleep spindles). For the superior precision over the compared more recent methods for this task, namely VALMOD [40], EMMA [41], SetFinder [42] and Learning Motifs [43], and the convenience of automatically determining

suitable values for the input parameters, we use this method in our motif discovery step.

3 Background

3.1 Motiflets

Discovering k-Motiflets can be achieved either by using the exact or approximate solution. Both solutions take as input the motif length l and the number of its occurrence k (previously determined by the proposed extensions of the method), and compute the k-Motiflets. Both the exact and approximate algorithm first generate the candidate motif sets by pairing each subsequence of a time series T with its non-overlapping $k-1$ nearest neighbors (NNs). Then the extent d is calculated for each of these sets, representing the maximum pairwise distance among the subsequences. The exact solution considers all subsets of subsequences of T of size k together with a heuristic on pruning to reduce the number of candidate sets, however, it obtains an exponential complexity on k . The approximate solution represents a greedy approach which computes the z-normalized pairwise distance matrix and applies pruning to limit candidate sets based on an upper bound on the current best extent d . This algorithm assumes that the NNs of a core subsequence are part of the final k-Motiflet, and returns the subsequences with the smallest d . For every candidate set where at least k subsequences are within d , it extracts the closest non-trivial subsequences and calculates their pairwise extent $dist$. Then, d is updated in case it is larger than the newly obtained $dist$. In the worst case, the complexity of the approximate solution is $O(kn^2) + O(nk^2)$ where n is the length of time series. The best case regarding complexity would be the case where the first subsequence is the top k-Motiflet and would allow pruning all further computations, leading to a complexity of $O(n^2) + O(k^2)$ [27].

Determining meaningful values for l and k are based on the extent function (EF), where $EF(k)$ returns the extent of the top k-Motiflet set, S_k , with fixed length l .

3.2 Learning meaningful k

Looking at the outputs of EF over increasing values of k , allows for inspecting the points with a significant increase of the slope of the line between consecutive values of k , known as *elbow points*. These points indicate that considering the next larger value of k causes a significant increase in the extent. The potential values for k are determined by using a threshold for detecting the elbow points. An example of the elbow method is shown in Figure 3, where the elbow points in 5, 9 and 17 reveal the top motifs.

3.3 Learning meaningful l

To determine l , the authors propose computing the normalized area under EF curve, AU_{EF} , which considers the list of $k-1$ extents for a given length l and the number of

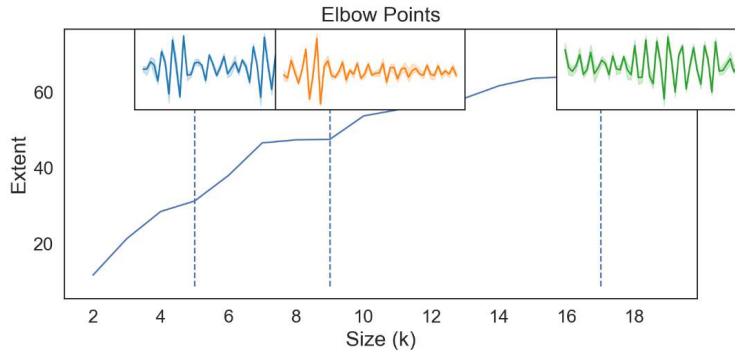


Figure 3: Example plot of the elbow method of choosing the top motifs from the electrode’s alpha band with a motif length of 2s, using the *plot_elbow* [9].

elbow points. The intuition is that elbow points are followed by typically steeper slopes, resulting in a larger AU_{EF} . Thus we want to choose the l minimizing the AU_{EF} score.

4 Proposed Method

Our primary objective was to propose a method workflow using k-Motiflets for a reliable classification in the most challenging data basis of MDD patients, compared to four mentioned data sets with other psychiatric conditions. We describe first this EEG data basis and then the method workflow. The details about the preprocessing and transformations of the data are in Appendix A.1. The patient database contains recordings of 176 patients who are treated for MDD with antidepressants. We use 10-min sessions of 19-channel EEG recordings per patient (resting state with the eyes closed), recorded on the 7th day of the treatment. The electrodes in the 10-20 standard EEG setup are placed on the patient’s scalp with channels: Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, P4, T3, T4, T5, T6, O1 and O2.

4.1 Labels

To evaluate the patient’s response to the treatment, psychiatric experts used the MADRS score and determined it before the treatment, and then on the 28th day after the start of the treatment. MADRS score ranges from 0 to 60 indicating the level of severity of the depression. The response to the treatment is determined based on the score difference between the two evaluations, indicating the level of improvement/worsening of the patient’s well-being. A patient is a responder (Label 1) if the MADRS score in 28th day reduced by 50% compared to the initial MADRS score, otherwise is a non-responder (Label 0). The dataset consists of 84 responders and 92 non-responders, 48 male and 128 female patients.

4.2 Frequency bands

Some research papers indicate that the alpha [34] [35], beta [14] and theta band [34] [36] contain representative biomarkers for depression detection or depression treatment outcome, hence the experiments are conducted on these frequency bands of each of the electrodes. We used the Python MNE library [11] to extract the frequency bands theta (4-8 Hz), alpha (8-12 Hz), and beta (12-30 Hz). A short segment of the extracted bands of an electrode can be found in Figure 9 in the Appendix A.1. The complete workflow from motif extraction to building a classifier is conducted for each frequency band separately and consists of several key steps: motif discovery for motif extraction using k-Motiflets, then selecting the motifs with the highest discriminatory power using the motif selection criteria, and then preparing the feature matrix for the classification step.

