Czech Technical University in Prague

Faculty of Electrical Engineering
Department of Circuit Theory

Multivariate Methods for Identification of Clinical Episodes in Bipolar Affective Disorder from Actigraphy

Master's Thesis

Bc. Carmen-Anna Konicarová

Master's programme: Medical Electronics and Bioinformatics
Branch of study: Signal Processing
Supervisor: **Ing. Eduard Bakštein, PhD.**

Prague, January 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Konicarová Carmen-Anna** Personal ID number: **483185**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Circuit Theory**

Study program: **Medical Electronics and Bioinformatics**

Specialisation: **Signal processing**

II. Master’s thesis details

Master’s thesis title in English:

Multivariate Methods for Identification of Clinical Episodes in Bipolar Affective Disorder from Actigraphy

Master’s thesis title in Czech:

Mnohorozměrné metody identifikace klinických epizod z aktigrafie u bipolární afektivní poruchy

Guidelines:

Bipolar disorder is a serious mental illness in which the clinical condition is associated with disturbances in activity and sleep. In this project, we examine the relationships between clinical status, as measured by clinical self-rating scales, and actigraphy variables such as changes in sleep and physical activity. Previous studies have shown significant but heterogeneous associations between actigraphic parameters and clinical status. In this project, we focus on changes in the structure of the relationships between multiple actigraphic parameters as clinical status changes.
1. Study the issue of motor symptoms and actigraphy vs clinical status in bipolar affective disorder (e.g., Schneider 2021)
2. Study methods for clustering multivariate time series based on covariance structure (e.g., Hallac 2017)
3. Propose an appropriate method and criterion for comparing clustering results and clinical status over time
4. Evaluate the effect of window length and other parameter values on classification accuracy on suitably simulated data for the methods selected in Section 2.
5. Apply the selected methods to data from long-term actigraphic follow-up of patients with bipolar affective disorder. Critically evaluate and discuss the results.

Bibliography / sources:

[1] Schneider, J.: 2021, Long-Term Actigrafphy in Bipolar Disorder:Processing, Analysis, and Applications in Diagnostics, PhD thesis, CTU Prague, https://dspace.cvut.cz/handle/10467/94363
[2] Hallac, D., Vare, S., Boyd, S., & Leskovec, J. (2017). Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. 215–223. https://doi.org/10.1145/3097983.3098060
[3] Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning. Springer, 2009
[4] Schneider, J., Bakštein, E., Kolenič, M., Vostatek, P., Correll, C. U., Novák, D., & Španiel, F. (2022). Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls. CNS Spectrums, 27(1), 82–92. https://doi.org/10.1017/S1092852920001777
[5] Krane-Gartiser K, Henriksen TEG, Morken G, Vaaler A, Fasmer OB. Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder. PLoS One. 2014;9(2). doi:10.1371/journal. pone.0089574

Name and workplace of master's thesis supervisor:

Ing. Eduard Bakštein, Ph.D. Analysis and Interpretation of Biomedical Data FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **26.01.2023** Deadline for master's thesis submission: **09.01.2024**

Assignment valid until: **22.09.2024**

Ing. Eduard Bakštein, Ph.D.	doc. Ing. Radoslav Bortel, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
Supervisor's signature	Head of department's signature	Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student's signature

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, January 2024

Bc. Carmen-Anna Konicarová

Acknowledgment

First of all, I would like to express my sincere gratitude and many thanks to my supervisor Ing. Eduard Bakštein, Ph.D. for a lot of valuable and inspiring advices and constant support during the whole time of working on this thesis. I would also like to thank the whole Mindpax company for providing us with all the data.

Bc. Carmen-Anna Konicarová

Abstract

Previous studies have established relationships between actigraphy-measured activities (amount of activity, sleep and their fragmentation) and clinical status in patients with bipolar disorder (BD). The main aim of this study was to investigate whether the covariance structure between different actigraphy parameters varies across different clinical conditions. First, we investigated an unsupervised approach based on Toeplitz inverse covariance-based clustering. After a comprehensive evaluation of this method on a simulated dataset we found it to be unsuitable for our real data due to short duration of the individual clinical states and low sampling frequency.

Next, we investigated a supervised approach to analyse changes in covariance structure across changing clinical state. To compare the predictive value of covariance structure to that of median and variance of individual actigraphy-derived variables, we validated them with a total of three approaches.

