universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Classification of treatment response in depression patients using motif discovery“

verfasst von / submitted by

Melanija Kraljevska

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2024 / Vienna, 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 645
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Data Science UG2002
Betreut von / Supervisor:	Univ.-Prof. Dipl-Inform. Univ. Dr. Claudia Plant
Mitbetreut von / Co-Supervisor:	RNDr. CSc. Katerina Schindlerova

Acknowledgements

My sincere gratitude goes to my supervisors Prof. Claudia Plant and Dr. Katerina Schindlerova (Kateřina Hlaváčková-Schindler) for their support in the creation of this thesis. I deeply appreciate the insights and dedication of Dr. Schindlerova, which were crucial in the process of the thesis development. I am immensely thankful to Dr. Lukas Miklautz, for his invaluable feedback and guidance, which significantly contributed to the success of this work.

I want to express my deepest gratitude to my family, for their encouragement during this journey. A heartfelt thank you goes to Pavel, for all his love and constant belief in me.

This work is part of the international project "Learning Synchronization Patterns in Multivariate Neural Signals for Prediction of Response to Antidepressants", together with the Data Mining and Machine Learning Research Group of the University of Vienna and the Czech Academy of Sciences.

Abstract

In recent years, researchers have become increasingly interested in the utilization of EEG signals to discover characteristics and patterns that relate to a certain psychiatric disease. Patients undergoing depression treatment must wait four to six weeks before a clinician assesses medication response due to the delayed noticeable effects of antidepressants. The identification of treatment response closer to its start has the potential of introducing several benefits for people suffering from depression, by reducing the emotional and economic burden of depression patients, as treatments that are not effective can be replaced weeks earlier.

In this thesis, we approach the prediction of patient response to treatment as a classification problem, by utilizing the dynamic properties of EEG recordings of depression patients undergoing antidepressant treatments. We investigate the application of state-of-the-art motif discovery algorithms SCRIMP++ and OSTINATO to EEG recordings and propose a workflow from motif extraction to building a classifier with high predictive performance. Motifs with different lengths are extracted from three frequency bands: alpha, beta and theta, and used as features in simpler and more interpretable models. The proposed feature extraction process consists of motif selection criteria and handling of class and gender imbalances. The database consists of 176 patients in total and is divided into a training and a separate testing set. We investigated four classifiers, out of which the SVM classifier had the best performance with an accuracy score of 0.738 and an F1 score of 0.744 on the testing set. The results demonstrate that the dynamic properties of the EEGs potentially hold information that could aid in discriminating between responders and non-responders.

This master thesis is part of the international research project "Learning Synchronization Patterns in Multivariate Neural Signals for Prediction of Response to Antidepressants", a joint international project by the University of Vienna, the Czech Academy of Sciences, and the National Institute of Mental Health of the Czech Republic.

Kurzfassung

In den letzten Jahren hat das Interesse der Forschung an der Nutzung von EEG-Signalen zur Entdeckung von Merkmalen und Mustern in Verbindung mit einer bestimmten psychiatrischen Erkrankung zugenommen. Patienten, die sich einer Depressionsbehandlung unterziehen müssen vier bis sechs Wochen warten, bevor ein Arzt das Ansprechen auf die Medikamente beurteilen kann, da die Wirkung von Antidepressiva erst mit Verzögerung eintritt. Die Erkennung des Ansprechens auf die Behandlung zu einem früheren Zeitpunkt hat das Potenzial, mehrere Vorteile für Menschen mit Depressionen mit sich zu bringen, indem die emotionale und wirtschaftliche Belastung von Depressionspatienten verringert wird, da Behandlungen, die nicht wirksam sind, Wochen früher ersetzt werden können.

In dieser Masterarbeit betrachten wir die Vorhersage des Ansprechens von Patienten auf die Behandlung als ein Klassifikationsproblem, indem wir die dynamischen Eigenschaften von EEG-Aufzeichnungen von Depressionspatienten nutzen, die sich einer antidepressiven Behandlung unterziehen. Wir untersuchen die Anwendung der führenden Motiverkennungsalgorithmen SCRIMP++ und OSTINATO auf EEG-Aufzeichnungen und schlagen einen Prozess von der Motivextraktion bis zur Erstellung eines Klassifikators mit hoher Vorhersageleistung vor. Motive mit unterschiedlichen Längen werden aus drei Frequenzbändern extrahiert: Alpha, Beta und Theta, und als Merkmale in einfacheren und besser interpretierbaren Modellen verwendet. Der vorgeschlagene Prozess der Merkmalsextraktion umfasst Kriterien für die Motivauswahl und die Behandlung von Klassen- und Geschlechterungleichgewichten. Die Datenbank besteht aus insgesamt 176 Patienten und ist in einen Trainings- und einen separaten Testsatz unterteilt. Wir untersuchten vier Klassifikatoren, von denen der SVM-Klassifikator mit einer Genauigkeit von 0,738 und einem F1-Wert von 0,744 in der Testgruppe, die beste Leistung erzielte. Die Ergebnisse zeigen, dass die dynamischen Eigenschaften der EEGs potenziell Informationen enthalten, die bei der Unterscheidung zwischen Ansprechern und Nicht-Ansprechern helfen könnten.

Diese Masterarbeit ist Teil des internationalen Forschungsprojekts "Learning Synchronization Patterns in Multivariate Neural Signals for Prediction of Response to Antidepressants", einem gemeinsamen internationalen Projekt der Universität Wien, der Tschechischen Akademie der Wissenschaften und des Nationalen Instituts für Psychische Gesundheit der Tschechischen Republik.

Contents

Acknowledgements i

Abstract iii

Kurzfassung v

List of Tables ix

List of Figures xi

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Problem description . . . 2
- 1.3 Outline of the solution . . . 4

2 Literature Overview 5
- 2.1 Classification using EEG data in MDD . . . 5
  - 2.1.1 Predicting response to MDD treatment . . . 6
- 2.2 Motif Discovery Algorithms . . . 8

3 Background 11
- 3.1 Motif Discovery via Matrix Profiles . . . 11
- 3.2 Motif Discovery via Consensus Motifs . . . 13

4 Methodology 17
- 4.1 EEG Database . . . 17
- 4.2 Workflow . . . 20
- 4.3 Motif Discovery . . . 24

5 Experiments 27
- 5.1 Motif extraction . . . 27
- 5.2 Classification . . . 31
- 5.3 Evaluation . . . 38

6 Discussion 43
- 6.1 Results . . . 43
- 6.2 Limitations of the approach . . . 44

7 Conclusion **45**
7.1 Contributions 45
7.2 Future work 46

Bibliography **47**

List of Tables

2.1 Overview of different motif discovery algorithms and their properties. **Exact**: ✓- exact solution, ✗- approximate solution, ✓* both; **Variable-length**: ✓- finds all motif lengths, ✗- the motif length is fixed; **Multivariate**: ✓- more than 2 time series, ✗- single time series; **Large-scale**: ✓- efficient and scalable, ✗- otherwise. 9

4.1 Class and gender distribution in the training and test sets. 18
4.2 Class and gender distribution in the train and validation split of the whole training set. The labels '0' and '1' denote non-responders and responders, respectively. 20

5.1 Table with the best training and evaluation results for each frequency band and motif length, evaluated by the accuracy and F1 score. 34
5.2 Table with the best results on the final testing set for each frequency band and motif length, evaluated by the accuracy and F1 score. 38

List of Figures

3.1 Short segment of the exact and approximate computation of the matrix profile of an electrode, computed by SCRIMP++, having the approximate matrix profile computed by considering 5% of all pairwise distances. . . . 13
3.2 Example of a short segment of the matrix profile of an electrode, together with the original time series of the electrode. The global minima of the matrix profile are marked with dashed lines, revealing the positions of the motif and its closest subsequence neighbor. 14
3.3 Example of a consensus motif found across multiple electrodes from the same channel in the alpha band using OSTINATO. 15

4.1 The standard 10-20 EEG setup, with the 19 electrodes (channels): Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, P4, T3, T4, T5, T6, O1, and O2. 17
4.2 Short segment of the EEG signal in the first electrode (channel Fp1) as well as the extracted frequency bands: beta, alpha and theta. 19
4.3 Visualization of an example output from the method *match*, which returns the distances to the best matches (orange) of a motif (red) within a given signal. 21

5.1 Example of a motif from the beta band and length 250 with a higher difference score, together with three examples matches across the two classes. 28
5.2 Example of a motif from the beta band and length 250 with a lower difference score, together with three examples matches across the two classes. 28
5.3 Example of a motif from the alpha band and length 250 with a higher difference score, together with three examples matches across the two classes. 29
5.4 Example of a motif from the alpha band and length 500 with a lower difference score, together with three examples matches across the two classes. 29
5.5 Heatmap presenting the mean difference scores across each electrode. The y-axis is sorted by the frequency band, where the rows represent the different motif lengths (sorted in ascending order). 30
5.6 Heatmap presenting the mean difference scores across each electrode (y-axis). The x-axis is sorted and grouped by the motif length, where each group has all three frequency bands. 30

5.7 Heatmaps presenting the mean difference scores across each electrode within each frequency band. The y-axis represents the different motif lengths (sorted in ascending order). 31
5.8 Example of the distribution of the projected values of an imbalanced training feature matrix of the alpha band, motif length 500. 32
5.9 Examples of the distribution of the projected values of the training feature matrix of each band. 33
5.10 Accuracy scores of the training (left) and validation (right) from the k-Fold Cross Validation 35
5.11 Accuracy scores of the training (left) and validation (right) from the k-Fold Cross Validation 36
5.12 Frequency count of the selected motifs across all classifiers for both algorithms (SCRIMP++ on the plots on the left, OSTINATO on the plots on the right). 37
5.13 The number of times the model (x-axis) obtained the best accuracy score on the final testing set for each band and motif length combination. 39
5.14 Confusion matrices for the final evaluation of the best classifier. The y-axis represents the true labels, while the x-axis the predicted labels. 40
5.15 The frequency of motifs across their electrode of origin, used for the best model for the alpha band with length 1000, grouped by class (left) and gender (right). 41

1 Introduction

In this chapter, we present a short overview of the motivation and the objective we are addressing in this research. In Section 1.1, we describe the current antidepressive treatment for depression and its drawbacks. In Section 1.2 we explain the research question in detail, as well as the challenges. Section 1.3 gives an overview of the contents of each chapter in the thesis.

1.1 Motivation

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 characterized by sad and depressive moods, reduced interests, cognitive dysfunction, and physical symptoms such as appetite or sleep disturbances [OGP+16]. Depression differs from regular mood changes and negative emotions from everyday life, as it can have an impact on various aspects of one's life, including relationships with family, friends, and society in general. It occurs twice as often in women than in men and it occurs in approximately one out of every six adults at some point in their lives [Org23].

It is estimated that 3.8% of the population suffers from depression, which amounts to approximately 280 million people worldwide. Depression affects 5% of adults and 5.7% with age above 60. Additionally, the risk of suicide in people with MDD is about 20 times that of the general population [oHME23].

The treatments for depression include psychological treatment and antidepressant medication, nevertheless, due to limited resources, antidepressants are employed more frequently than psychological interventions. On average, the response rate of antidepressants, i.e., the percentage of cases where there is an improvement of depression symptoms, is in the range of 42-53% [TSB+21]. 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 [GJG+17]. To assess the effectiveness of the antidepressant treatment and monitor changes over time, it is a common practice for healthcare providers to use the Montgomery-Åsberg Depression Rating Scale (MADRS) [MÅ79]. The MADRS questionnaire measures the severity of depression in the individual; a higher score indicates worse symptoms. If the MADRS score obtained after 4 to 6 weeks of treatment shows no improvement in the patient's symptoms, then the treatment needs to be changed or adjusted.

Although antidepressive treatments have been effective in treating depression for some patients, there are several limitations and downsides associated with the common clinical practice. 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. This includes worsening of the patient's state and increases the likelihood of occurrence of possible side effects, impacting the patient's quality of life. Besides the patient's well-being, a major issue with depression treatment are the costs, which represent a burden to the economy. According to a study [GFS+21] with data of MDD patients in the US, the incremental economic burden of adults with MDD has increased by 37.9%, from $US236.6 billion to $US326.2 billion between the years 2010 and 2018. The costs consist of direct costs, such as medical services, medication purchases, etc., and indirect costs, such as problems in the workplace (decreased work productivity, missed workdays, unemployment, etc.). Approximately 40% of this rise is linked to an increase in the number of individuals with MDD, whereas the remaining 60% is the result of increased costs per patient. The identification of treatment response closer to its start has the potential of introducing several benefits for people suffering from depression, by reducing the emotional and economic burden of depression patients, as treatments that are not effective can be replaced weeks earlier.

Most studies that have addressed the prediction of treatment response to MDD, have utilized baseline clinical data [CBD+21], which resulted in variations in prediction performance. Recently, there has been a growing interest in the use of neurophysiological markers, such as electroencephalography (EEG) signals, which present a cost-effective and potentially scalable clinical tool. According to a recent review [WPR+22] of machine learning approaches that use EEG data for addressing the prediction of treatment response in MDD, there is a lack of consistency among feature selection and extraction methods used. This could be partially due to the properties of EEGs in depressive patients, as they do not usually contain characteristic episodes that are, for instance, observable in EEGs of epileptic patients, making the EEG signal different with and without seizures. The study highlights that the large variation among the current approaches hinders the establishment of well-defined individual biomarkers that can aid in choosing the appropriate treatment for MDD. As a result, there is a need to develop a standardized and reliable way of producing features that can effectively discriminate between respondents and non-respondents of MDD treatment.

1.2 Problem description

In this thesis, we approach the prediction of patient response to treatment as a classification problem, by utilizing the dynamic properties of EEG recordings of depression patients undergoing antidepressant treatments. In order to build a classifier for earlier prediction of responsiveness of treatment, features extracted from the EEGs of the 7th day after the start of the treatment are used, together with the responsiveness of the treatment from the 28th day as the target variable. In this work, 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 any pre-existing knowledge about their location and shape [TL17]. Motifs have been defined in the literature as short time series that represent reoccurring patterns, frequent trends, or approximately repeated sequences [CKL03][UBA04][MIES07]. Motifs can capture characteristic temporal changes and allow for the identification and representation of relevant patterns at different scales and lengths over time. In the case of EEG, motifs can respond to specific brain activities or states, e.g., different stages of being asleep can be differentiated using motifs detected in EEG sleep data [KMRP00]. Detecting such reoccurring patterns in EEGs could aid in understanding different brain behaviours between patient groups, as they can be easily visualized and inspected by domain experts.

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 [TL17].

In the case of EEG, we have multivariate time series and therefore, we want to discover the so-called consensus motifs that represent conserved patterns that occur in a single time series as well as across other time series. Thus, having a proper aggregation and selection criterion might be needed to narrow down the list of potential motifs to those that are most likely to be important or informative and lead to a better interpretation and accuracy.

There are several key aspects when tackling the classification problem. We need to develop an interpretable classifier, which can provide insight into the decision-making process and the underlying factors that influence the classification. Another crucial property is reliability, due to its role in decision-making when managing depression treatments. The input space of the model should have a simple representation in order to avoid high dimensionality, which can affect the model's performance. A desired property of the classifier is scalability, meaning it is able to 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.

Based on the above-mentioned potential benefits of using motifs as dynamic and interpretable features in this classification setting, the research question that we want to address in the thesis is the following:

How can we extract motifs that exclusively characterize both groups of respondents and non-respondents from patient EEG signals and develop a binary classifier?

This objective consists of several subtasks. We have to effectively extract motifs from multivariate EEG time series data from depression patients and identify the most informative and relevant motifs. This means that we need a selection approach for choosing a subset of the discovered motifs that can be useful for this classification problem. In order to achieve high predictive performance and interpretability, an appropriate feature space that is derived from the identified motifs is needed. This includes investigating how different motif-based feature engineering approaches compare in terms of distinguishing between patient groups, i.e. predictive performance.