4.3 Motif Discovery

The main and most extensive part of the whole workflow process is the motif discovery for different frequency bands: alpha, beta and theta. We use the k-Motiflets algorithm together with its proposed heuristics for determining the motif length l and occurrence k . In the first step of the motif discovery, we conduct the search for the most suitable motif length l . This is done by using the *find_au_ef_motif_length* [9] method, which takes the list of potential motif lengths L as input, together with the maximum value of k , k_{max} . In order to pick an appropriate list of motif lengths for determining l , we carried out several initial experiments with different lengths, and settled for list of potential motif lengths ranging from 0.2 seconds to 8 seconds. As for k_{max} , we used 20, in accordance to the experiments conducted by the authors. Once l is obtained, we run the *search_k_motiflets_elbow* [9] method, which for each k up to k_{max} computes the EF and returns the list of motif set candidates for each k , together the distances for each occurrence, as well as the elbow points which point to the best candidates. The simplified pseudocode of this approach is presented in Algorithm 1.

Data: E^1, \dots, E^{19} , where E^i is a set of all i^{th} electrodes of all patients, L - set of potential motif lengths, k_{max} - maximum number of motif occurrences

Result: motifs[] - a list of extracted motifs, ks[] - a list of their corresponding k

```

motifs ← []
ks ← []
for i = 1 to 19 do
  for each e in Ei do
    l ← find_au_ef_motif_length(e, L, kmax)
    dists, candidates, elbow_points
    ← search_k_motiflets_elbow(e, l, kmax)
    for each motiflet in candidates[elbow_points] do
      k ← len(motiflet)
      idx ← motiflet[0]
      motif ← e[motif_idx : motif_idx + l]
      motifs.append(motif)
      ks.append(k)

```

Algorithm 1: **Motif extraction with k-Motiflets**

4.4 Motif selection criteria

After the motif discovery step, we obtain a large set of motifs. To filter out motifs that do not have any discriminatory power to differentiate between responders and non-responders, we need to check which motifs appear to be typical for one class, i.e., are not common patterns for both classes. To assess this, we need to check the presence of a motif within a given signal. This involves computing the closest subsequence within the signal, as well as identifying the nearest neighbors to this subsequence - we referred to these closest subsequences to the motif as the *motif matches*. For identifying the closest match of a motif within a time series, we use the *match* method within Python library STUMPY [32]. The method uses the MASS [29] algorithm and provides a fast way to compute the sliding window dot product between the query subsequence (in our case the motif) and all the subsequences in the given time series. We use the method to obtain a list of k closest matches, including the z-normalized Euclidean distance between the motif and the match, see an example in Figure 4. The maximum distance for which a subsequence in time series T of length n is considered a match for a given motif Q of length m is defined as: $f(D) = \max((\text{mean}(D) - 2 \cdot \text{std}(D)), \min(D))$ where D is an array with a size of $n - m + 1$ and represents the distance profile of Q with T . Hence, function $f(D)$ returns at least the closest match.

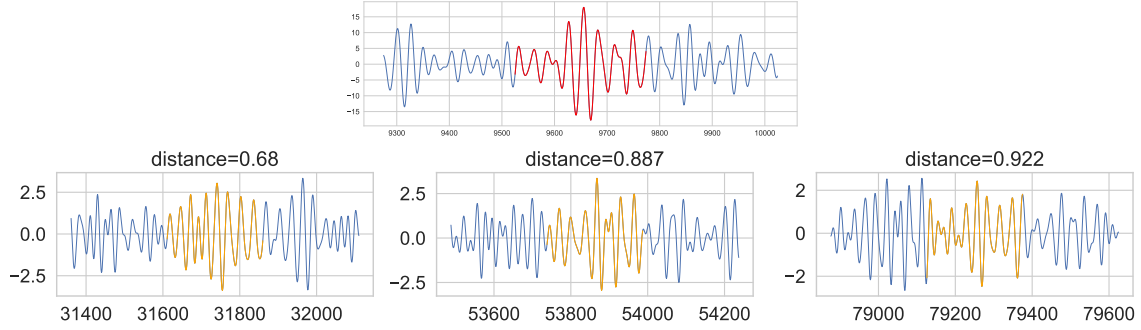


Figure 4: An example output of *match* returning the distances to the best matches (orange) to a motif (red).

Data: Q - motif, E^j - list of all j^{th} electrodes, L - binary list of corresponding labels, *percentage* - the percentage of lowest distances to consider

Result: $score_{diff}$

$distances \leftarrow \text{dict}()$

$distances[0] \leftarrow []$

$distances[1] \leftarrow []$

for $i = 1$ to $\text{len}(E^j)$ **do**

$electrode \leftarrow E^j[i]$

$label \leftarrow L[i]$

$distances_{matches} \leftarrow \text{match}(Q, electrode)$

$distances[label].append(\text{mean}(distances_{matches}))$

$best_distances_0 \leftarrow \text{get_best}(distances[0], \text{percentage})$ $best_distances_1 \leftarrow$

$\text{get_best}(distances[1], \text{percentage})$ $score_{diff} \leftarrow \text{abs}(best_distances_0 - best_distances_1)$

Algorithm 2: Compute difference score of a motif

For checking the motif presence, we have to consider the distances of the subsequences within the motif match to the given motif and decide on a suitable threshold for determining which distance (or average distance) is acceptable for considering the found match represents an instance of the motif. Finding such a threshold is difficult, as it may also vary depending on the motif length and the frequency band. Therefore, instead of a threshold, we use the average distances within the match to represent how distant is the found match to the given motif. Motifs with a greater discriminatory power between the two classes are expected to have close matches in one class and more distant matches within the other class. To quantify this discriminatory power, we compute the *difference score* in Algorithm 2 by considering the distances of the motif to its matches within each class. If there is a greater difference in the average distances between the classes, that could indicate that

the motif is mainly present in one of the classes. For computing the average distance of a motif to its class, instead of considering the average match distances to all patients of the class, we consider a *percentage* of the lowest average match distances. This allows for a low difference score for motifs that are typical for a subset of patients within one class. It is important to mention that when computing this score, we consider only the electrodes that correspond to the motif’s electrode of origin, i.e. if the motif originates from a patient’s j^{th} electrode, we consider all j^{th} electrodes.