We applied three methods using a Support Vector Machine (SVM) classifier. These methods, based on covariance matrix (COV), median and standard deviation (SD) values of actigraphic features, were combined with two dimension reduction techniques, principal component analysis (PCA) and our own (max-diff) method based on the maximum difference of COV pairs (or median or SD values) of features between clinical conditions. Three validation scenarios were used to evaluate these methods: leave-one-patient-out cross-validation (LOOCV), which distinguished clinical states without time dependence, and classifying each time series in a 7-day window across all patients and for individuals separately. Three different types of distributions were used to partition the training and test data sets: patient, state, and time-based splits. All these validations revealed that the mania-remission was easier to determine than the depression-remission distinction. The results also show a high variability between patients. It is noteworthy that using LOOCV, an average accuracy of 72% for mania-remission and 67% for depression-remission was achieved using the SD method with our dimension reduction method. Time-series classification over a 7-day window on the patient-based split dataset achieved lower accuracy for mania using the SD value method with PCA of 67% and depression using the COV PCA of 59%. The state-split datasets provided significantly better results with 85% accuracy for mania and 59% for depression using the median method with diff feature selection. The time-based split dataset, which was the most dependent, naturally showed the highest accuracy, reaching 90% for mania and 80% for depression. In the case of classification, within each patient separately, especially for mania, the methods achieved almost always perfect classification, and of depression states slightly lower but still relatively high accuracy of around 81%.

While the covariance structure of actigraphy-derived variables shows changes between clinical states in BD, the performance did not substantially exceed that of the compared methods, based on standard deviation and median of the individual variables.

Keywords: bipolar affective disorder, actigraphy, multivariate machine learning methods, feature space dimensionality reduction

Abstrakt

Předchozí studie prokázaly vztah mezi aktivitami měřenými aktigrafií (míru aktivity, spánku a jejich fragmentaci) a klinickým stavem u pacientů s bipolární poruchou (BD). Hlavním cílem této studie bylo zjistit, zda se struktura kovariance mezi různými parametry aktigrafie liší u různých klinických stavů. Nejprve jsme zkoumali nesupervizovaný přístup založený na Toeplitzově inverzním kovariančním shlukování. Po komplexním vyhodnocení této metody na simulovaném souboru dat jsme zjistili, že je pro naše reálná data nevhodná, vzhledem ke krátkému trvání jednotlivých klinických stavů a nízké frekvenci vzorkování.

Dále jsme zkoumali supervizovaný přístup k analýze změn kovarianční struktury při měnícím se klinickém stavu. Abychom porovnali prediktivní hodnotu kovarianční struktury navíc s mediánovou hodnotou a hodnotou rozptylu jednotlivých proměnných odvozených z aktigrafie, validovali jsme je celkem třemi přístupy.

Použili jsme tři metody využívající metodu podpůrných vektorů (Support Vector Machine, SVM). Tyto metody založené na kovarianční matici (COV), mediánu a směrodatné odchylce (SD) hodnot aktigrafických znaků byly kombinovány se dvěma technikami redukce dimenzionality, analýzou hlavních komponent (PCA) a naší vlastní metodou (max-diff) založenou na maximálním rozdílu COV dvojic (nebo mediánů či hodnot SD) příznaků mezi klinickými stavy. K vyhodnocení těchto metod byly použity tři validační scénáře: leave-one-patient-out (LOOCV), který rozlišoval klinické stavy bez časové závislosti, a klasifikace každé časové řady v 7denním okně u všech pacientů a u jednotlivců zvlášť. K rozdělení trénovacích a testovacích souborů dat byly použity tři různé typy dělení: podle pacientů, stavů a času. Všechna tato ověřování odhalila, že rozlišení mánie-remise bylo snadnější než rozlišení deprese-remise. Výsledky také ukazují velkou variabilitu mezi jednotlivými pacienty. Za zmínku stojí, že při použití LOOCV bylo dosaženo průměrné přesnosti 72 % pro mánii a 67 % pro depresi pomocí metody SD s naší metodou redukce dimenze. Klasifikace časových řad v sedmidenním okně na souboru rozdělených dat od pacientů dosáhla nižší přesnosti, pro mánii pomocí metody SD hodnot s PCA 67 % a pro depresi pomocí COV PCA 59 %. Soubory dat rozdělených podle stavu poskytly výrazně lepší výsledky s přesností 85 % pro mánii a 59 % pro depresi při použití metody mediánu s výběrem diferenčních příznaků. Časově rozdělená datová sada, která byla nejvíce závislá, přirozeně vykazovala nejvyšší přesnost, 90 % pro mánii a 80 % pro depresi. V případě klasifikace v rámci každého pacienta zvlášť, zejména u mánie, dosáhly metody téměř vždy dokonalé klasifikace a u stavů deprese o něco nižší, ale stále poměrně solidní přesnosti kolem 81 %.