1.3 Outline of the solution

We present a methodology for addressing the prediction of treatment outcomes using motifs obtained from the patients' EEGs as features. The proposed workflow consists of two main steps: *motif extraction*, where we apply a motif discovery algorithm to extract motifs, and *classification*, where we utilize the motifs in the feature engineering step, to produce a feature space that would allow the classifier to distinguish between the two patient groups.

The related work and challenges associated with addressing this problem are described in Chapter 2, as well as an extensive overview of current motif discovery methods and their properties.

Within the scope of the experiments, we use two motif discovery algorithms, SCRIMP++ and Ostinato, described in Chapter 3. The algorithm SCRIMP++ [ZYZ+18] represents a similarity search algorithm for finding fixed-length motifs in a single time series. It allows for approximated search which provides a much faster convergence, hence it is suitable for time series with a lot of samples. The algorithm Ostinato [DPVH20] is designed to detect fixed-length motifs that are present among multiple time series, called consensus motifs. In this setting, we aim to detect consensus motifs among the signals within the same patient group. A more detailed explanation regarding the application of the motif discovery algorithms on the EEG data is presented in Chapter 4.

EEG recordings are analyzed within different frequency bands, hence in this thesis, we cover several frequency bands: alpha, beta, and theta in a separate experiment. In addition, since both algorithms require the motif lengths as an input, for every frequency band we conducted several experiments for different motif lengths. Each experiment workflow consists of the application of one motif discovery algorithm, where a large number of motifs are extracted. To obtain a smaller subset containing motifs that have a discriminatory power to help differentiate between the two classes, we propose a motif ranking criteria. The motif ranking helps in deciding which motifs to keep and use in the feature engineering step, where for every patient, we construct a feature vector with the closest distances to the chosen subset of motifs. We use four common classifiers, such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression. The results from the experiments are summarized in Chapter 5. The results and contributions are discussed in Chapter 6, followed by a summary of the obtained conclusions and ideas for future work in Chapter 7.

2 Literature Overview

In this chapter, we present a summary of the classification challenge involving EEG signals in MDD, which can be found in Section 2.1. Subsection 2.1.1 specifically addresses the classification of treatment response in MDD. Lastly, Section 2.2 contains an overview of the current motif discovery methods and their properties.

2.1 Classification using EEG data in MDD

In recent years, researchers have become increasingly interested in using EEG signals to discover characteristics and patterns that relate to a certain psychiatric disease. Such findings are often referred to as biomarkers, which can be useful in identifying the presence of a disease, discovering pathophysiological mechanisms, and predicting outcomes of treatment.

In a machine-learning setting, these classification problems require a suitable feature extraction technique that can construct informative features from the EEG signals. Due to the properties of EEG signals, being non-stationary, non-linear, non-Gaussian, and having multiple channels, the process of feature extraction presents a major challenge.

Nonetheless, EEG signals have been shown to have effective discriminative power over differentiating MDD patients and healthy subjects, according to a review [GMC+21] that investigates EEG features for detecting MDD. Significant differences in frequency band power have been shown to be useful for diagnosis, however, the findings are inconsistent even within the same frequency band. One consistent result is the hyper-activation of the right prefrontal cortex in MDD, associated with withdrawal behaviors. In regard to time-related changes, MDD patients also exhibit different values in event-related potentials (ERP), such as lower amplitudes or shorter latency. Differences in ERPs related to stimuli type and feedback are also found, although the results are not conclusive. Complexity metrics indicate higher values of nonlinear parameters in MDD, reflecting fractal and unpredictable characteristics of the data. MDD is associated with increased EEG coherence, indicating heightened neurophysiological connectivity, and abnormal graph properties. However, summarizing the results and findings obtained across different papers, the authors conclude that specific brain areas linked to changes in EEG activity remain uncertain and require additional exploration.

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. Čukić et al. [ČSSP20] present the effectiveness of two non-linear measures: Higuchi's fractal Dimension (HFD) and Sample Entropy (SampFN) and explore multiple

classification models, reaching an accuracy in the range of 90.24% to 97.56%. A shortcoming of this research is, however, the relatively small size of the dataset, consisting of only 23 patients.

Besides the prediction of the presence of depression, EEGs have been utilized for other cases of classification. Bučková et al. [BBBH20] have used the same dataset of EEGs that is also used in this thesis and their objective is to classify biological sex from the EEG recordings. The aim is to test the hypothesis of the presence of higher beta power in women compared to men, as well as in the presence of depression. The authors have obtained the best performance with a convolutional neural network model, achieving 81% accuracy.

Deep learning approaches have been also used in the diagnosis of depression disease. An exhaustive review [CFS+18] of existing neural networks-based approaches for diagnosing MDD and bipolar disorder using EEG signals emphasizes that EEG-based methods have substantial potential for assessing and monitoring both diseases. The authors conclude that although deep learning models are capable of achieving good predictive performance, they still remain "black boxes", as their lack of interpretability poses a significant challenge for physicians, who may be unable to explain to patients how their diagnosis was determined. This prevents the community from further developing reproducible and deterministic protocols and achieving clinically useful results.

2.1.1 Predicting response to MDD treatment

The persistent clinical challenge in optimizing the treatment management of depression patients has led to an increasing focus on building predictive models using machine learning and information extracted from EEG signals to determine whether the patient responds to the treatment.

Alik et al. [WBM+19] provide an overview of the implementation and the reliability of potential EEG biomarkers, and investigate the features used for the prediction of treatment response in MDD. The research papers that were considered in the review focused on analyzing the changes in alpha and theta rhythms, as well as low-frequency EEG power and cordance, being the most heavily represented features. The sensitivity and specificity of these features were 0.72 and 0.68, respectively, with an area under the curve of 0.76, however, the study sizes are generally small, with a median of 25. Additionally, the review highlights that such features are not clinically reliable, due to underreporting of negative results, lack of out-of-sample validation, and insufficient replication of previous findings.

According to a recent machine learning meta-analysis review [WPR+22] on this topic, features such as nonlinear features, spectral entropy, and cordance have shown promising performance in predicting treatment response. The authors point out that EEG features are better at capturing predictors of clinical non-response rather than predictors of clinical response across different treatment modalities. EEG signals consist of various frequency components that can be categorized into different bands or rhythms, such as alpha, beta, gamma, and other waves - each of them linked to various brain states and cognitive processes. Alpha, theta, and gamma power in frontal electrodes were

regarded as significant features according to the review, as well as, coherence between frontal and temporal electrodes, and baseline power at specific electrodes. However, the studies exhibited variations in the number of electrodes, electrodes of interest, and methods for extracting features. Regarding the choice of the classification methods, Support Vector Machine (SVM) [HDO+98] with a radial kernel demonstrated the best performance, among the covered articles, in predicting treatment response using EEG power and asymmetry features. The advantage of SVM outperforming other algorithms like Random Forest in most cases, lies in dealing with high-dimensional data.

Working with EEG signals often implies dealing with high-dimensional data, due to the number of electrodes, duration of recordings, and sampling rate. A theoretical study [DS19] reveals that as the dimensionality increases, the number of reproduction events decreases exponentially, leading the system to enter an infinite cover-delete cycle. As a result, it is more difficult for the classifiers to produce highly general and accurate rules beyond a certain number of dimensions. Althian et al. [AAAB+21] investigate the impact of dataset size on the performance of several widely used supervised machine learning models in the medical domain. Their interesting findings show that model performance depends more on the data representation itself rather than the dataset size. Moreover, they argue that a robust model that utilizes a limited dataset does not guarantee the best performance compared to other models.

The current attempts to develop approaches to address the issue of treatment response in MDD using EEG signals mainly focus on developing and utilizing suitable feature extraction methods. They predominantly focus on the static property of the signal, e.g., statistic properties, application of non-linear methods, time and frequency domain features, etc. Mueen et al. [MKBS09] comment on the possible usage of motifs as features for classification. In the paper, the authors set up a disk-aware algorithm to find fixed-length motifs in multi-gigabyte databases containing massive amounts of time series which detects motifs of a specific length in the extracted independent components of the signal. They suggest extracting the features using a similarity measure between motifs and using the computed distance between motifs as features. To our best knowledge, the usage of motifs as building blocks of features from EEGs has not been used yet, especially not in the case of predicting the diagnosis of depression, nor in predicting the outcome of the anti-depressive treatment.

In the scope of this project, "Learning Synchronization Patterns in Multivariate Neural Signals for Prediction of Response to Antidepressants" [oV21, NIoMHotCRN22], a thesis work [PPS23] investigates the application of Graphical Granger Causality by Information-Theoretic Criteria, by computing Granger-causal networks in different frequency bands. The training was done for each gender separately. The classifier that obtained the best evaluation score was the Decision Tree Classifier, with an F1 score of 0.61 on the test set for female subjects, while for male subjects, the F1 score on the test set was only 0.2. The result of this could be that separating the dataset by gender meant using less data for both trainings, hence the classifier could not learn to generalize well.

2.2 Motif Discovery Algorithms

Motifs are a recurring theme in the literature on time series analysis and have been variously referred to as patterns, trends, sequences, shapes, episodes, or frequent subsequences, among other names [TL17].

Motif discovery algorithms vary in their approach depending on the specific application. Some of these algorithms can identify exact or approximate motifs, motifs with fixed lengths, or variable lengths. Dealing with either univariate or multivariate time series data is also important to consider when finding motifs, as multivariate data is more complex to analyze and identify meaningful patterns across different channels. In short, when choosing a motif discovery algorithm, it is crucial to consider the algorithm's properties and ensure they align with the specific requirements of the application at hand.

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. Algorithms that address the problem of finding motifs with a fixed length, and are referred to as state-of-the-art, are the MK [MKZ+09] and the QuickMotif [LLYG15] algorithm. The shortcomings of these algorithms are that in less ideal situations, both algorithms can degenerate to brute-force search and require a lot of memory compared to more recent methods. Yeh et al. introduce an algorithm called STAMP [YZU+16a] which utilizes a fast similarity search algorithm to find exact fixed-length motifs in a time series of length l, with time complexity $O(l^2 log(l))$. The authors further introduce STOMP [ZZS+16], which reduces the time complexity of STAMP by $O(log(l))$. Both algorithms have time and space complexities that are independent of the given motif length. However, despite its slower computational time, STAMP is often preferred over STOMP due to its rapid convergence. In most cases, running STAMP to a partial completion is sufficient to obtain a precise estimation of the desired solution. The algorithm SCRIMP++ [ZYZ+18] combines the features of both STOMP and STAMP, having a time complexity of $O(l^2)$, and requires almost the same time as STOMP to reach full convergence.

However, determining the optimal length to search for a motif can be a challenging task. The challenge with variable-length motif search is the scalability, as its computational complexity compared to the fixed-length search problem can be ten times larger. To address this, approaches such as VALMOD [LZPK18], HIME [GL19b], and MOEN [MC15] have been proposed to search for motifs of all lengths within a given range. The limitation of these algorithms is that their search for variable-length motifs is limited to one or at most two time series.

Recently, there has been an increase in research on discovering consensus motifs in multiple time series. Unlike traditional motifs, which are the most similar subsequence pairs from one or two time series, consensus motifs are common patterns among multiple time series. The Ostinato algorithm [DPVH20] is designed to detect fixed-length motifs across n time series, with memory complexity of $O(l)$ and the time complexity of $O(n^2 l^2 log(l))$, where n is the average time series length. It achieves this by introducing a technique to compute the distances among various subsequences extracted from different time series using a rapid pruning strategy. It then selects the seed subsequence with the smallest

distance as the precise consensus motif.

The algorithms CHIME [GL19a] and VACOMI [ZWW22] enable the search of variable-length motifs in a set of more than two-time series within a specified range of motif length. The most significant factor that dominates the time complexity of CHIME is the pairwise comparison post-processing, which takes $O(nl^2m)$ time, where n is the number of time series (dimensions), l is the length of a single time series and m is the length of the detected motif. In the case of CHIME, to avoid redundant subsequences, the recurring symbols from each dimension are combined together and used for the symbol-matching results collectively across all dimensions. VACOMI introduces a lower bound on the distance between multiple subsequences extracted from distinct time series, which can be computed in linear time. Additionally, the authors present an auto-tuning technique that helps prune more sequences for processing, thus significantly reducing the computational time.

Algorithm	Exact	Variable length	Multivariate	Large-scale
MK [MKZ+09]	✓	✗	✗	✗
Quick-motif [LLYG15]	✓	✗	✗	✗
MOEN [MC15]	✓	✓	✗	✗
STAMP [YZU+16a]	✓	✗	✗	✗
STOMP [ZZS+16]	✓	✗	up to 2	✗
SCRIMP++ [ZYZ+18]	✓*	✗	✗	✓
VALMOD [LZPK18]	✓	✗	up to 2	✓
HIME [GL19b]	✗	✓	up to 2	✓
CHIME [GL19a]	✗	✓	✓	✓
OSTINATO [DPVH20]	✓	✗	✓	✓
VACOMI [ZWW22]	✓	✓	✓	✓
k-Motiflets [SL22]	✓*	✗	✗	✓

Table 2.1: Overview of different motif discovery algorithms and their properties.
Exact: ✓- exact solution, ✗- approximate solution, ✓* both;
Variable-length: ✓- finds all motif lengths, ✗- the motif length is fixed;
Multivariate: ✓- more than 2 time series, ✗- single time series;
Large-scale: ✓- efficient and scalable, ✗- otherwise.

There are also different approaches to the motif discovery problem. Schäfer et al. [SL22] an algorithm for finding *k-Motiflets* with an exact and approximate solution. These motiflets are defined as the set of exactly k occurrences of a motif of length l with minimal maximum pairwise distance. The authors argue that setting the parameter k of a motif set is more intuitive and easier to set than the distance threshold r. The authors propose two extensions to learn the input parameters k and l from the data and have used an experiment with EEG sleep signals to find the two largest motif sets, which correspond to well-known motifs in sleep EEG data (K-Complex and sleep spindles). The shortcoming

of this algorithm is that currently it only works for only one time series. Table 2.1, gives a short summary of the mentioned motif discovery algorithms and their properties.

Some of the authors of the above-mentioned motif discovery algorithms have tested the algorithms by additionally conducting classification experiments with the obtained motifs. For instance, the authors of CHIME use the identified subdimensional motifs in a binary prediction problem. They construct a distance feature vector with a simple decision tree classifier, resulting in around 70% accuracy by only using the top 3 motifs. A recent paper introduces an interpretable approach IRMAC [YPS+23] that uses the motifs obtained from STOMP, to predict two classes of electricity users. However, in general, the research regarding the utilization of motifs in classification is limited.

In this thesis, we use the SCRIMP++ and Ostinato algorithms for discovering motifs, due to their available implementation in Python, within the STUMPY library [Law19], as well as their suitability for our classification problem. The SCRIMP++ provides a robust approximate computation of the matrix profiles, which is suitable for large time series for our case. On the other hand, Ostinato has a larger computation time but provides the identification of consensus motifs, which in our case means finding motifs that are common within the signals belonging to the same class. The algorithm CHIME combines both fast approximating solutions and handling multiple time series as input, by identifying multidimensional motifs (i.e., motifs that occur simultaneously across multiple signals). Although this approach was publicly available and had been implemented, it proved problematic when executed within the context of our study. Despite our best efforts, we encountered errors and complications that made it difficult to extract meaningful or presentation-worthy results to report.

3 Background

In this section, we provide a more detailed description of the motif discovery algorithms that are used in this thesis. In Section 3.1, we present the theory behind the matrix profile computation and how it is used within the SCRIMP++ algorithm, while in Section 3.2 we extend this theory to the discovery of consensus motif by the Ostinato algorithm.