Using these scores significantly reduces the pool of motif candidates. In order to have a balanced representation of each class and gender, we choose the same number of motifs for each class-gender combination. Hence, we preserve the best n motifs for each class-gender combination based on the difference score.

4.5 Features

The feature matrix consists of rows - patients and the columns - the motifs. Patient i is represented as a d -dim vector v , where d is the number of motifs and v_j is the average distance of the closest matches to the j^{th} motif. Similarly as above, for computing this value we use method *match*, considering the patient’s electrode that corresponds to the motif. Essentially, the feature matrix represents a distance matrix.

4.6 Classification

For predicting the class label, i.e., the treatment outcome of the patients, we use several common classification methods. To combat overfitting, we have chosen simpler methods, whose hyperparameters allow for regularization. Interpretability is also a crucial property in decision-making in the medical domain. Considering both preferences, we selected the support vector machine (SVM) with two different kernels, decision tree, random forest, logistic regression, and multi-layer perceptron from the scikit-learn library [37]. The selection of the hyperparameters of these classifiers can be found in Appendix B.2. As for the performance evaluation we use the F1 measure as it accounts for class imbalances as well as accuracy. To further reduce the feature space and retain the most important motifs, we use the wrapper method Recursive Feature Elimination (RFE) [38], which iteratively removes the least important feature until the optimal number of features is attained based on a prediction metric.

5 Experiments

The experiments consist of the identification and selection of discriminatory motifs for a concrete data set and training and evaluation of the classifiers. Our Python code can be found under <https://anonymous.4open.science/r/motif-discovery-eeg-EF02/>.

5.1 Motif Discovery

The experiments were conducted on frequency bands beta, alpha, and theta and on different motif lengths. Due to the large number of extracted motifs for each combination of frequency band and electrode, for each of the motifs we compute the difference score (Algorithm 2). An example of the score of a motif from class 0 in the depression data basis is depicted in Figure 5 and Figure 6, where a motif for the beta band is shown together with its three closest matches from each class (for readability reasons, we do not show all of the matches within one signal, but rather only the closest match). The motifs with higher difference scores in one class indicate that they have close matches (measured with the normalized Euclidean distance) within the class and more distant matches in the other class. Figure 6 is an example of a motif with a score close to 0, which indicates that there are close matches in both classes, therefore the motif might not be a good candidate for discriminating between the classes.

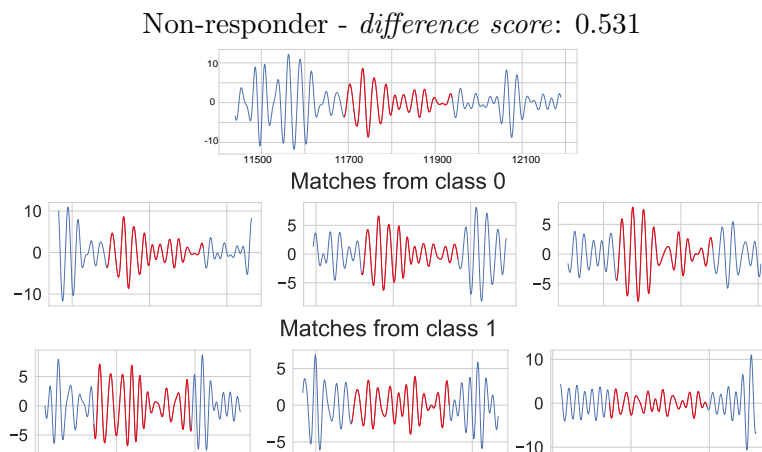


Figure 5: Example of a motif from the beta band and with a higher difference score, together with three examples matches across the two classes.

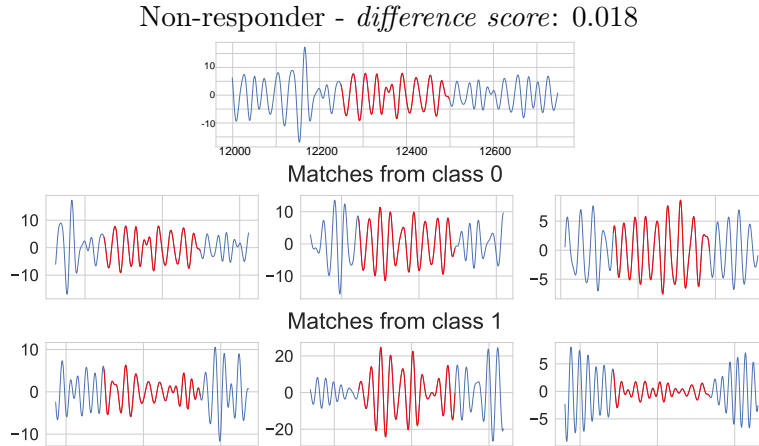


Figure 6: As in Fig 5 with a lower difference score.

5.2 Classification

We construct a feature matrix with a subset of the obtained motifs, having the largest discriminatory power, i.e. difference score.

5.3 Imbalance

The dataset is imbalanced with respect to the class label and gender. Thus, it is of crucial importance to have each combination of class and gender equally represented in the feature matrix. When selecting a smaller subset of the motifs, we choose the same number of motifs for each combination. We would like the number of motifs to be as low as possible, to avoid overfitting, but on the other hand, we would like to include as much information as possible. Hence, after conducting several experiments, we decided to start with the best 20 motifs for each combination (ranked by the difference score) to construct the (initial) feature matrix. We additionally use a feature selection technique to further reduce the feature space and keep the most descriptive motifs. To obtain a better overview of the constructed feature space, we project the obtained patient profiles to a lower-dimensional space using Linear Discriminant Analysis (LDA) [39]. If the projected points from different classes are well-separated in the lower-dimensional space, it indicates that LDA has successfully captured the discriminative information, and the classes are easily distinguishable. This suggests that a simple linear classifier (e.g., logistic regression or linear SVM) trained on the transformed data is likely to perform well. For each of the bands and motif lengths, we visualize the values from the obtained projection to gain further insights into possible linear separability between the two classes.

The boxplot on the left of Figure 7 represents the distribution of the projected values among the two classes, while the scatter plot on the right depicts the projected values,

separated by gender. One can see that there is some degree of linear separability between the classes (the corresponding plots for beta and theta bands are in Appendix B.1).

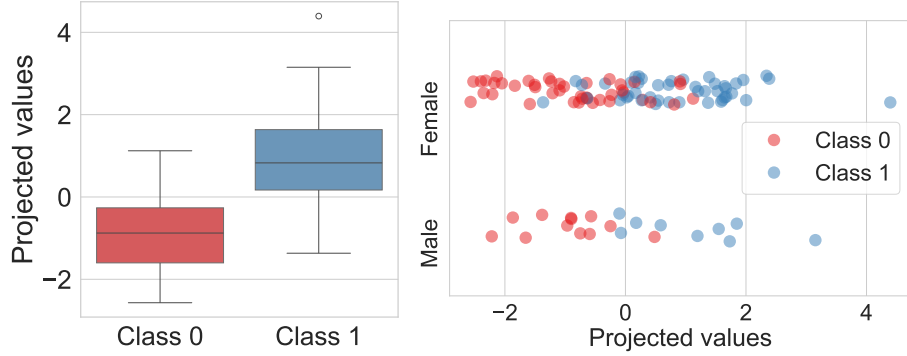


Figure 7: Example of projected values of the feature matrix for the alpha band.

5.4 Evaluation

To have a fair evaluation of the classifiers, we apply a 5-fold cross validation (CV) on the entire dataset of 176 patients and use mean F1 score of the F1 scores obtained from each validation fold. Figure 8 shows the mean F1 from CV for each of the three frequency bands, where each box plot depicts the 5 pairs of training and validation sets. One can see that the training F1 scores are quite stable, the classifiers learn and generalize well, and the validation scores behave similarly to the training scores. Balanced scores often result from models that have learned meaningful features and relationships within the data. Similar plots showing the accuracy results for all frequency bands can be found in Appendix B.3.

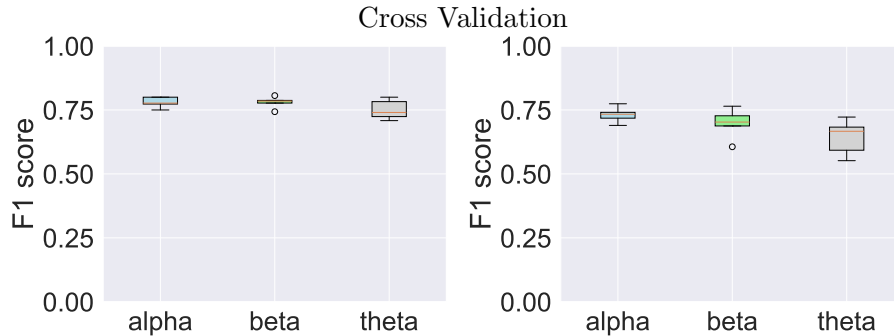


Figure 8: Mean F1 of training (left) and validation (right) for all three freq. bands of the best classifier.

Table 5.4 contains the results for each band, obtained by the best classifier (ranked by

the highest mean F1 score on the 5-fold CV). One can see that all three bands have similar scores and also scores are similar for both training and validation sets.

Band	Training		Validation	
	Accuracy	F1	Accuracy	F1
Alpha	0.785	0.78	0.731	0.722
Beta	0.751	0.779	0.673	0.713
Theta	0.738	0.751	0.62	0.647

Table 1: Training and validation precision in MDD set for each freq. band - mean accuracy and F1 from CV.

Best performing classifier Overall, the best performing model is the decision tree classifier using the decision rule criterion Gini and a maximum depth of 3. The best score is obtained on the alpha band with a mean validation F1 of 0.722 and an validation accuracy of 0.731.

Feature importance The final feature sets obtained from the feature selection step, contain 8 motifs for the alpha band (4 for responders, 4 for non-responders), and 5 motifs for both beta and theta band (with 1 for non-responders and 4 for responders). The motifs from the electrode O2 seem to be the most dominant for the alpha band, having 4 of the motifs originating from this channel. The second most important electrode from which originate 2 motifs, is the Fz channel.