3.1 Motif Discovery via Matrix Profiles

The computation of the matrix profile addresses motif discovery by all-pairs-similarity search. One can trivially compute all top-k motifs and motifs of different ranges if one has access to the matrix profile [YZU+16a]. In order to formally define the matrix profile, based on [YZU+16a], we first introduce the following terms:

- A time series T is a sequence of real-valued numbers $t_i : T = t_1, t_2, \ldots, t_n$ where n is the length of T.
- A sub-sequence $T_{i,m}$ of a T is a continuous subset of the values from T of length m starting from position i: $T_{i,m} = t_i, t_{i+1}, \ldots, t_{i+m-1}$, where $1 \leq i \leq n - m + 1$.
- A distance profile D_i is a vector of the Euclidean distances between a given query $T_{i,m}$ and each sub-sequence in an all-sub-sequences set in a given time series T. The j^{th} element of D_i, represents the distance between $T_{i,m}$ and $T_{j,m}$.

  Once we have D_i, we can determine the nearest neighbor of $T_{i,m}$ within T. If $T_{i,m}$ is a subsequence of T, the i-th position in the distance profile D_i will be zero, and the values that lie close to i will be close to zero. Hence, we establish an exclusion zone of length $\frac{m}{4}$ before and after i, in order to avoid such situations, commonly referred to as "trivial matches" in the literature. In practice, we assign $d_{i,j}$ to infinity for $i - \frac{m}{4} \leq j \leq i + \frac{m}{4}$. Consequently, by computing the minimum value in D_i we can identify the nearest neighbor of $T_{i,m}$.

  The matrix profile and the matrix profile index hold the information about the nearest neighbor for every subsequence T:

- A matrix profile P of time series T is a vector of the Euclidean distances between every subsequence of T and its nearest neighbor in T. We can formally define P as: $P = [\min(D_l), \min(D_2), \ldots, \min(D_{n-m+l})$, where $D_i(l \leq i \leq n - m + l)$ is the distance profile D_i corresponding to query $T_{i,m}$ and time series T. The i^{th} element in the matrix profile P tells us the Euclidean distance from subsequence $T_{i,m}$ to its nearest neighbor in time series T.

- A matrix profile index stores the location of the nearest neighbor. Formally, I of time series T is a vector of integers: $I = [I_1, I_2, \dots I_{n-m+I}]$, where $I_i = j$ if $d_{i,j} = \min(D_i)$.

In order to perform motif discovery using the matrix profile, we use SCRIMP++ [ZYZ+18], which consists of two parts. In the first part, PreSCRIMP (also introduced within the same paper), serves as a fast preprocessing step in which it employs the Consecutive Neighborhood Preserving (CNP) property of time series subsequences. This property divides the matrix profile indexes into sections where consecutive subsequences tend to have consecutive subsequences as their nearest neighbors. In other words, the property suggests that if $T_{i,m}$ and $T_{i+1,m}$ are subsequences, then there is a high probability that their nearest neighbors are $T_{j,m}$ and $T_{j+1,m}$. The CNP property enables PreSCRIMP to rapidly generate an approximate matrix profile, by sampling subsequences from the time series at fixed intervals and finding the exact nearest neighbor for each sampled subsequence. The number of pairwise distances that are sampled are controlled by the sampling rate s. The larger this parameter is, the better the approximation, hence if we use 100% of the pairwise distances, we obtain the exact computation. It is worth noticing that for the approximation case, even though fewer pairwise distances are being computed, no pairwise distance is being approximated.

In the second part, the SCRIMP algorithm builds upon the matrix profile produced by PreSCRIMP, further updating it until it reaches either a runtime threshold or an exact solution. SCRIMP computes the distance matrix for each pair of time series subsequences by using the z-normalized Euclidean distance measure [WMD+13]. The z-normalized Euclidean distance of the time series subsequences $T_{i,m}$ and $T_{j,m}$ is defined as:

$$d_{i,j} = \sqrt{2m\left(1 - \frac{Q_{i,j} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)}$$

where $Q_{i,j}$ is the dot product of the two subsequences, μ_i and μ_j are the means of $T_{i,m}$ and $T_{j,m}$, while σ_i and σ_j are the standard deviations of $T_{i,m}$ and $T_{j,m}$. This distance measure enables the comparison of the shape of the subsequence since it is invariant to the amplitude changes and phase shifts.

In the case of dealing with EEG data, we have relatively large time series, where a single electrode per patient can hold up to 130,000 data points. For the purpose of speeding up the computation, we compute the approximate matrix profile. Figure 3.1 shows a short segment of the matrix profile computation of the exact (using all pairwise distances) and the approximate approach (using a small percentage of all the distances). One can observe that even by considering a small portion of them, the approximate case captures the true matrix profile quite well.

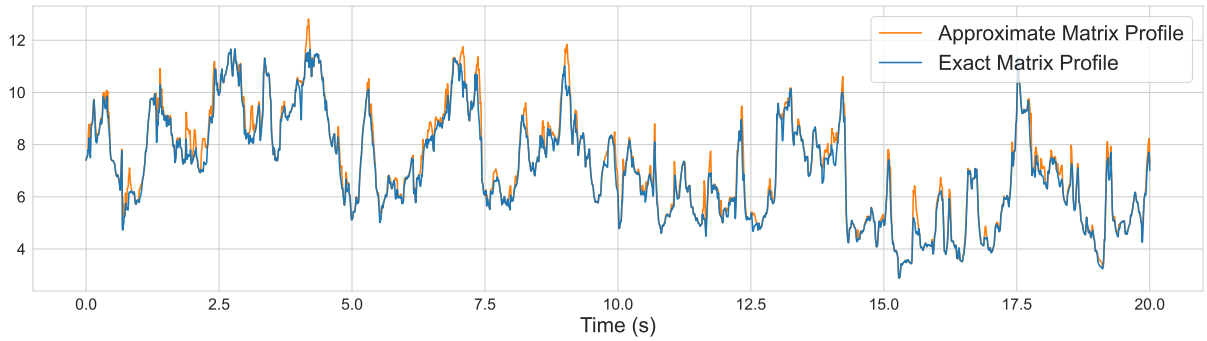


Figure 3.1: Short segment of the exact and approximate computation of the matrix profile of an electrode, computed by SCRIMP++, having the approximate matrix profile computed by considering 5% of all pairwise distances.

As explained above, the matrix profile is essentially a vector that keeps track of the z-normalized Euclidean distance between any subsequence within a time series and its closest neighbor. Consequently, when the matrix profile values are low, it suggests the potential existence of a pattern, while a reference subsequence with a high matrix profile value may indicate the presence of an anomaly. Figure 3.2, shows an example of two similar subsequences in a short segment of an electrode (potential motifs), alongside the computed matrix profile.

3.2 Motif Discovery via Consensus Motifs

When referring to consensus motifs, we approach the problem of motif discovery in a slightly different matter, i.e., by discovering repeated subsequences in a set of time series. The algorithm OSTINATO [DPVH20] introduces the definition of consensus motifs and a scalable approach for their discovery. In order to describe this approach, we begin with the necessary definitions adopted from [DPVH20]:

- The radius r of a subsequence $T^i_{j,m}$ of time series T^i with respect to a sequence of time series $T^1, ...T^k$ represents the maximum distance between $T^i_{j,m}$ and its neighbor in each of $T^1, ..., T^k$.

  The intuition behind the radius r is that we consider the subsequences within a set of k time series as points in an m-dimensional space. Each subsequence can be surrounded by a hypersphere, with the sphere's radius r increasing until it encompasses at least one subsequence from each of the time series. The consensus motif is the subsequence that possesses the smallest value of the radius among all subsequences.

- Given a sequence of k time series $T_1, \ldots, T_k$, the *consensus motif* is the subsequence which possesses the smallest radius of any subsequence appearing in any of time series $T_1, \ldots, T_k$.

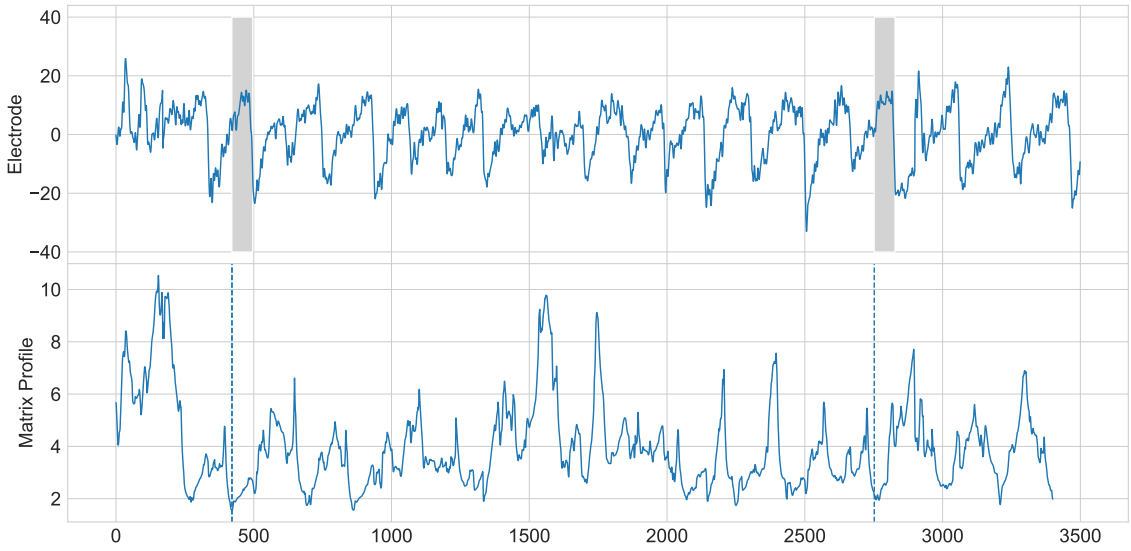


Figure 3.2: Example of a short segment of the matrix profile of an electrode, together with the original time series of the electrode. The global minima of the matrix profile are marked with dashed lines, revealing the positions of the motif and its closest subsequence neighbor.

These two definitions assume that a highly "conserved" subsequence is present in all k time series, which might not always be the case. Hence, in OSTINATO, when determining the radius, it chooses this set of k time series out of any subset of P time series. By selecting k time series from the set P, we can effectively exclude the remaining $|P| - k$ time series that may be considered outliers.

In order to prune the search space when computing the radius for each subsequence, OSTINATO first computes a lower bound for each candidate subsequence using the $ABJoin$, formally defined as:

- An $ABJoin$ [YZU+16a] is a composite time series that provides annotations for each overlapping subsequence, denoted as $A_{i,m}$, in the sequence A. These annotations indicate the distance to the nearest neighbor of the corresponding subsequence within sequence B.

The algorithm computes the lower bound of each subsequence in the time series using a fast $ABJoin$ method [YZU+16b]. This is followed by k searches, each for one of the k time series. Candidate subsequences from a single time series are sorted in ascending order based on their estimated lower bound, within each search. These candidates are evaluated until the best one within that time series is found. Once the estimated radius of the next candidate exceeds the best radius found so far, the algorithm removes the remaining candidates from that time series from the search.

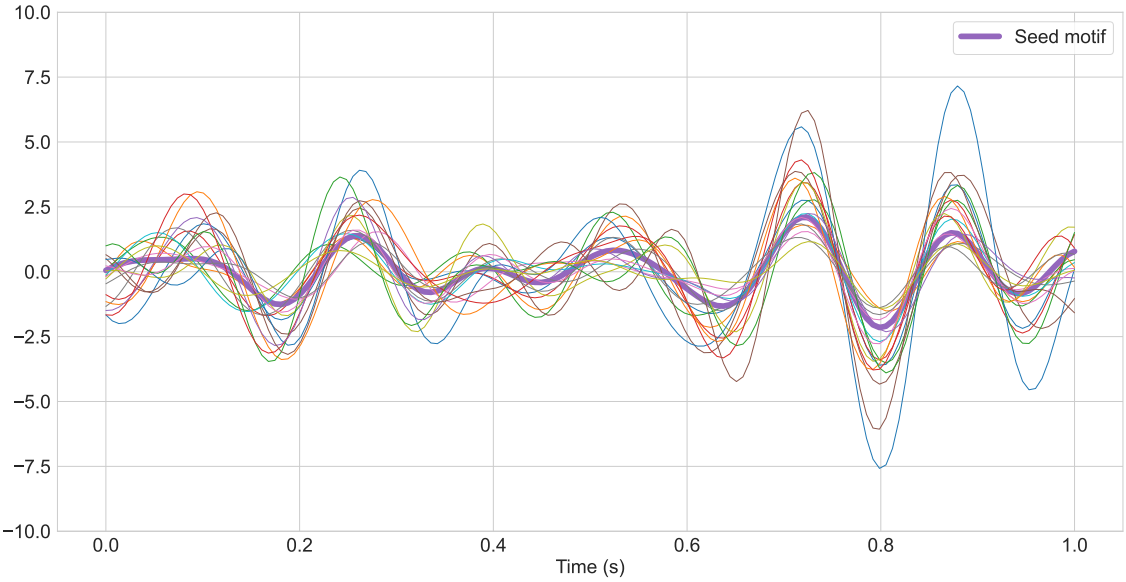


Figure 3.3: Example of a consensus motif found across multiple electrodes from the same channel in the alpha band using OSTINATO.

In the final step, the algorithm returns the candidate with the smallest radius as a triple consisting of its radius, the index of its location, and its subsequence index within the time series. Figure 3.3 depicts the discovered subsequences within the smallest radius. The most central subsequence is also called *seed motif*.

4 Methodology

In this chapter, we present the developed methodology workflow for the classification of response to treatment. Before describing the approach in detail in Section 4.2, we provide an overview of the process of obtaining the EEG recordings as well as their preprocessing and transformations, which can be found in Section 4.1. Section 4.3 contains a more detailed explanation of how the motif discovery algorithms are applied to the EEG data.

4.1 EEG Database

The patient database is recorded and provided by the Czech Academy of Sciences within the Synchronization project. It contains EEG recordings of 228 patients who are treated for Major Depressive Disorder (MDD) with antidepressants. There are two sessions of EEG recordings per patient which are relevant for our task, where the first one takes place before the start of the treatment, whereas the second one is recorded on the 7th day after the start of the treatment.

The recordings were obtained during a 10-minute resting state with the eyes closed, using 19 electrodes. The electrodes are placed on the patient's scalp according to the 10-20 standard EEG setup (their names and corresponding locations are presented in Figure 4.1).

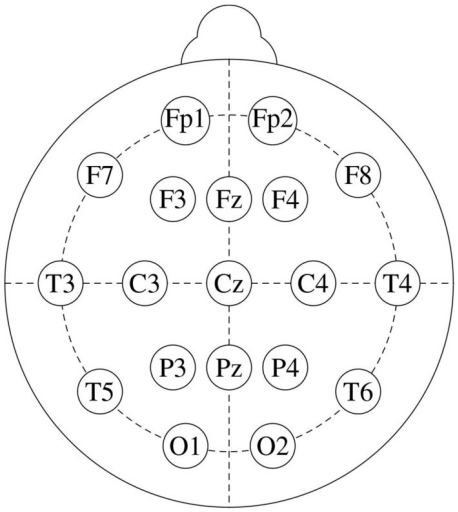


Figure 4.1: The standard 10-20 EEG setup, with the 19 electrodes (channels): Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, P4, T3, T4, T5, T6, O1, and O2.

The initial preprocessing of the EEG signals and the class labels determination have been provided by the project members and are described below:

Preprocessing The preprocessing was done in MATLAB using the open-source toolbox EEGLAB [DM04]. 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 [Off50] 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 in order 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.)

Labels In order to determine whether a patient has responded well to the treatment, psychiatric experts have evaluated the patient's status using the MADRS score. The MADRS score has values from 0 to 60, indicating the level of severity of the depression. The examination was first done before the start of the treatment, and then on the 28th day after the start of the treatment. The response to the treatment is determined based on the score difference between the two evaluations, indicating the level of improvement or worsening of the patient's well-being.

Hence, the class labels indicate whether a patient is a responder or a non-responder. A patient is considered a responder if the MADRS score obtained after 4 weeks has reduced by 50% compared to the initial MADRS score.

	Training Set			**Test Set**		
	Male	Female	Total	Male	Female	Total
Responders	13	53	66	4	13	17
Non-responders	18	50	68	13	12	25
Total	31	103	134	17	25	42

Table 4.1: Class and gender distribution in the training and test sets.