5.5 Evaluation of the Workflow on Other Psychiatric Data Sets

We evaluated the proposed workflow on four publicly available EEG datasets of patients with different psychiatric conditions and their results are summarized in Table 2. The dataset of 16-channel EEG of healthy adolescents and those with symptoms of schizophrenia from [45] contains 84 patients in total (45 with schizophrenia and 39 healthy). Our framework found 9 motifs for the healthy and 11 motifs for the schizophrenia group. The dataset from [46] contains multiple EEG recordings of 23 pediatric patients with intractable seizures. For classifying the presence of seizures in a recording, we used two recordings per patient (one containing a seizure and one without), having 46 samples in total. Our framework found 1 motif for healthy and 3 motifs for recordings with seizures. The dataset from [47] contains EEG recordings of 88 patients, having 36 diagnosed with Alzheimer’s disease, 23 with fronto-temporal dementia, and 29 healthy patients. We used both groups of patients with Alzheimer’s and dementia separately to train two binary classifiers that distinguish healthy patients from patients with one of the diseases. Our framework found 10 motifs for Alzheimer’s group and 6 motifs for the control group, and 8 motifs for dementia and 4 motifs for the control group. Results for other frequency bands for each data set can be found in Appendix B.4.

Dataset	Band	Training		Validation	
		Accuracy	F1	Accuracy	F1
Schizophrenia	Alpha	0.927	0.931	0.820	0.821
Seizures	Theta	0.834	0.831	0.787	0.784
Alzheimer	Theta	0.868	0.847	0.845	0.823
Dementia	Theta	0.954	0.951	0.87	0.865

Table 2: Evaluation on other psychiatric datasets on the band with the highest mean F1 on CV

One can see from Table 2 that our method achieves in average comparable or even better classification precision on all these data sets than on the MDD data set.

6 Discussion

The best classification precision on our MDD data closely aligns with a recent meta-analysis review [44] on the prediction of treatment response using EEG in MDD, where the best performing classifier achieves an accuracy score in the range of 0.782 to 0.826. The approaches used in the review paper have a median size of 86.5, and 60% of them use CV to report on the final results. However a direct comparison of the precision values is problematic, since our database is three times bigger than the data sets in this paper.

6.1 Comparison of classification precision on MDD to other psychiatric data sets

We postulate that the relatively lower classification precision on the MDD data set is not in the method but in the data set. Firstly, the MDD data set is compared to the other data sets larger, which can influence the classification precision. And secondly, the classification precision is also influenced how correctness of the labeling. The class labelling in the MDD data set is done by the MADRS questionnaire, filled in subjectively by the patients on the 28th day of the treatment. On the other hand, the class labelling in the data sets with schizophrenia, seizures and Alzheimer conditions were done by medical professionals. Thus the lower classification in the MDD set might be related to the fact that a correct labelling for depression treatment classification is a real challenge.

There are some limitations of the approach: The experimental workflow analyzes motifs up to 8 seconds. Also, we trained the classifier using motifs within the same frequency band. The possibility of combining them is limited, since for filtering the motifs that are good candidates we use the proposed difference score. The difference score is based on the distances calculated by the z-normalized Euclidean distance metric, which scales with the length of the motif. In addition, the signals are different within each frequency band,

hence in order to be able to do a fair ranking of the motifs across all bands and lengths, the difference score has to be invariant to these two parameters.

7 Conclusions and Future Work

We proposed a novel framework including feature selection, motif discovery and classification to be used in general for classification of psychiatric patients based on their EEG. The best classification precision was obtained for data sets with schizophrenia and dementia which in both cases have training score in F1 and accuracy above 0.92 and by testing above 0.82. Accounting not only for the class imbalance but also for the gender imbalance proved to be crucial in achieving good performance in the case of MDD. In our proposed workflow we represent each patient group (class and gender) by the same number of motifs when computing the feature matrix. We have also considered the gender information in the experiments in the other psychiatric data sets where it was available, i.e., for the Alzheimer’s disease and dementia.

Possible future work The motif ranking computation, i.e., the difference score, represents a complex problem, given that it depends on several parameters. The z-normalized Euclidean distance, used by the difference score uses, is not appropriate for comparison of similarity between pairs of time series of different lengths. In future we want to use a length invariant score, as well as to explore the frequency bands gamma and delta and to investigate longer motifs that can also differentiate the two patient groups. Given that EEG channels are recorded simultaneously, one can consider motif discovery as a multi-dimensional motif search, having each electrode as a separate dimension. For this case, other motif discovery algorithms can be selected.

Acknowledgements

This work was supported by the Austrian Science Foundation FWF (project I5113).

References

- [1] World Health Organization. 2023. Depressive disorder (depression). <https://www.who.int/news-room/fact-sheets/detail/depression>. Accessed: 29.04.2023.
- [2] D. Taliaz, A. Spinrad, R. Barzilay, Z. Barnett-Itzhaki, D. Averbuch, O. Teltsh, R. Schurr, S. Darki-Morag, and B. Lerer. 2021. *Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data*. *Translational Psychiatry* 11, 1 (2021), 381.
- [3] S. Gautam, A. Jain, M. Gautam, V.N. Vahia, and S. Grover. 2017. Clinical practice guidelines for the management of depression. *Indian Journal of Psychiatry* 59, Suppl 1 (2017), S34.

- [4] S. A. Montgomery and M. Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry* 134, 4 (1979), 382–389.
- [5] S. Torkamani and V. Lohweg. 2017. Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 2 (2017), e1199.
- [6] B. Chiu, E. Keogh, and S. Lonardi. 2003. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 493–498.
- [7] A. Udechukwu, K. Barker, and R. Alhajj. 2004. Discovering all frequent trends in time series. In *Proceedings of the Winter International Symposium on Information and Communication Technologies*. 1–6.
- [8] D. Minnen, C. Isbell, I. Essa, and T. Starner. 2007. Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 601–606.
- [9] P. Schäfer, U. Leser. Motiflets - Simple and Accurate Detection of Motifs in Time Series. GitHub repository, <https://github.com/patrickzib/motiflets>.
- [10] J. Kohlmorgen, K.R. Müller, J. Rittweger, and K. Pawelzik. 2000. Identification of non-stationary dynamics in physiological recordings. *Biological Cybernetics* 83, 1 (2000), 73–84.
- [11] A. Gramfort, M. Luessi, E. Larson, D.A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al. 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 267 (2013).
- [12] C. Greco, O. Matarazzo, G. Cordasco, A. Vinciarelli, Z. Callejas, and A. Esposito. 2021. Discriminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Systematic Review. *IEEE Access* 9 (2021), 112850–112870.
- [13] M. Čukić, M. Stokić, S. Simić, and D. Pokrajac. 2020. The successful discrimination of depression from EEG could be attributed to proper feature extraction and not to a particular classification method. *Cognitive Neurodynamics* 14 (2020), 443–455.
- [14] B. Bučková, M. Brunovský, M. Bareš, and J. Hlinka. 2020. Predicting sex from EEG: validity and generalisability of deep-learning-based interpretable classifier. *Frontiers in Neuroscience* 14 (2020), 589303.
- [15] A. Cipriani, T.A. Furukawa, G. Salanti, A. Chaimani, L.Z. Atkinson, Y. Ogawa, S. Leucht, H.G. Ruhe, E.H. Turner, JPT Higgins, et al . 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Focus* 16, 4 (2018), 420–429.

- [16] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A.B. Dris, N. Alzakari, A. Abou El-wafa, and H. Kurdi. 2021. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences* 11, 2 (2021), 796.
- [17] K. Hlaváčková-Schindler, C. Pacher, C. Plant, M. Lazarenko, M. Paluš, J. Hlinka, M Brunovský. 2023. Pattern Discovery in an EEG Database of Depression Patients: Preliminary Results. In *2023 14th International Conference on Measurement* (pp. 80-83). IEEE.
- [18] C. Pacher. 2023. Analysis of an EEG database of depression patients by means of graphical Granger causality. Master thesis, Faculty of Computer Science, University of Vienna.
- [19] A. Stähle. 2024. Major depressive disorder treatment prediction using entropy-based measures. Master thesis, Faculty of Computer Science, University of Vienna.
- [20] Siddiqui, M. K., Morales-Menendez, R., Huang, X., Hussain, N. 2020. A review of epileptic seizure detection using machine learning classifiers. *Brain Informatics*, 7(1), 5.
- [21] Sairamya, N. J., Subathra, M. S. P., and George, S. T. 2022. Automatic identification of schizophrenia using EEG signals based on discrete wavelet transform and RLNDiP technique with ANN. *Expert Systems with Applications*, 192, 116230.
- [22] F.F. Offner. 1950. The EEG as potential mapping: the value of the average monopolar reference. *Electroencephalography and Clinical Neurophysiology* 2, 213–214.
- [23] A. Delorme, S. Makeig. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods* 134, 1, 9–21.
- [24] Miltiadous, A., Tzamourta, K. D., Giannakeas, N., Tsiouras, M. G., Afrantou, T., Ioannidis, P., and Tzallas, A. T. 2021. Alzheimer’s disease and frontotemporal dementia: A robust classification method of EEG signals and a comparison of validation methods. *Diagnostics*, 11(8), 1437.
- [25] A. Mueen, E. Keogh, and N. Bigdely-Shamlo. 2009. Finding time series motifs in disk-resident data. In *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 367–376.
- [26] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. 2009. Exact discovery of time series motifs. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 473–484.

- [27] Schäfer, Patrick and Leser, Ulf, 2022. Motiflets: Simple and Accurate Detection of Motifs in Time Series. In Proceedings of the VLDB Endowment, 725–737.
- [28] Y. Li, H.U. Leong, M.L. Yiu, and Z. Gong. 2015. Quick-motif: An efficient and scalable framework for exact motif discovery. In 2015 IEEE 31st International Conference on Data Engineering. IEEE, 579–590.
- [29] C.C.M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H.A. Dau, D.F. Silva, A. Mueen, and E. Keogh. 2016. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 1317-1322.
- [30] Y. Zhu, Z. Zimmerman, N.S. Senobari, C.C.M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh. 2016. Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 739-748.
- [31] Y. Zhu, C.C M. Yeh, Kamgar K. Zimmerman, Z., and E. Keogh. 2018. Matrix profile XI: SCRIMP++: time series motif discovery at interactive speeds. In 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 837–846.
- [32] S.M. Law. 2019. STUMPY: A powerful and scalable Python library for time series data mining. *Journal of Open Source Software* 4, 39 (2019), 1504.
- [33] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh. 2013. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery* 26 (2013), 275–309.
- [34] A.S. Widge, M.T. Bilge, R. Montana, W. Chang, C.I. Rodriguez, T. Deckersbach, L.L. Carpenter, N.H. Kalin, and C.B. Nemeroff. 2019. Electroencephalographic Biomarkers for Treatment Response Prediction in Major Depressive Illness: A Meta-Analysis. *American Journal of Psychiatry* 176, 1 (2019), 44–56. <https://doi.org/10.1176/appi.ajp.2018.17121358> arXiv:<https://doi.org/10.1176/appi.ajp.2018.17121358> PMID: 30278789
- [35] L. Zhao, L. Yang, B. Li, Z. Su, and C. Liu. 2021. Frontal alpha EEG asymmetry variation of depression patients assessed by entropy measures and Lemple–Ziv complexity. *Journal of Medical and Biological Engineering* 41 (2021), 146–154.
- [36] M. Arns, A. Etkin, U. Hegerl, L.M. Williams, C. DeBattista, D.M. Palmer, P.B. Fitzgerald, A. Harris, R. deBeuss, and E. Gordon. 2015. Frontal and rostral anterior cingulate (rACC) theta EEG in depression: Implications for treatment outcome? *European Neuropsychopharmacology* 25, 8 (2015), 1190–1200.