The final dataset used in this thesis consists of 176 patients, after cleaning and removing unstable recordings with technical issues or an excessive number of artifacts. The number of responders is 84, while the number of non-responders is 92. Regarding biological sex, there are 48 male and 128 female patients. The distribution of the classes and genders within the training and test set are depicted in Table 4.1. One can observe that the data is balanced regarding the classes, but quite imbalanced by gender, as the training set contains considerably fewer male patients.

Frequency bands In addition to using the EEG signals as given (with frequencies from 1 to 40 Hz), experiments are conducted on individual frequency bands of each of the electrodes. As pointed out in Chapter 2, several research papers point to the alpha [WBM+19] [ZYL+21], beta [BBBH20] and theta band [WBM+19] [AEH+15], as containing representative biomarkers for depression detection or depression treatment outcome. Hence, our research focuses on these three frequency bands.

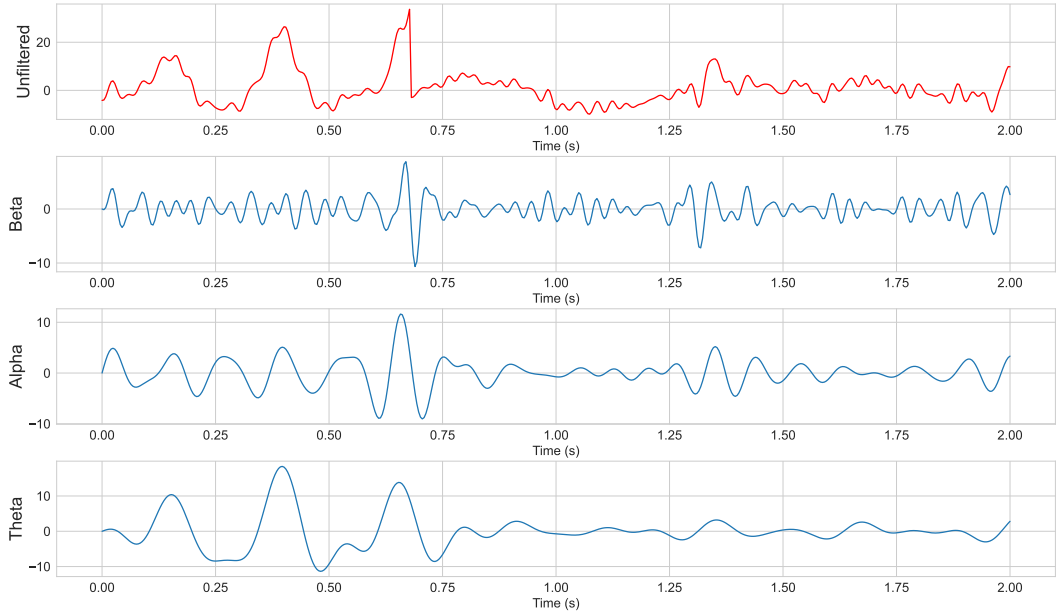


Figure 4.2: Short segment of the EEG signal in the first electrode (channel Fp1) as well as the extracted frequency bands: beta, alpha and theta.

The extraction of the frequency bands theta (4-8 Hz), alpha (8-12 Hz), and beta (12-30 Hz) are computed using the Python MNE library [GLL+13]. Figure 4.2 shows a short segment of the first electrode of a patient across different frequency bands, as well as the unfiltered (raw) signal.

4.2 Workflow

The complete workflow from motif extraction to building a classifier consists of several key steps. The motif discovery part, which includes the application of the motif discovery algorithms, is described in detail in Section 4.3, for each algorithm separately. Each experimental setup that is conducted consists of a varying motif discovery step, since we have different motif discovery algorithms, different frequency bands, and different motif lengths. The remaining steps of the workflow include methods and approaches that are part of every experiment and are described in the following paragraphs:

Dataset split We first split the training data (consisting of 134 patients) in a way that is appropriate for evaluation later in the classification step. Since we are planning to extract motifs and evaluate their predictive power, we would need a separate dataset (i.e., an evaluation set) from the training set to conduct model tuning and feature selection. The testing set is used in the end for the final evaluation.

We separate the training set into two sets: training and evaluation set, as presented in Table 4.2, by preserving a balance on the class label, as well as the gender. From this point on, when we refer to the training set, we refer to the one obtained after the split (108 patients). Thus, the motif discovery is applied to the time series data of this training set.

Set	Label	Female	Male	Total
Training	0	42	12	54
	1	45	9	54
Validation	0	8	6	14
	1	8	4	12

Table 4.2: Class and gender distribution in the train and validation split of the whole training set. The labels '0' and '1' denote non-responders and responders, respectively.

Motif selection criteria After the motif discovery step we obtain a large set of motifs. We want to filter out motifs that do not have any discriminatory power to differentiate between responders and non-responders. To do so, we need to check which motifs appear to be typical for one class, i.e., are not common patterns for both classes.

To be able to assess this, we need to define the process of checking 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's library STUMPY [Law19], which takes a motif and a time series as input. The method returns a list of closest matches, including the z-normalized Euclidean distance of the motif and

the match. 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:

$$f(D) = \max\left(\left(\text{mean}(D) - 2 \cdot \text{std}(D)\right), \min(D)\right)$$

where D is an array with a size of $n - m + 1$ and represents the distance profile of Q with T. Hence, the function $f(D)$ returns at least the closest match. An example of an output of the method *match* (the obtained matches together with their distances) to the motif is shown in Figure 4.3:

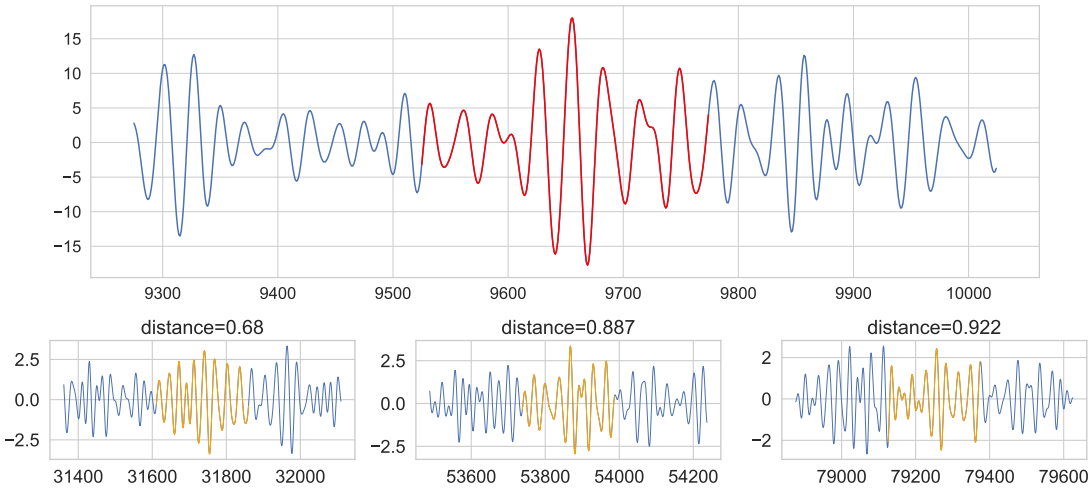


Figure 4.3: Visualization of an example output from the method *match*, which returns the distances to the best matches (orange) of a motif (red) within a given signal.

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. However, finding an appropriate threshold is difficult, as it would also vary depending on the motif length and possibly the frequency band. Therefore, instead of using a threshold, we use the average distances within the match to represent how distant is the found match to the given motif. Motifs with 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* (presented in Algorithm 1) by considering the distances of the motif to its match in the corresponding electrodes for 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. Considering that a motif could be typical for a subset of patients within one class, we take only a *percentage* of the best (lowest) average match distances from both classes and then compute the absolute difference. 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.

Algorithm 1 Compute the difference score of a motif

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

Output: score_{diff}

distances $\leftarrow dict()$
$distances[0] \leftarrow []$
$distances[1] \leftarrow []$
for $i = 1$ *to len(*E^j*)* **do**
    electrode $\leftarrow$E$^j[i]$
    label $\leftarrow$L[i]
    $distances_{matches} \leftarrow match(Q, electrode)$
    $distances[label].append(mean(distances_{matches}))$
end
best_distances_0 $\leftarrow get_(distances[0], percentage)$
$best\_distances\_1 \leftarrow get_(distances[1], percentage)$
$score_{diff} \leftarrow abs(best\_distances\_0 - best\_distances\_1)$

With the help of these scores, we can significantly reduce 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.

Features The feature matrix consists of rows that represent each patient, while the columns represent the motifs. Patient i is represented as a d-dimensional 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 the method *match*, considering the patient's electrode that corresponds to the motif. Essentially, the feature matrix represents a distance matrix.

Classification For predicting the class label, i.e., the treatment outcome of the patients, we use several 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 have used the following models: Support Vector Machine (SVM) with different kernels, Decision Tree, Random Forest, and Logistic Regression, from the scikit-learn library [PVG+11]. As evaluation metrics, we use the accuracy and F1 score, as well as a confusion matrix to further check for class imbalances.

For choosing the most optimal hyperparameters for the model, a 5-fold cross-validation is used. For each model, the hyperparameters that were tuned, are listed bellow:

- 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]
  - min_samples_split: [1, 2, 3, 4]
- Logistic Regression
  - C: [0.1, 0.5, 0.7, 1]
  - penalty: ["l1", "l2"]

Note: In the case of SVM, we consider two implementations provided by the scikit-learn library: SVC (Support Vector Classification) and LinearSVC (Linear Support Vector Classification), which have different loss functions set by default, and different handling of intercept regularization, as the LinearSVC penalizes the size of the bias. Hence, in the case of SVC, we tune the 'C' and 'kernel' parameters, as it uses the 'l2' penalty by default. In the case of LinearSVC, we explore different penalty settings: 'l1' and 'l2', as the L1 norm imposes a stricter regularization.

To further reduce the feature space and retain the most important motifs, we use Recursive Feature Elimination (RFE) [GWBV02]. This wrapper method iteratively chooses features by successively reducing the feature set. Initially, the estimator undergoes training with the complete set of features, where also the significance of each feature is determined, typically using a particular attribute or callable method. Subsequently, the least important features are removed from the current set. This process is repeated recursively on the trimmed feature set until the desired number of selected features is attained. The number of optimal features to keep is obtained based on the highest prediction score on the evaluation set.

4.3 Motif Discovery

The main and most extensive part of the whole workflow process is the motif discovery. As mentioned in Section 4.1, we conduct the identification of motifs among different frequency bands: alpha, beta and delta. The motif discovery algorithms that were used are SCRIMP++ and OSTINATO, both of which have the motif length as an input. In order to pick an appropriate set of motif lengths for conducting the experiments, we have carried out several initial experiments with different lengths. Motifs with very short lengths did not seem to capture meaningful patterns, while motif lengths with very large lengths were not considered due to a large increase in computational time. Thus, we have settled for motif lengths ranging from 50 to 2000 samples (corresponding to a duration from 0.2 seconds to 8 seconds). The implementations of these algorithms and helper functions are available in Python's STUMPY[Law19] library. The application of such algorithms to EEG data is presented in the following paragraphs:

SCRIMP++ Since this algorithm can be used for finding motifs of length m in a single time series, we need to compute the approximate matrix profile for every electrode per patient. This is done using STUMPY's method *scrump*. The method has an additional parameter *percentage*, which controls how many pairwise distances will be considered for computing the matrix profile. Experimental results showed that using just 5% already gives a relatively accurate approximation, as presented in Section 3.1.

Once the matrix profile is obtained, we extract the best motif using the method *motifs*. The method takes the time series and the matrix profile of the time series as input and returns the location of the best motifs, as well as their distances to their neighbors. Considering that we are using this algorithm for every single electrode, we only keep the single best motif obtained by the method (hence, we set the parameter max_motifs to 1). By default, a subsequence becomes a candidate motif if it has at least one neighbor (i.e., a similar subsequence with a distance below a certain threshold). In order to impose stricter criteria for motif candidates, we require at least three neighbors for the subsequence to become a candidate motif. The (simplified) pseudocode of this approach is presented in 2.

Algorithm 2 Motif extraction with SCRIMP++

Input: m - motif length, $E^1, ..., E^{19}$, where E^i is a set of all i^{th} electrodes
Output: motifs[] - a list of extracted motifs

for $i = 1$ *to 19* **do**
  for *each* e *in* E^i **do**
    approx_mp $\leftarrow scrump(e, m, percentage = 0.05)$
    $motif_{idx} \leftarrow motifs(e, approx\_mp, min\_neighbours = 3, max\_motifs = 1)$
    $top\_motif \leftarrow e[motif_{idx} : motif_{idx} + m]$
    $motifs.append(top\_motif)$
  end
end

OSTINATO This algorithm finds a consensus motif of length m across n time series. The input parameters of the OSTINATO algorithm are a set of time series for which we want to obtain the consensus motif and the desired motif length. In our case, this translates to finding a consensus motif across the same electrodes within each class, since we would like to extract motifs that are common for each patient group. In order to add another separation level in the grouping of the time series for the consensus search, we additionally separate the electrodes by the patients' gender.

This setting implies that for each of the 19 electrodes, we run the algorithm five times, for each gender and class combination, resulting in fewer motif candidates. To have the possibility of obtaining more than one motif within a class and among the same gender of an electrode, we run the OSTINATO several times by sampling one-third of the time series before feeding it as an input the to algorithm. STUMPY's method *ostinato* receives a list of the chosen time series and the desired motif length as an input and computes the radius, the index of the time series of the seed motif, and its location within the time series. The pseudocode for this approach is presented in Algorithm 3.

Algorithm 3 Motif extraction with OSTINATO

Input: m - motif length, $P = P_1, \ldots, P_n$, where P_i contains all 19 electrodes of the i^{th} patient, $G = G_1, \ldots, G_n$ is a binary array where G_i represents the gender of the i^{th} patient, repetitions - the number of times to perform sampling

Output: motifs[] - a list of extracted motifs

for $i = 1$ *to 19* **do**
  for j *in {"female", "male"}* **do**
    electrodes $\leftarrow$ $get\_electrodes(P, G, electrode = i, gender = j)$
    for $k = 1$ *to repetitions* **do**
      $\text{electrodes}_{subset} \leftarrow sample(electrodes, rate = \frac{1}{3})$
      $radius, electrode_{idx}, location_{idx} \leftarrow ostinato(electrodes_{subset}, m)$
      $motif_{seed} = electrodes_{subset}[electrode_{idx}][location_{idx} : location_{idx} + m]$
      $motifs.append(motif_{seed})$
    end
  end
end

5 Experiments

In this Chapter, we present the results of the experimental work conducted in the thesis. The experiments can be divided into two main parts; in Section 5.1, we focus on the identification and selection of discriminatory motifs. In Section 5.2, we depict the patient profiling and present the results of the classifiers that are trained to distinguish between the responders and non-responders.

5.1 Motif extraction

As mentioned in Chapter 4, the experiments are conducted on different frequency bands (beta, alpha, and theta) and different motif lengths. Since both motif discovery algorithms, SCRIMP++ and OSTINATO require the motif length as an input parameter, in the first stage we analyze how well the discovered motifs differentiate between the two classes for different motif lengths. This initial higher-level overview of the obtained motifs and their computed difference scores helps in obtaining a sense of which combinations of frequency band and motif length appear promising later in the classification stage.

For each frequency band, we run the motif discovery algorithms on the following motif lengths: 50, 100, 250, 500, 1000, and 2000. Due to the relatively large number of motif extraction results for each combination of frequency band and motif length, we compute and analyze the *difference score* (described in Algorithm 1) of the motifs to get an overview of the discriminatory power within each frequency band and motif lengths, as well as to spot any potential differences within the bands. The following plots refer to the results obtained from the SCRIMP++ algorithm, as both algorithms yield similar results regarding the distribution of the difference score, hence analyzing the results from SCRIMP++ serves the purpose of obtaining a better understanding of the score. Additionally, we have normalized the difference score from 0 to 1 for visualization purposes.

An example of the score of a motif from class 0 is depicted in Figure 5.1 and Figure 5.2, where a motif is shown together with its three closest matches from each class (for readability and clarity, we do not show all of the matches within one signal, but rather only the closest match). Similarly, examples of the scores from class 1 are depicted in Figure 5.3 and Figure 5.4. 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. In Figures 5.2 and 5.4, we have 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.

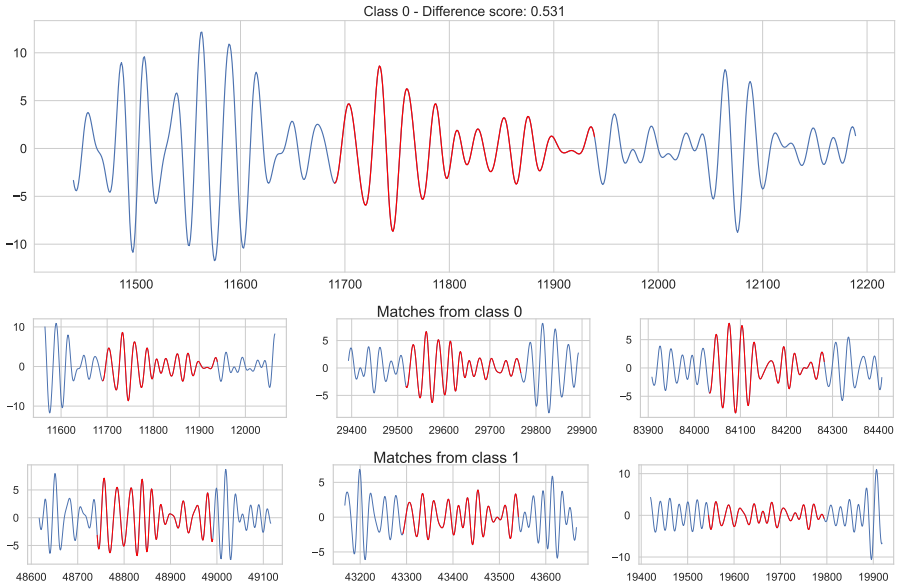


Figure 5.1: Example of a motif from the beta band and length 250 with a higher difference score, together with three examples matches across the two classes.

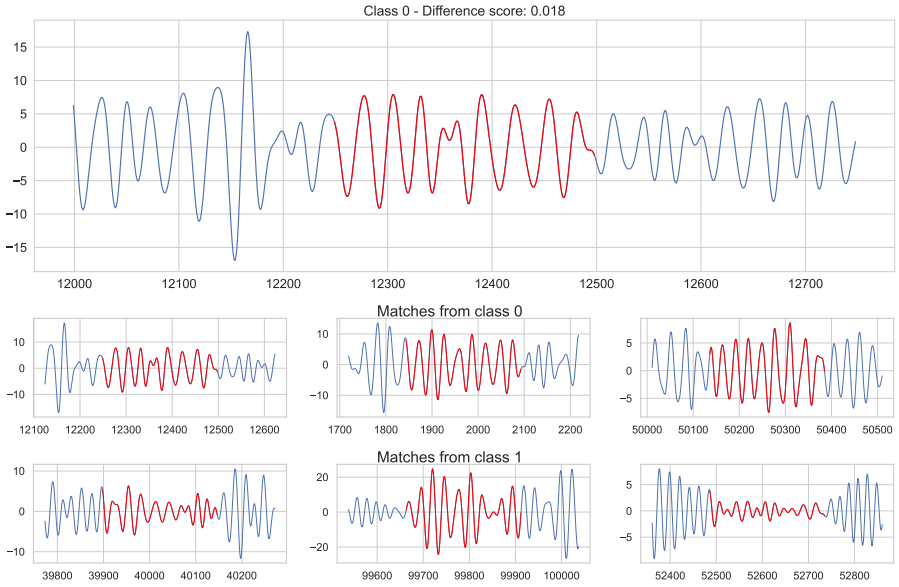


Figure 5.2: Example of a motif from the beta band and length 250 with a lower difference score, together with three examples matches across the two classes.

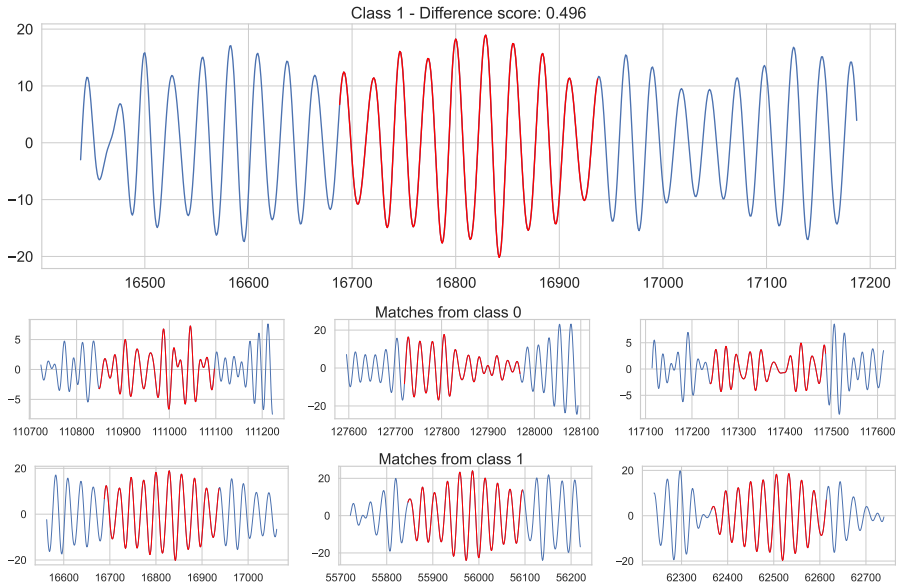


Figure 5.3: Example of a motif from the alpha band and length 250 with a higher difference score, together with three examples matches across the two classes.

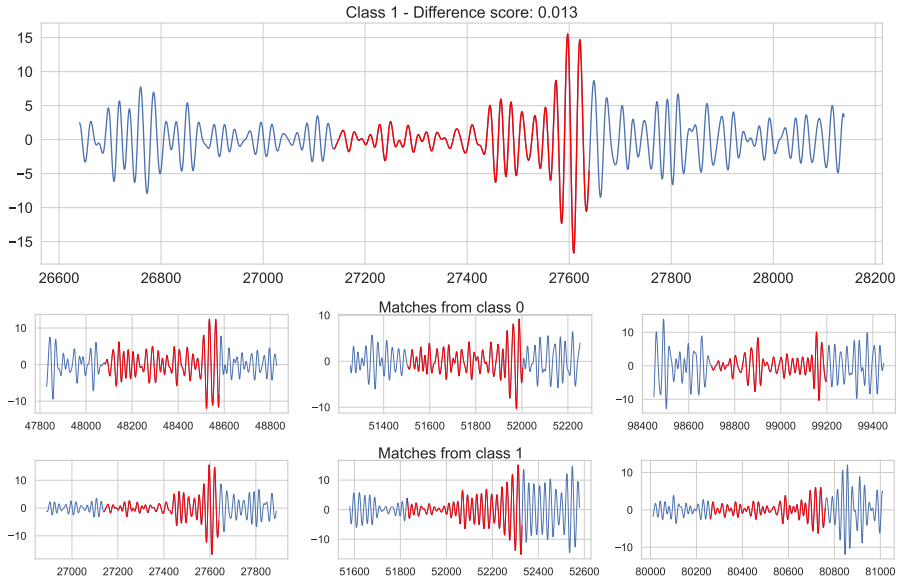


Figure 5.4: Example of a motif from the alpha band and length 500 with a lower difference score, together with three examples matches across the two classes.

In order to obtain a general overview of the *difference scores*, we look at the mean of the scores for each band and motif length. Before computing the mean, we first group the motif by its electrode of origin, in order to obtain further insights. This overview is depicted in Figure 5.5, where one can already spot differences between the bands. We can immediately observe that alpha has the overall highest difference score, especially from motifs with a length 1000. It is important to highlight that the difference score is proportionate to the motif length itself since we are using the normalized Euclidean distance in the computation, hence for shorter motifs we would expect lower values. Hence, if we compare the values for the shorter motifs across the bands, we can observe that the beta band has on average relatively higher discriminatory power than the rest of the bands. The comparison of the difference scores within each motif length can be spotted in Figure 5.6, where the x-axis is sorted by the motif length and shows the frequency bands next to each other, for easier comparison.

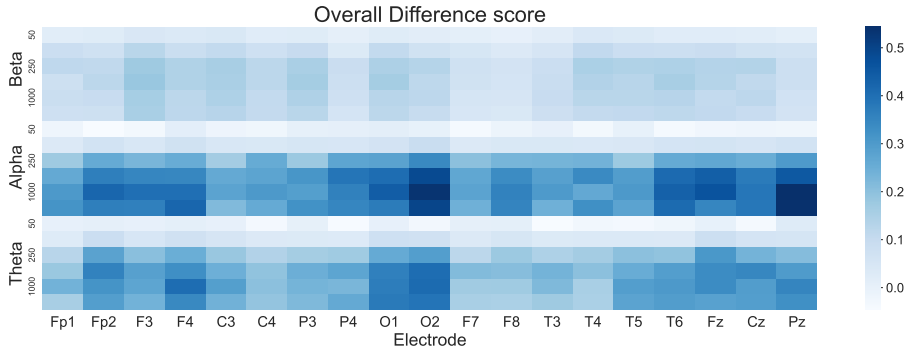


Figure 5.5: Heatmap presenting the mean difference scores across each electrode. The y-axis is sorted by the frequency band, where the rows represent the different motif lengths (sorted in ascending order).

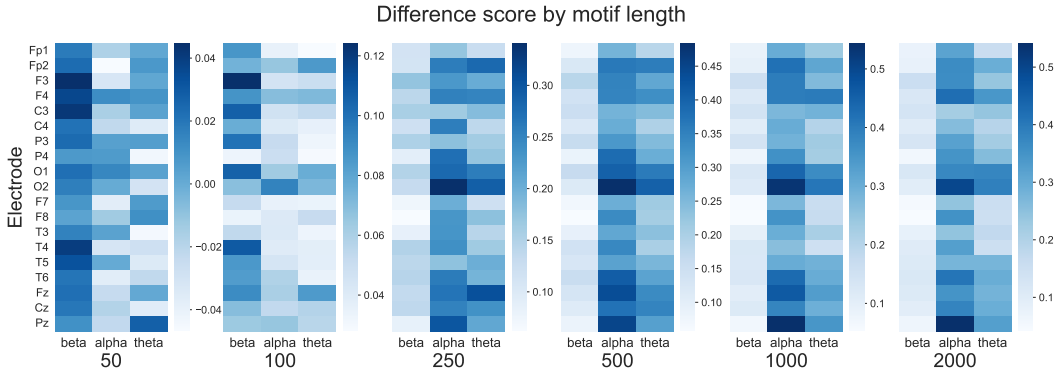


Figure 5.6: Heatmap presenting the mean difference scores across each electrode (y-axis). The x-axis is sorted and grouped by the motif length, where each group has all three frequency bands.

In order to obtain a better understanding of which motif length would have potentially a larger discriminatory power for each frequency band separately, in Figure 5.7 we visualize separate heatmaps for each band. As for the beta band, the motif lengths 250, 500, and 1000 contain higher difference scores, while for the alpha and theta bands, higher values can be spotted for motifs of a larger length.

We can also observe that there are differences across the scores across the electrodes; for the beta band, the highest values can be spotted in the 3rd electrode (channel F3), followed by electrodes: C3, P3, O1, T4 and T6; for the alpha band we can see that the electrodes O2 and Pz have the highest scores, while for the theta band, we have the electrodes: F4, O1 and O2.

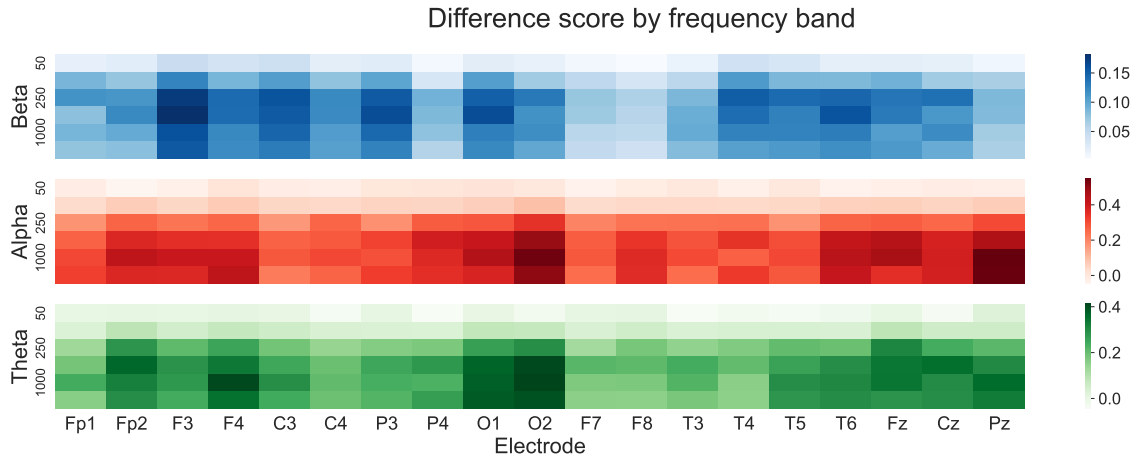


Figure 5.7: Heatmaps presenting the mean difference scores across each electrode within each frequency band. The y-axis represents the different motif lengths (sorted in ascending order).

5.2 Classification

In this stage, we construct a feature matrix with a subset of the obtained motifs. Since for every combination of frequency band and motif length, there is a large set of motifs, we first have to reduce this number by using the ones that have the largest discriminatory power, i.e., difference score.

Imbalance The training dataset is imbalanced regarding the distribution of the class label and gender. Therefore, it is of crucial importance to have each combination of class and gender equally represented in the feature matrix. This means that when selecting a smaller subset of the motifs, we choose the same number of motifs for each combination. In the following paragraphs, we also include an example of not considering the label and gender imbalance, by constructing the feature space based on the overall best (highest) difference score (Figure 5.8).

Since the feature space increases with the number of chosen motifs, ideally we would want this number 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 have 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.

In order to obtain a better overview of constructed feature space, we project the obtained patient profiles to a lower dimensional space using Linear Discriminant Analysis (LDA) [Fis36]. 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.

To highlight the impact of the gender imbalance in the data, we show an example of a constructed feature matrix using the motifs from the alpha band with the highest difference score. Figure 5.8 depicts the projected values of an imbalanced feature matrix. The boxplot on the left represents the distribution of the projected values among the two classes and one might conclude there is some degree of linear separability between the classes, but still some overlap. The visualization on the right includes a scatter plot of the projected values, separated by gender. Here, we can observe that the classes are not well separated for both genders.

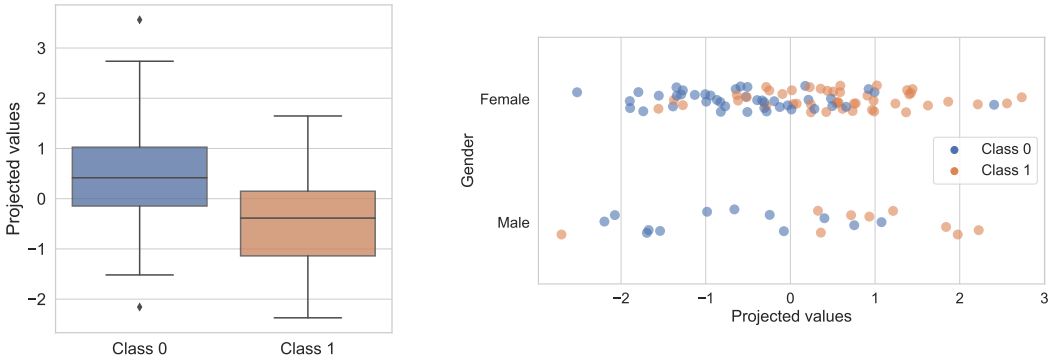


Figure 5.8: Example of the distribution of the projected values of an imbalanced training feature matrix of the alpha band, motif length 500.