- [37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research* 12 (2011), 2825-2830.
- [38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46 (2002), 389-422.
- [39] R.A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 2 (1936), 179-188.
- [40] Michele Linardi, Yan Zhu, Themis Palpanas, and Eamonn Keogh. 2018. VALMOD: A suite for easy and exact detection of variable length motifs in data series. In *Proceedings of the 2018 International Conference on Management of Data*. 1757- 1760.
- [41] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel. 2002. Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*. 53-68.
- [42] Anthony Bagnall, Jon Hills, and Jason Lines. 2014. Finding motif sets in time series. *arXiv preprint arXiv:1407.3685* (2014)
- [43] Josif Grabocka, Nicolas Schilling, and Lars Schmidt-Thieme. 2016. Latent timeseries motifs. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 1 (2016), 1-20
- [44] D. Watts, R.F. Pulice, J. Reilly, A.R. Brunoni, F. Kapczinski, and I.C. Passos. 2022. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational Psychiatry* 12, 1 (2022), 332.
- [45] N.N. Gorbachevskaya and S.V. Borisov. 2019. EEG of healthy adolescents and adolescents with symptoms of schizophrenia. http://http://brain.bio.msu.ru/eeg_schizophrenia.htm
- [46] J. Guttag. 2010. CHB-MIT Scalp EEG Database (version 1.0.0). <https://physionet.org/content/chbmit/1.0.0/>
- [47] A. Miltiadous, K.D. Tzamourta, T. Afrantou, P. Ioannidis, N. Grigoriadis, Angelidis P. Tsalikakis, D.G., M.G. Tsipouras, E. Glavas, N. Giannakeas, and A.T. Tzallas. 2023. A dataset of EEG recordings from: Alzheimer’s disease, Frontotemporal dementia and Healthy subjects. <https://openneuro.org/datasets/ds004504/versions/1.0.6>.

Appendix

A More details to the EEG data sets

A.1 Preprocessing of the EEG Signal

The initial preprocessing of the EEG signals has been provided by the project members. The preprocessing was done in MATLAB using the open-source toolbox EEGLAB [23]. The following steps were performed:

- Downsample the signals with 1000 Hz sampling rate to 250 Hz, by keeping every 4th sample.
- Remove the first and last 30 seconds of the signal, as this period can contain a high number of artifacts.
- Use the Average Reference method [22] to transform the EEG recordings.

(An EEG signal quantifies the electrical potential difference between the recording electrode and a reference electrode. In the Average Reference method, the reference electrode represents the average signal across all electrodes.)

- Apply a bandpass filter to keep only frequencies from 1 to 40 Hz.
- Remove segments of a signal that contain high-power artifacts.
(The segments are determined by a window with no overlap and the ones containing problematic data are removed entirely. The length of the window is set to 2 seconds, as this ensures that every 2-second segment of the signal is continuous.)
- Since we are using each frequency band separately, we further downsample each signal for each band to its own Nyquist rate to save computation time. This means downsampling the theta, alpha, and beta bands to 16, 24, and 60 Hz respectively.

A.2 Class and Gender Distribution

	Male	Female	Total
Responders	17	66	83
Non-responders	31	62	93
Total	48	128	176

Table 3: Class and gender distribution of the patients in the EEG database

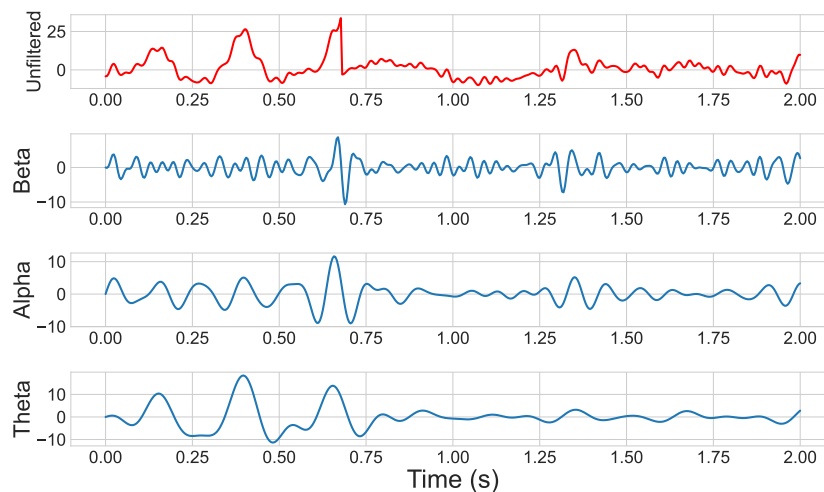


Figure 9: Short segment of the EEG signal in channel Fp1 and the extracted frequency bands: beta, alpha and theta.

B Details to the experiments

B.1 Feature separability

B.2 Hyperparameter tuning

For choosing the most optimal hyperparameters for the model, a 5-fold cross-validation is used. Each model had the following hyperparameters that were tuned:

- SVM
 - kernel: ["linear", "rbf"]
 - C: [0.0005, 0.001, 0.01, 0.1, 0.5]
 - penalty: ["l1", "l2"]
- Decision Tree
 - criterion: ["gini", "log_loss"]
 - max_depth: [3, 4, 5, 10, 20]
- Random Forest
 - n_estimators: [5, 10, 15, 20]
 - max_depth: [3, 4, 5, 10, 20]
 - min_samples_leaf: [1, 2, 3, 4]