The case of a balanced feature matrix is depicted in Figure 5.9, where for each band we show one example of the visualization of the projected values, choosing the motif length that obtained a promising linear separability. One can conclude that the two classes have little overlap in the transformed space, except for the male patients in the case of the alpha band.

The plots are consistent with the observation spotted in the previous section, that the shorter motif lengths appear to be suitable for the beta band, while the longer motif lengths could be more suitable for the alpha and theta bands.

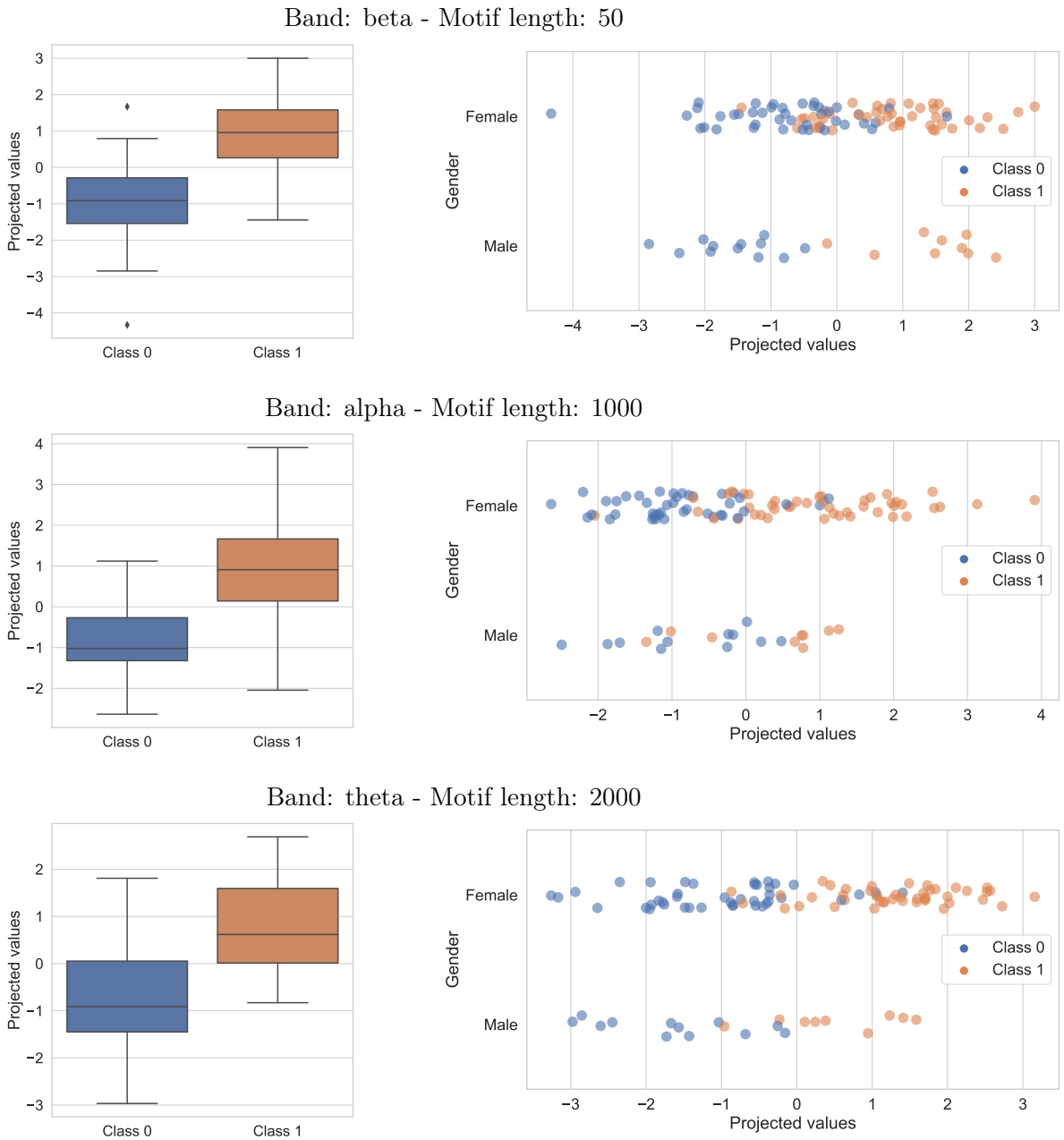


Figure 5.9: Examples of the distribution of the projected values of the training feature matrix of each band.

Training and Evaluation When creating the feature matrices, it is important to highlight that the patient profile is constructed using the reduced set of motifs obtained from only the training set (as described in Section 4.2). The same applies to the patients from the evaluation and final test set. The hyperparameter tuning is done by cross validation on the training set and then later evaluated on the evaluation set. The evaluation set is additionally used in the feature selection process, where we decide on the optimal number of features to keep. Once we obtain the best hyperparameters from the cross-validation, we train different classifiers for each band and length combination, together with the reduced feature space.

Table 5.1 contains the results for each combination obtained by the best classifier (ranked by the highest accuracy score on the evaluation set). To have better clarity and readability, we will include one visualization per frequency band that has the best results. According to the table, in the case of the alpha and beta frequency bands, we have the best evaluation score for the motif length 500, while in the case of the beta band, we have the motif length 1000.

Band	Length	Training		Validation	
		Accuracy	F1	Accuracy	F1
Alpha	50	0.639	0.649	0.692	0.667
	100	0.769	0.790	0.769	0.769
	250	0.741	0.754	0.731	0.741
	500	**0.778**	**0.806**	**0.769**	**0.786**
	1000	0.685	0.702	0.615	0.583
	2000	0.750	0.769	0.692	0.733
Beta	50	0.824	0.800	0.615	0.583
	100	0.815	0.825	0.615	0.643
	250	0.769	0.783	0.654	0.667
	500	**0.806**	**0.811**	**0.692**	**0.714**
	1000	0.602	0.590	0.654	0.640
	2000	0.815	0.821	0.631	0.620
Theta	50	0.750	0.787	0.615	0.615
	100	0.741	0.731	0.769	0.769
	250	0.778	0.793	0.731	0.759
	500	0.694	0.718	0.692	0.714
	1000	**0.833**	**0.845**	**0.769**	**0.789**
	2000	0.694	0.686	0.731	0.696

SCRIMP++

Band	Length	Training		Validation	
		Accuracy	F1	Accuracy	F1
Alpha	50	0.611	0.7	0.538	0.647
	100	0.824	0.84	0.654	0.667
	250	**0.787**	**0.8**	**0.769**	**0.727**
	500	0.787	0.813	0.577	0.667
	1000	0.806	0.796	0.692	0.667
	2000	0.759	0.783	0.615	0.667
Beta	50	0.796	0.796	0.538	0.6
	100	0.787	0.793	0.615	0.545
	250	0.843	0.857	0.538	0.538
	500	**0.796**	**0.804**	**0.692**	**0.667**
	1000	0.611	0.72	0.5	0.649
	2000	0.863	0.863	0.631	0.596
Theta	50	0.611	0.644	0.654	0.69
	100	0.583	0.516	0.654	0.609
	250	0.787	0.785	0.769	0.769
	500	0.796	0.804	0.692	0.714
	1000	0.833	0.82	0.731	0.759
	2000	**0.806**	**0.814**	**0.769**	**0.8**

OSTINATO

Table 5.1: Table with the best training and evaluation results for each frequency band and motif length, evaluated by the accuracy and F1 score.

Since we have a binary classification problem, an accuracy score higher than just 0.5 is important because such score can be achieved by a random guessing. From the training and evaluation results, one can observe that overall, the classifiers learn and generalize well, aso given that the evaluation scores are similar to the training scores. Balanced scores often result from models that have learned meaningful features and relationships within the data.

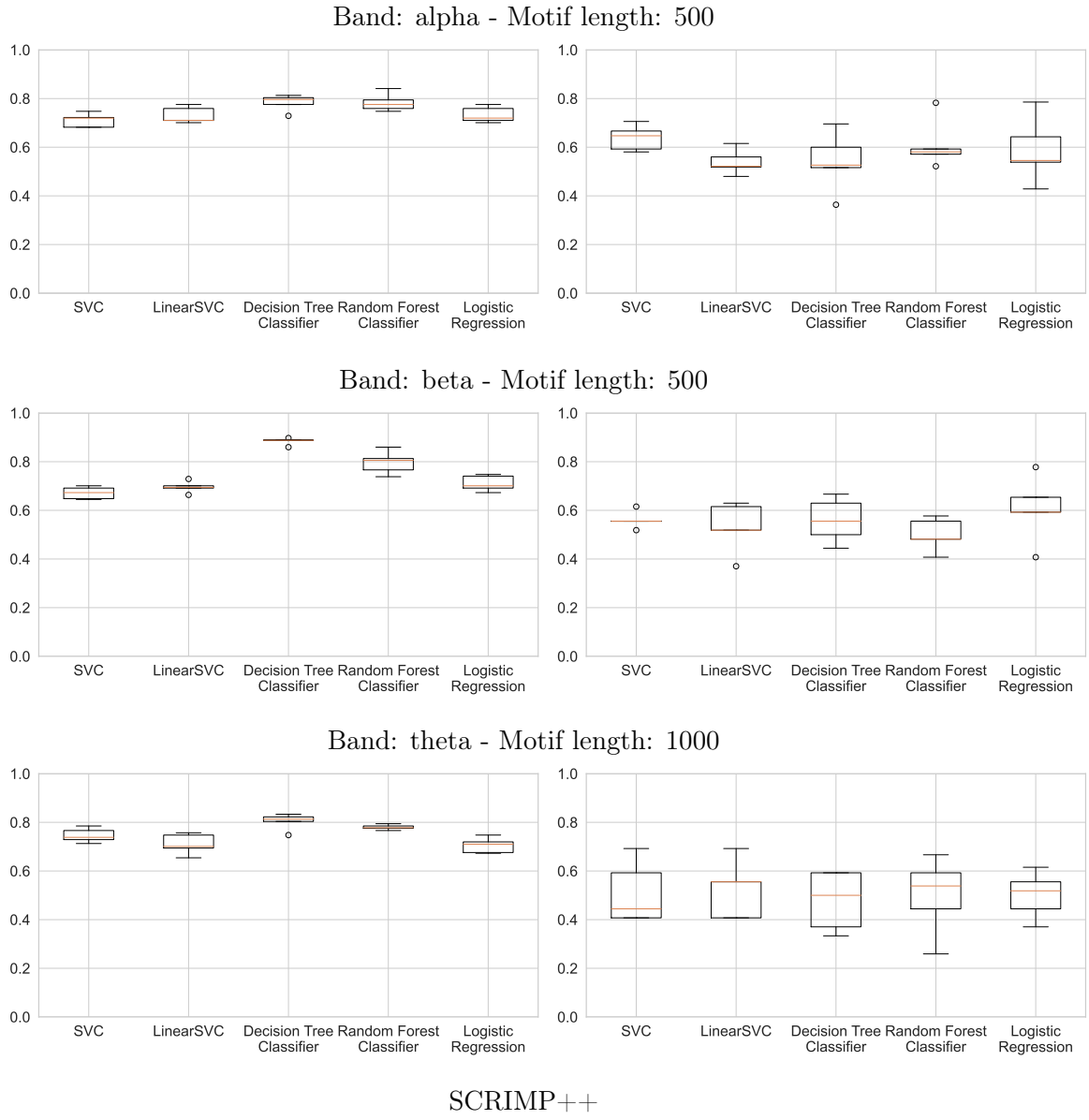


Figure 5.10: Accuracy scores of the training (left) and validation (right) from the k-Fold Cross Validation

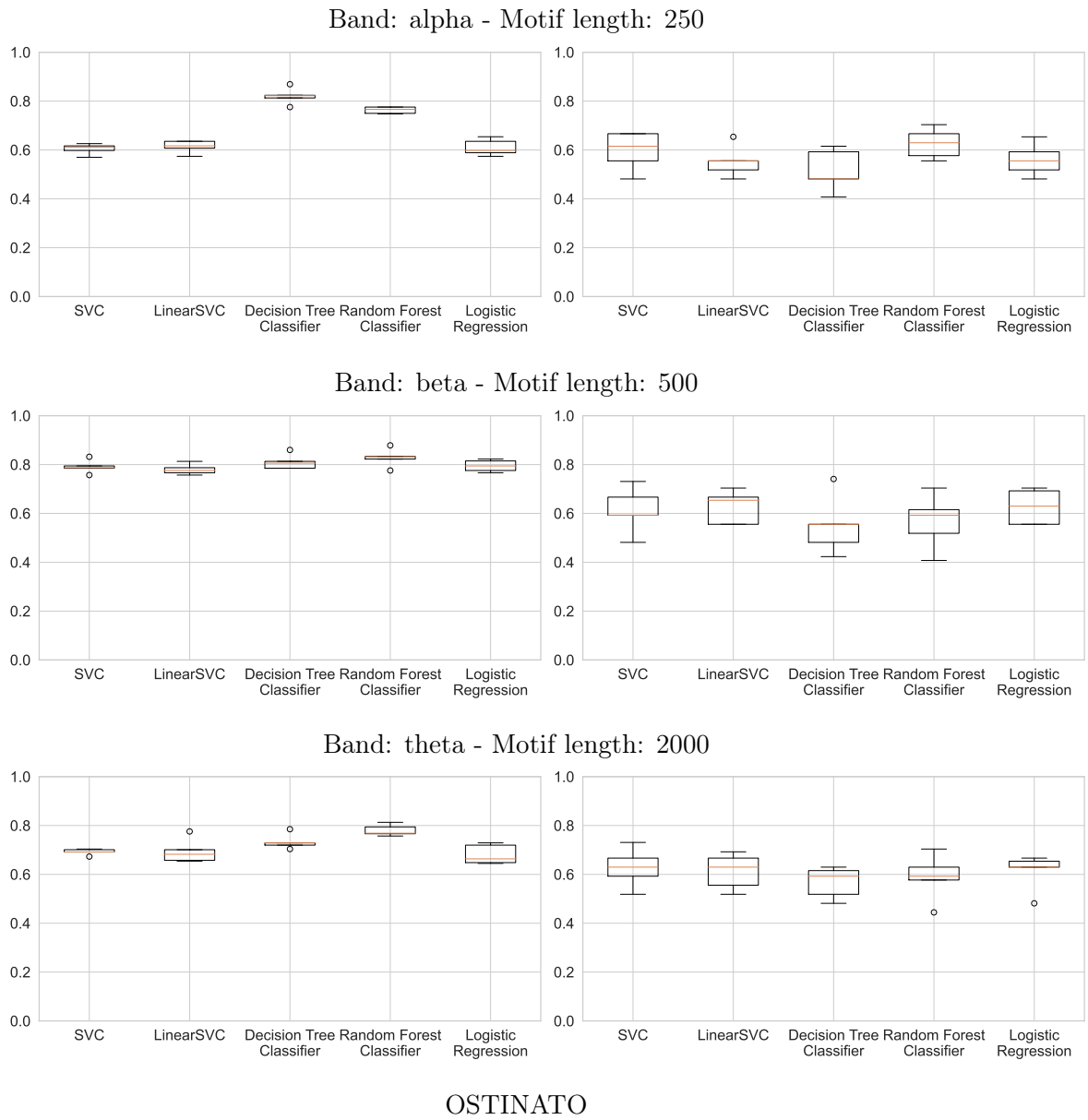


Figure 5.11: Accuracy scores of the training (left) and validation (right) from the k-Fold Cross Validation

By examining the variation in the evaluation scores across the K folds, we can have an idea of the classifier's stability and variance in performance. Figures 5.10 and 5.11 show the results obtained from K-fold cross-validation (using $k = 5$), where each boxplot depicts the K pairs of training and validation sets. We can observe that the training scores seem quite stable, with some cases with greater variance in the evaluation accuracies.

Feature selection In Figure 5.12, one can observe the number of times motifs from each electrode were chosen among the most important features for both motif discovery algorithms. The electrodes with the most count, i.e., with the most important motifs originating from them, are relatively consistent among the two algorithms. The motifs from the electrode (channel) O2 seem to be the most important within the alpha band, while in the theta band, we have O2 and Fp2. For the beta band, we have the electrodes F3, C3, and P3 as the ones with the highest count.

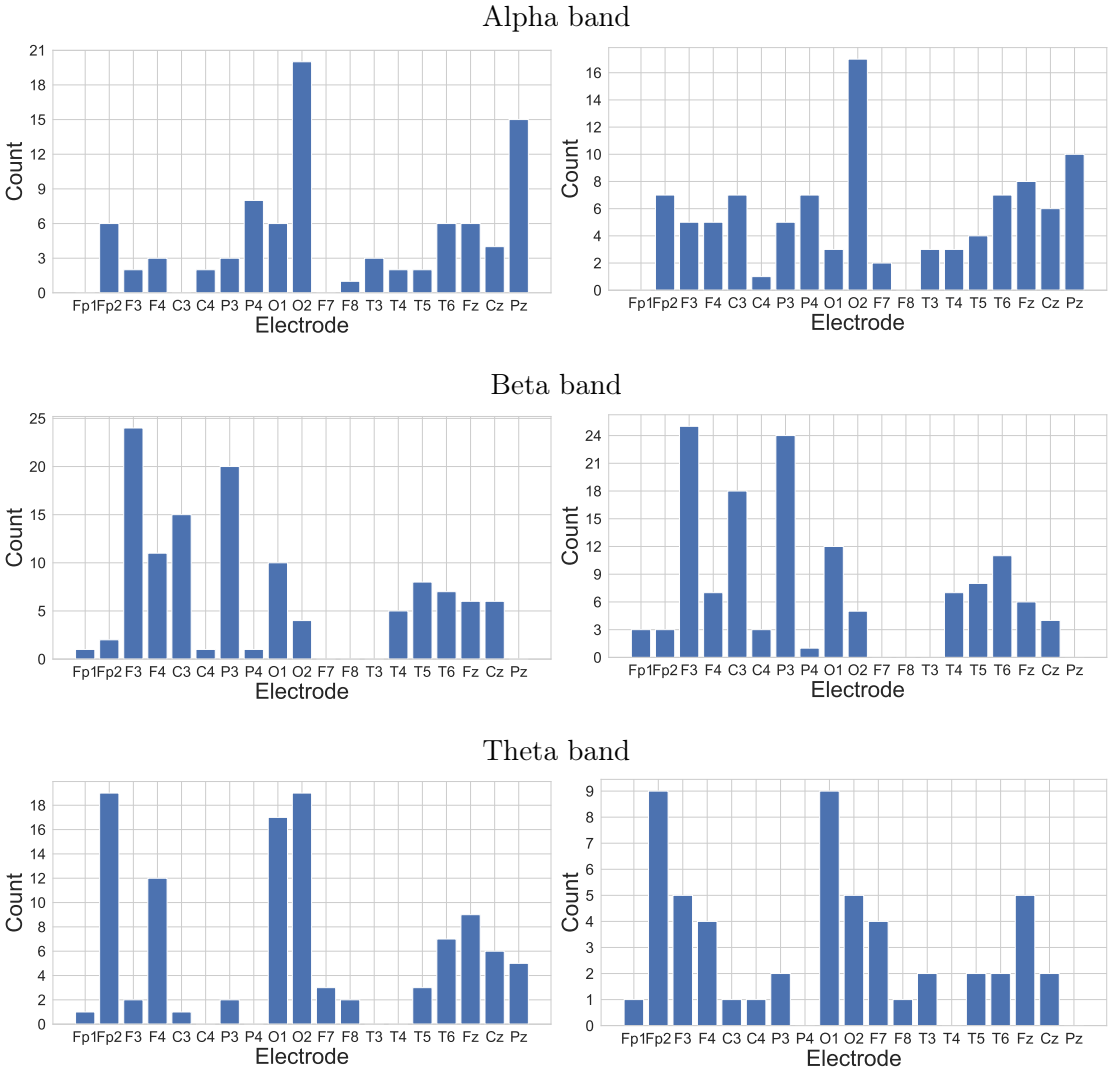


Figure 5.12: Frequency count of the selected motifs across all classifiers for both algorithms (SCRIMP++ on the plots on the left, OSTINATO on the plots on the right).

5.3 Evaluation

The final evaluation for the best models is conducted on a separate EEG recording set, as described in Section 4.1. It is important to highlight that this dataset was not used in the motif extraction process, nor in the model training and parameter tuning. In this step, we perform the final training using the best parameters from the cross-validation and the reduced feature space on both the training and validation sets. The results of the final evaluation can be found in Table 5.2.

Band	Length	Training		Final validation	
		Accuracy	F1	Accuracy	F1
Alpha	50	0.507	0.486	0.595	0.475
	100	0.791	0.821	0.595	0.514
	250	0.612	0.552	0.500	0.323
	500	0.687	0.687	0.595	0.622
	1000	**0.716**	**0.716**	**0.738**	**0.744**
	2000	0.619	0.648	0.548	0.578
Beta	50	0.507	0.521	0.544	0.531
	100	0.612	0.612	0.429	0.400
	250	0.634	0.620	0.548	0.537
	500	0.672	0.656	0.571	0.526
	1000	**0.604**	**0.500**	**0.595**	**0.513**
	2000	0.746	0.742	0.571	0.526
Theta	50	0.507	0.386	0.595	0.212
	100	0.507	0.371	0.500	0.284
	250	0.649	0.624	0.502	0.303
	500	0.828	0.848	0.548	0.512
	1000	0.791	0.759	0.595	0.387
	2000	**0.694**	**0.549**	**0.643**	**0.285**

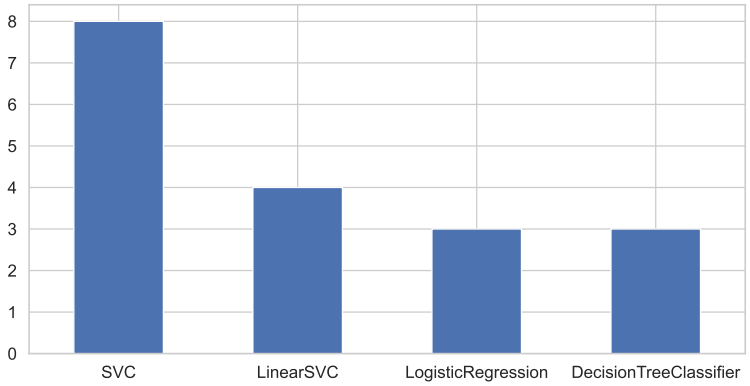
SCRIMP++

Band	Length	Training		Final validation	
		Accuracy	F1	Accuracy	F1
Alpha	50	0.59	0.682	0.5	0.588
	100	0.776	0.805	0.595	0.622
	250	0.604	0.576	0.476	0.476
	500	0.672	0.676	0.595	0.585
	1000	**0.754**	**0.756**	**0.619**	**0.6**
	2000	0.619	0.648	0.548	0.578
Beta	50	0.761	0.784	0.452	0.489
	100	0.619	0.605	0.476	0.421
	250	0.634	0.62	0.548	0.537
	500	**0.754**	**0.748**	**0.548**	**0.558**
	1000	0.664	0.737	0.333	0.481
	2000	0.843	0.84	0.452	0.378
Theta	50	0.507	0.0	0.595	0.0
	100	0.507	0.487	0.595	0.281
	250	0.791	0.816	0.476	0.421
	500	0.619	0.611	0.5	0.4
	1000	0.522	0.521	0.595	0.334
	2000	**0.687**	**0.533**	**0.643**	**0.211**

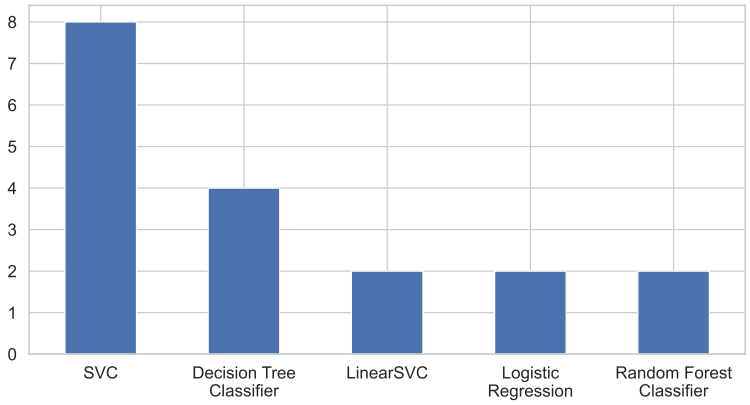
OSTINATO

Table 5.2: Table with the best results on the final testing set for each frequency band and motif length, evaluated by the accuracy and F1 score.

Overall, we can observe that for some cases the training holds slightly worse results compared to the results from the training process. We have the best results for the alpha band with a motif length of 1000, with an accuracy of 0.738 and an F1 score of 0.744. The second best result for this band is the length 500, with an F1 score of 0.622. In the case of the beta band, we have the best accuracy result for the length 1000, with 0.595 and a slightly worse F1 score of 0.513. The theta has the best-performing classifier with a motif length of 2000, with an accuracy of 0.643, however, the F1 score of 0.285 indicates that it performs well on only one of the classes. As a second-best result, we can consider the motif length of 2000, with an accuracy of 0.595 and an F1 score of 0.387.



SCRIMP++



OSTINATO

Figure 5.13: The number of times the model (x-axis) obtained the best accuracy score on the final testing set for each band and motif length combination.

Figure 5.13 shows which model has performed the best among all motif lengths and bands. In the case of SCRIMP++, both SVM classifiers, the SVC and LinearSVC provide the best performance in most cases, i.e., in two-thirds of the cases. As for OSTINATO, both SVM and Decision Tree classifiers are most frequently the best models. Regarding the hyperparameters (which were set in the training phase), in the case of the LinearSVC model, using the L1 norm for all cases gave the highest accuracy and F1 score, while for the SVC model, the setting which resulted in the best performance was using the linear kernel and very low values of the regularization parameter C (from 0.005 to 0.01). The lower the value of parameter C, the stronger the regularization, hence using such values was necessary to combat overfitting.

Best performing classifier Overall, the best performing model is the SVM classifier, obtained by the motifs with a length of 1000 extracted by the SCRIMP++ algorithm from the alpha band. The evaluation on the final testing set reveals an accuracy score of 0.738 and an F1 score of 0.744 on the final testing set. In this subsection, we provide a further analysis of the classifier's performance and its feature set.

Considering the scores by gender, in the case of female patients, we have an accuracy and F1 score of 0.72 and 0.667, while for males 0.766 and 0.818 respectively. These results are depicted in the confusion matrices found in Figure 5.14, where it can be seen that the predictive model has better results for recognizing responders than non-responders. Based on the overall confusion matrix on the top, we have a specificity, i.e., a negative rate of 0.64. This means that considering only the true non-responders, 16 out of 25 non-responders are classified in the correct class. In the case of true responders, we have a recall of 0.882, since 15 out of 17 responders are classified with the correct class.

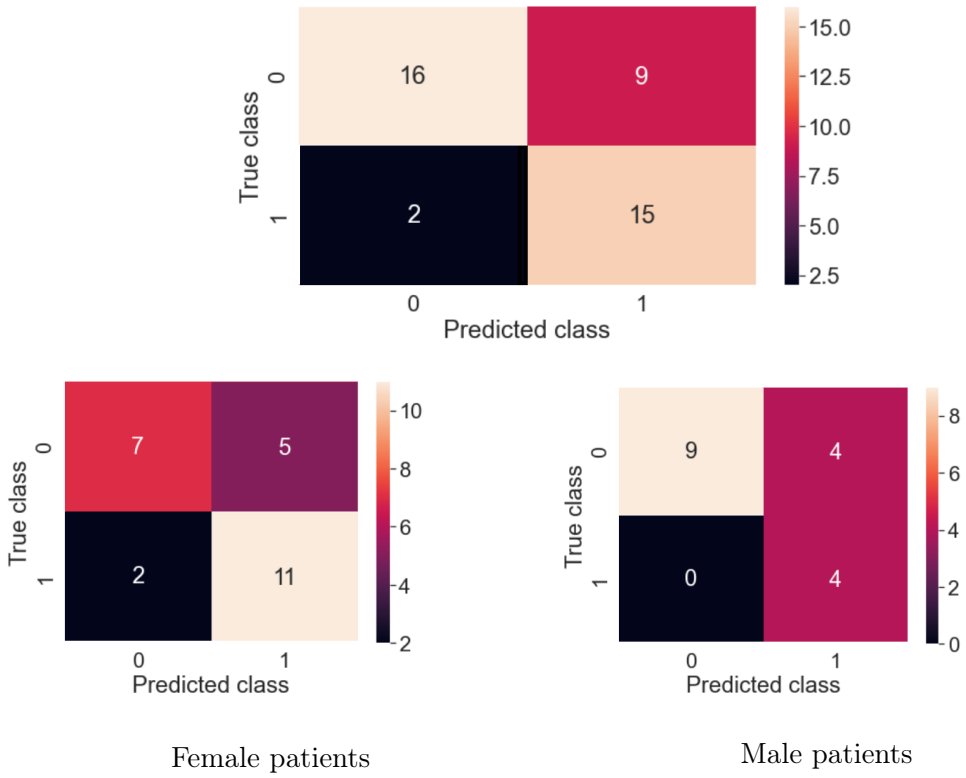


Figure 5.14: Confusion matrices for the final evaluation of the best classifier. The y-axis represents the true labels, while the x-axis the predicted labels.

The final feature set, obtained by the feature selection in the training process, consists of 25 motifs. Figure 5.15 visualizes the counts of the electrode of the motifs, separated by the patient's group (the class label) and the patient's gender in which the motif was found. We have 12 motifs originating from class 0 and 13 motifs from class 1. We can observe that the motifs originating from the non-responders (class 0) are extracted mainly from the channels Pz, O2, and T6, while for the responders - channels O2, O1, and FP2. Regarding gender, there are 8 motifs originating from female patients and 17 from male patients. It can be observed that in the case of male patients, the motifs are extracted mainly from the O2, Pz, and O2 channels, while for females predominantly from O2.

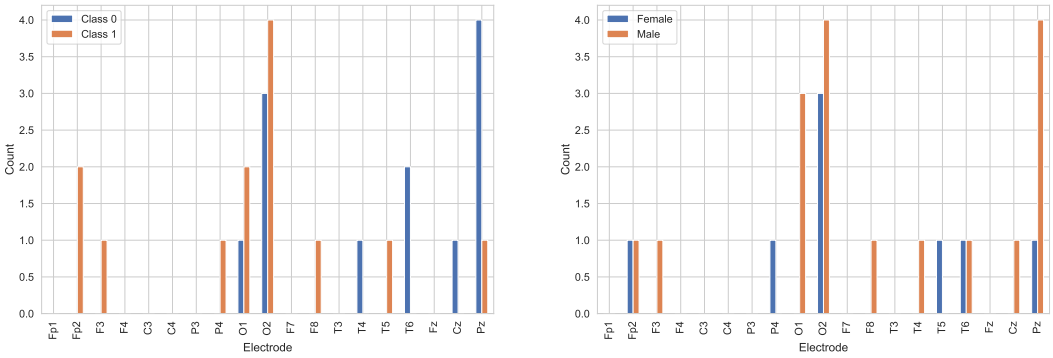


Figure 5.15: The frequency of motifs across their electrode of origin, used for the best model for the alpha band with length 1000, grouped by class (left) and gender (right).

6 Discussion

6.1 Results

The experiments that were conducted in the scope of this thesis considered three frequency bands of the EEG signals, as well as different motif lengths, ranging from 50 up to 2000 samples, which corresponds to 0.2 to 8 seconds. Two motif discovery algorithms were used, SCRIMP++ and OSTINATO.