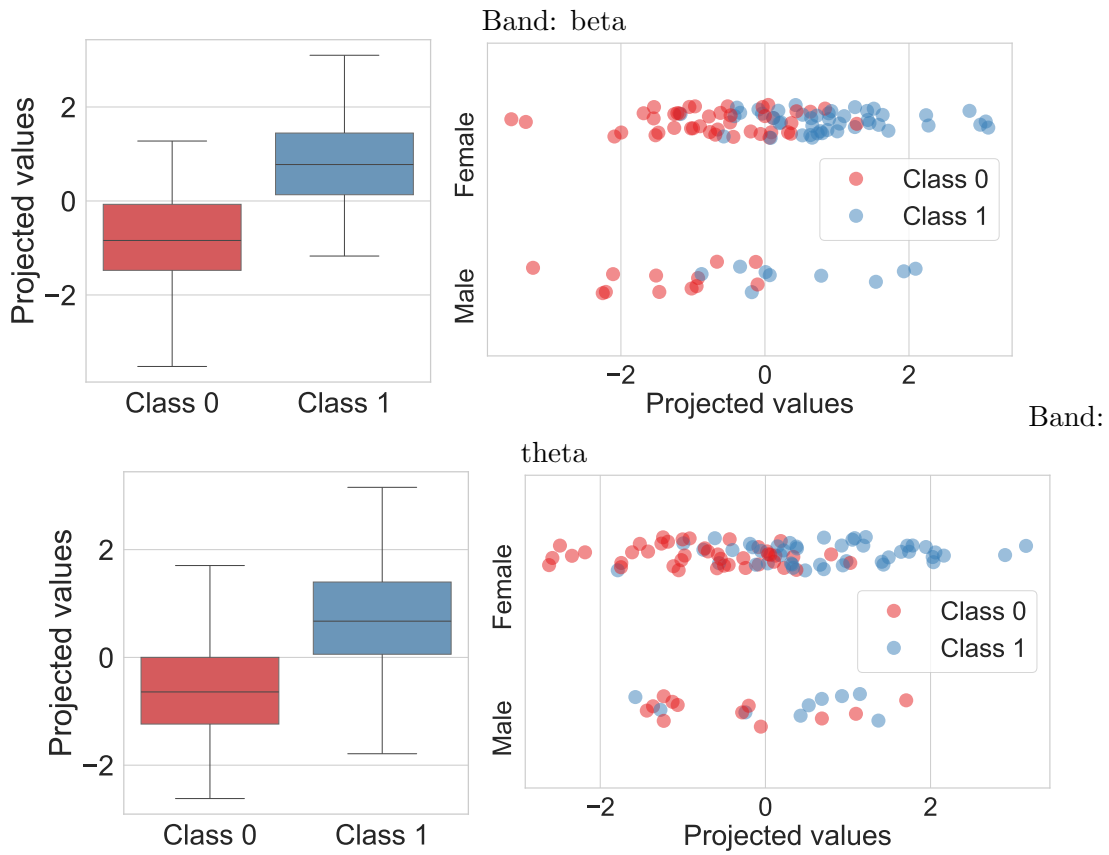


Figure 10: Examples of the distribution of the projected values of the feature matrix of the beta and theta band.

- min_samples_split: [2, 3, 4]
- Logistic Regression
 - C: [0.1, 0.5, 0.7, 1]
 - penalty: ["l1", "l2"]
- Multi-layer perception
 - alpha: [0.01, 0.05, 0.1, 0.5]
 - hidden_layer_sizes: [(5, 10, 5), (10, 10), (20, 10, 20), (20, 20)]

B.3 Cross Validation

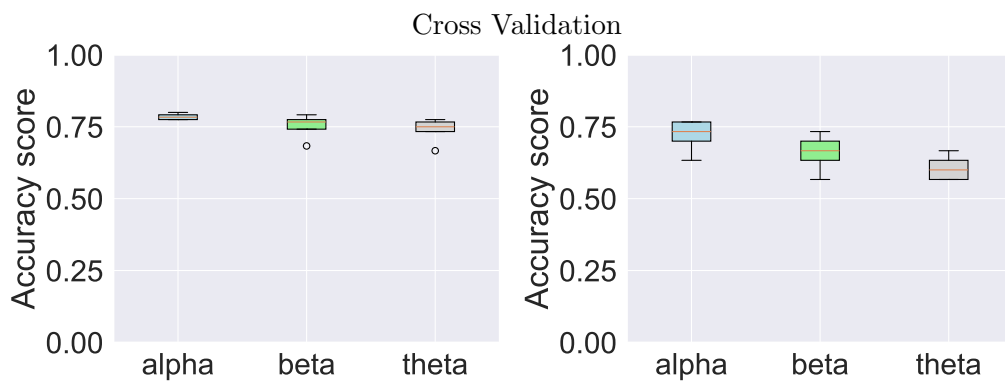


Figure 11: Mean Accuracy scores of training (left) and validation (right) for all three frequency bands of the best performing classifier.

B.4 Precision on other psychiatric data sets

The results can be found in Table 4.

Table 4: The best evaluation results on other psychiatric data sets within each band.

Dataset	Band	Training		Validation	
		Accuracy	F1	Accuracy	F1
Schizophrenia	alpha	0.927	0.931	0.82	0.821
	beta	0.898	0.907	0.857	0.877
	theta	0.767	0.799	0.762	0.789
Seizures	alpha	0.875	0.875	0.675	0.617
	beta	0.711	0.749	0.624	0.681
	theta	0.834	0.831	0.787	0.784
Alzheimer's	alpha	0.896	0.88	0.8	0.774
	beta	0.9	0.886	0.861	0.845
	theta	0.868	0.847	0.845	0.823
Dementia	alpha	0.797	0.827	0.808	0.835
	beta	0.971	0.974	0.807	0.82
	theta	0.954	0.951	0.87	0.865