Both algorithms obtained promising results in the training phase, where the highest F1 score was obtained within the alpha band, with 0.786 and 0.727 on the evaluation set for SCRIMP++ and OSTINATO respectively. The alpha band already showed potential of having the best motif candidates in the exploratory analysis of the difference scores, since it had the highest scores overall, across the motif lengths from 500 to 2000 within the electrodes O2 and Pz. The analysis also revealed that the beta band might be more suitable for shorter motif lengths (ranging up to 500), while the alpha and theta for motifs longer than 500. This was confirmed, considering the F1 scores obtained on the final testing set, having the highest scores for lengths 1000 and 2000 within the alpha and theta band, while 250 and 500 for the beta band.

Among the two algorithms, the electrodes with the highest count, implying their significance in generating motifs, show significant similarity. Specifically, within the alpha band, motifs originating from electrode O2 appear to hold the utmost importance. In the case of SCRIMP++, electrode Pz seems to hold a bigger importance than for OSTINATO, where motifs from channel F4 were more present. In the theta band, electrodes O2 and FP3 emerge as significant contributors, while for the beta band, electrodes F3, C3, and P3 are identified as having the highest motif count.

In regard to the final evaluation results, SCRIMP++ performs slightly better than OSTINATO. This might be possibly due to its ability to accept multiple tunable input parameters, offering room for improvement. In contrast, OSTINATO solely relies on a single input parameter, the motif length, leaving no room for further tunning. However, dealing with a greater number of tunable parameters presents challenges, as it entails conducting extensive experiments to identify and validate the optimal settings. Hence, in the case of SCRIMP++, it was only feasible to explore a limited selection of possible scenarios.

The best performing classifier was obtained using the motifs from the alpha band of length 1000, with an F1 score of 0.744 on the separate testing set. This closely aligns with a recent machine learning meta-analysis review [WPR+22] on the prediction of treatment response using EEG in Major Depressive Disorder (MDD), where the SVM is regarded as the model with the best performance, achieving an accuracy score in the range of 0.782 to

0.826 among the approaches predicting response to antidepressant medication mentioned in the review paper. However, the approaches covered in the review paper have a median size of 86.5, and 60% of them use cross-validation instead of a separate testing set to report on the final results.

In the scope of this project, a thesis work [PPS23] that uses the same EEG database and has the same goal of predicting the patient outcome, computes the Granger-causal graphs across the same frequency bands, training the classifiers for each gender separately. As we have significantly more female patients than male patients in the initial training set (approximately 70% female), the results on the testing set were 0.6 accuracy for females and 0.2 for males. The assertion that females and males exhibit distinctions is consistent with prior research. Earlier studies have not only identified substantial variations in EEG recordings between male and female subjects (e.g., [ZYL+21]), but they have also suggested that there could be divergent biomarkers for predicting treatment outcomes in men and women (e.g., [BBBH20]). Hence, addressing the gender imbalance in our methodology was crucial in obtaining more robust prediction results.

6.2 Limitations of the approach

The experimental workflow analyzes motifs up to a maximum length of 2000, which is 8 seconds. While SCRIMP++ offers significant speed improvements through approximate matrix profiling computation, OSTINATO lacks this feature, making it challenging to explore longer motif lengths. Although the literature is not consistent in stating which frequency bands are significant for major depressive disorder, we selected only three frequency bands, based on recent relevant work, as well as considering the constraints of conducting a large amount of experiments.

In our proposed application of OSTINATO, we apply the consensus motif search on a subset of the signals that belong to a group (same class and same gender). Hence, we are assuming that only patients with the same label and gender have common motifs. On the other hand, having these groups is necessary to significantly reduce the computation time, which was a major challenge when using OSTINATO. To further reduce the computation time and include the possibility of finding more motifs, we further reduce the search space by randomly sampling the signals. As each sample may yield different outcomes, i.e., allows the possibility of discovering more motifs, it might be challenging to provide consistent or stable estimates, as the seed motif can vary.

We trained the classifier exclusively using motifs of the same length and 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 behave differently 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 Conclusion

With the aim of predicting treatment outcomes of depressive patients, this thesis investigates the application of motif discovery algorithms on patient EEG recordings. The experiments were conducted on three frequency bands, alpha, beta and theta, and for different motif lengths.

The results from the final evaluation reveal that motifs appear to possess some level of discriminatory power over non-responsive and responsive patients, especially within the alpha frequency band. The motifs obtained within the alpha band that were chosen as the most important by the classifier originated mainly from the O2 and Pz channels.

7.1 Contributions

To the best of our knowledge, this thesis is the first to apply motif discovery for extracting features to predict the outcome of depression treatment. Recent attempts to classify the treatment outcome are based on feature extraction techniques that utilize the stationary properties of the EEG signals. The research in this thesis is a crucial step in figuring out how to apply state-of-the-art motif discovery algorithms and identify potential reoccurring patterns in the EEG signals that could indicate an earlier detection of treatment response.

Numerous studies that analyze EEG signals in regards to depression, report on smaller datasets. The database that we use consists of 176 patients, hence it contributes to the limited body of research examining how classification methods perform when applied to a larger dataset of individuals with depression. The database contains a separate testing set, which allows the assessment of the final classifiers using an independent test set. While existing literature often employs cross-validation methods to identify optimal classification techniques for similar problems, it frequently lacks validation on distinct test datasets.

Accounting not only for the class imbalance but also for the gender imbalance in the training phase proved to be crucial in achieving good performance across all experiments. In our proposed workflow we divide the initial training set into a training and validation set which accounts for the gender balancing. We further address this by representing each patient group (class and gender) by the same number of motifs when computing the feature matrix. Deciding on which motifs to include in the feature representation was done using the proposed motif ranking, by computing the *difference score*, which describes how discriminative a motif is between the two classes. The motif ranking proved to be a challenging task, considering the large amount of parameters that depend on it. The current proposed computation serves as an important milestone toward understanding and discovering which improvements could lead to more reliable and robust results.

7.2 Future work

As the field of data mining in psychiatry is continuously evolving, there remain several compelling areas for future investigation and improvement. Since we have covered only three frequency bands, exploring the frequency bands gamma and delta could be of interest, as well as investigating if there are motifs with a longer duration that can also differentiate the two patient groups.

The motif ranking computation, i.e., the difference score, represents a complex problem, given that it depends on several parameters. One of these parameters is the distance metric used, which in this case was the z-normalized Euclidean distance. This metric is proportionate to the length of the time series, hence it is not appropriate for conducting a comparison of similarity between pairs of time series of different lengths. Ideally improving the score would mean having it invariant to the length and having it bounded. In general, exploring other metrics could influence motif ranking in a new direction and lead to valuable insights.

Currently, we have explored the predictive power of the motifs separately by their frequency band and length. By already highlighting that there are motifs with high discriminatory power found among the three frequency bands, combining these motifs within one classifier might help improve the performance. This would increase the information we provide to the classifier and could help to increase robustness.

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 that address this kind of problem could be applied and lead to potential new findings.

Bibliography

[AAAB+21] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, 11(2):796, 2021.

[AEH+15] Martijn Arns, Amit Etkin, Ulrich Hegerl, Leanne M Williams, Charles DeBattista, Donna M Palmer, Paul B Fitzgerald, Anthony Harris, Roger deBeuss, and Evian Gordon. Frontal and rostral anterior cingulate (racc) theta EEG in depression: Implications for treatment outcome? *European Neuropsychopharmacology*, 25(8):1190–1200, 2015.

[BBBH20] Barbora Bučková, Martin Brunovskỳ, Martin Bareš, and Jaroslav Hlinka. Predicting sex from EEG: validity and generalizability of deep-learning-based interpretable classifier. *Frontiers in Neuroscience*, 14:589303, 2020.

[CBD+21] Adam M Chekroud, Julia Bondar, Jaime Delgadillo, Gavin Doherty, Akash Wasil, Marjolein Fokkema, Zachary Cohen, Danielle Belgrave, Robert DeRubeis, Raquel Iniesta, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2):154–170, 2021.

[CFS+18] Andrea Cipriani, Toshi A Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z Atkinson, Yusuke Ogawa, Stefan Leucht, Henricus G Ruhe, Erick H Turner, Julian PT Higgins, et al. 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):420–429, 2018.

[CKL03] Bill Chiu, Eamonn Keogh, and Stefano Lonardi. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–498, 2003.

[ČSSP20] Milena Čukić, Miodrag Stokić, Slobodan Simić, and Dragoljub Pokrajac. The successful discrimination of depression from EEG could be attributed to proper feature extraction and not to a particular classification method. *Cognitive Neurodynamics*, 14:443–455, 2020.

[DM04] Arnaud Delorme and Scott Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[DPVH20] Dieter De Paepe and Sofie Van Hoecke. Mining recurring patterns in real-valued time series using the radius profile. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 984–989. IEEE, 2020.

[DS19] Essam Debie and Kamran Shafi. Implications of the curse of dimensionality for supervised learning classifier systems: Theoretical and empirical analyses. *Pattern Analysis and Applications*, 22:536, 2019.

[Fis36] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[GFS+21] Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H Koenigsberg, and Ronald C Kessler. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *Pharmacoeconomics*, 39(6):653–665, 2021.

[GJG+17] Shiv Gautam, Akhilesh Jain, Manaswi Gautam, Vihang N Vahia, and Sandeep Grover. Clinical practice guidelines for the management of depression. *Indian Journal of Psychiatry*, 59(Suppl 1):S34, 2017.

[GL19a] Yifeng Gao and Jessica Lin. Discovering subdimensional motifs of different lengths in large-scale multivariate time series. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 220–229. IEEE, 2019.

[GL19b] Yifeng Gao and Jessica Lin. HIME: discovering variable-length motifs in large-scale time series. *Knowledge and Information Systems*, 61:513–542, 2019.

[GLL+13] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. MEG and EEG data analysis with mne-python. *Frontiers in Neuroscience*, page 267, 2013.

[GMC+21] Claudia Greco, Olimpia Matarazzo, Gennaro Cordasco, Alessandro Vinciarelli, Zoraida Callejas, and Anna Esposito. Discriminative power of EEG-based biomarkers in major depressive disorder: A systematic review. *IEEE Access*, 9:112850–112870, 2021.

[GWBV02] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[HDO+98] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4):18–28, 1998.

[KMRP00] Jens Kohlmorgen, K-R Müller, Jörn Rittweger, and Klaus Pawelzik. Identification of nonstationary dynamics in physiological recordings. *Biological Cybernetics*, 83(1):73–84, 2000.

[Law19] Sean M Law. STUMPY: A powerful and scalable python library for time series data mining. *Journal of Open Source Software*, 4(39):1504, 2019.

[LLYG15] Yuhong Li, Hou U Leong, Man Lung Yiu, and Zhiguo Gong. Quick-motif: An efficient and scalable framework for exact motif discovery. In *2015 IEEE 31st International Conference on Data Engineering*, pages 579–590. IEEE, 2015.

[LZPK18] Michele Linardi, Yan Zhu, Themis Palpanas, and Eamonn Keogh. Matrix profile x: VALMOD-scalable discovery of variable-length motifs in data series. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1053–1066, 2018.

[MÅ79] Stuart A Montgomery and MARIE Åsberg. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, 134(4):382–389, 1979.

[MC15] Abdullah Mueen and Nikan Chavoshi. Enumeration of time series motifs of all lengths. *Knowledge and Information Systems*, 45(1):105–132, 2015.

[MIES07] David Minnen, Charles Isbell, Irfan Essa, and Thad Starner. Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 601–606. IEEE, 2007.

[MKBS09] Abdullah Mueen, Eamonn Keogh, and Nima Bigdely-Shamlo. Finding time series motifs in disk-resident data. In *2009 Ninth IEEE International Conference on Data Mining*, pages 367–376. IEEE, 2009.

[MKZ+09] Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. Exact discovery of time series motifs. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 473–484. SIAM, 2009.

[NIoMHotCRN22] (21-14727k) National Institute of Mental Health of the Czech Republic (NIMH). Learning synchronization patterns in multivariate neural signals for prediction of response to antidepressants. `https://www.nudz.cz/en/research/clinical-research-program/grants-and-projects/struktury-synchronizace-v-mnohorozmernych-neuralnich-signalech-strojove-uceni-a-predikce`, 2022.

[Off50] Franklin F Offner. The EEG as potential mapping: the value of the average monopolar reference. *Electroencephalography and Clinical Neurophysiology*, 2(2):213–214, 1950.

[OGP+16] Christian Otte, Stefan M Gold, Brenda W Penninx, Carmine M Pariante, Amit Etkin, Maurizio Fava, David C Mohr, and Alan F Schatzberg. Major depressive disorder. *Nature Reviews Disease Primers*, 2(1):1–20, 2016.

[oHME23] Institute of Health Metrics and Evaluation. Global health data exchange. `https://vizhub.healthdata.org/gbd-results/`, 2023. Accessed: 29.04.2023.

[Org23] World Health Organization. Depressive disorder (depression). `https://www.who.int/news-room/fact-sheets/detail/depression`, 2023. Accessed: 29.04.2023.

[oV21] University of Vienna. Learning synchronization patterns in multivariate neural signals for prediction of response to antidepressants. `https://dm.cs.univie.ac.at/research/projects/project/347/`, 2021.

[PPS23] Christina Pacher, Claudia Plant, and Katerina Schindlerova. Analysis of an EEG database of depression patients by means of graphical granger causality, 2023.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of Machine Learning Research*, 12:2825–2830, 2011.

[SL22] Patrick Schäfer and Ulf Leser. Motiflets–fast and accurate detection of motifs in time series. *arXiv preprint arXiv:2206.03735*, 2022.

[TL17] Sahar Torkamani and Volker Lohweg. Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2):e1199, 2017.

[TSB+21] Dekel Taliaz, Amit Spinrad, Ran Barzilay, Zohar Barnett-Itzhaki, Dana Averbuch, Omri Teltsh, Roy Schurr, Sne Darki-Morag, and Bernard Lerer. Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data. *Translational Psychiatry*, 11(1):381, 2021.

[UBA04] Ajumobi Udechukwu, Ken Barker, and Reda Alhajj. Discovering all frequent trends in time series. In *Proceedings of the Winter International Synposium on Information and Communication Technologies*, pages 1–6, 2004.

[WBM+19] Alik S. Widge, M. Taha Bilge, Rebecca Montana, Weilynn Chang, Carolyn I. Rodriguez, Thilo Deckersbach, Linda L. Carpenter, Ned H. Kalin, and Charles B. Nemeroff. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: A meta-analysis. *American Journal of Psychiatry*, 176(1):44–56, 2019. PMID: 30278789.

[WMD+13] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26:275–309, 2013.

[WPR+22] Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational Psychiatry*, 12(1):332, 2022.

[YPS+23] Rui Yuan, S Ali Pourmousavi, Wen L Soong, Giang Nguyen, and Jon AR Liisberg. Irmac: Interpretable refined motifs in binary classification for smart grid applications. *Engineering Applications of Artificial Intelligence*, 117:105588, 2023.

[YZU+16a] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 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)*, pages 1317–1322. Ieee, 2016.

[YZU+16b] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 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)*, pages 1317–1322. IEEE, 2016.

[ZWW22] Mingming Zhang, Peng Wang, and Wei Wang. Efficient consensus motif discovery of all lengths in multiple time series. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part II*, pages 540–555. Springer, 2022.

[ZYL+21] Lulu Zhao, Licai Yang, Baimin Li, Zhonghua Su, and Chengyu Liu. Frontal alpha EEG asymmetry variation of depression patients assessed by entropy measures and lemple–ziv complexity. *Journal of Medical and Biological Engineering*, 41:146–154, 2021.

[ZYZ+18] Yan Zhu, Chin-Chia Michael Yeh, Zachary Zimmerman, Kaveh Kamgar, and Eamonn Keogh. Matrix profile xi: SCRIMP++: time series motif discovery at interactive speeds. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 837–846. IEEE, 2018.

[ZZS+16] Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. 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)*, pages 739–748. IEEE, 2016.