Ačkoli kovarianční struktura proměnných odvozených z aktigrafie vykazuje změny mezi klinickými stavy u BD, její výkonnost nijak výrazně nepřevyšovala výkonnost srovnávaných metod, založených na mediánu a směrodatné odchylce jednotlivých proměnných.

Klíčová slova: bipolární afektivní porucha, aktigrafie, vícerozměrné metody strojového učení, redukce dimenzionality prostoru příznaků

Contents

List of Tables

List of Figures

List of Acronyms

ACC accuracy. 24, 25, 52, 57, 62, 68, 73

APA American Psychological Association. 5

ARI Adjusted Rand Index. 13, 15, 41, 42

ASERT Aktibipo Self-rating EMA. 29, 33

BD Bipolar Affective Disorder. 1, 3–7, 10, 27, 29, 33, 35, 44, 79, 80, 83

CBT Cognitive Behavioral Therapy. 5

COV Covariance Matrices. xiii, 1, 10, 17–19, 44, 45, 50, 51, 54, 57, 59, 62, 67, 68, 70–72, 74, 76, 79, 83

LOOCV Leave-One-Person-Out Cross-Validation. 17, 18, 24, 50, 51

LTM Long-term Monitoring. 5

MADRS Montgomery-Åsberg Depression Rating Scale. 5, 29, 33

max-diff Maximal-Difference Method. 10, 19, 44, 50, 51, 54, 57, 59, 62, 64, 67–72, 74, 76, 79, 80

MI Mutual Information. 13

MRF Markov Random Field. 11, 12

NIMH National Institute of Mental Health. 29, 33

NMI Normalized Mutual Information. 13, 16

p25 25th percentile. 34

p75 75th percentile. 34

PCA Principal Component Analysis. 10, 17–19, 44, 50, 51, 54, 57, 67, 68, 70–72, 74, 76, 79–81, 83

PPV Positive Predictive Value. 13, 14

RBF Radial Basis Function. 10, 17, 23, 44

RI Rand Index. 15, 16

SD Standard Deviation. xiii, 1, 10, 17–20, 44, 45, 50, 51, 54, 57, 59, 62, 67–72, 74, 76, 79, 80, 83

SEN sensitivity. 24, 52, 57, 62, 68, 73

SPEC specificity. 24, 25, 52, 57, 62, 68, 73

SVM Support Vector Machine. 1, 10, 17, 20, 21, 23, 44

TICC Toeplitz Inverse Covariance-Based Clustering. 10–15, 17, 27, 39, 42, 43, 79

YMRS Young Mania Rating Scale. 5, 29, 33

Chapter 1

Introduction

Monitoring and recording large amounts of data about patients is an important part of modern medicine. Particularly in patients with Bipolar Affective Disorder (BD), whose clinical condition is closely linked to disturbances in activity and sleep behaviour, there is a great research interest in monitoring actigraphy and clinical status - usually measured using clinical self-assessment scales. Previous studies have already shown significant but heterogeneous associations between actigraphic parameters and clinical status. Therefore, in the present study, we decided to investigate changes in the structure of the relationships between actigraphic parameters as a function of changes in the patient's clinical status.

1.1 Goals of the Thesis

The main aims of this thesis are:

- To study the issue of motor symptoms and actigraphy against clinical status in BD.
- To investigate methods for clustering multivariate time series based on covariance structure [1].
- To propose an appropriate method and criterion for comparing the clustering results and clinical status over time.
- Evaluate the impact of the set parameter values on the identification accuracy using appropriate simulated data.
- Application of selected methods to data from long-term actigraphic follow-up of patients with BD and critical evaluation and discussion of the results.
- Beyond the original thesis assignment, the goals were extended with supervised methods – Support Vector Machine (SVM) classifier on COV, median and SD values of actigraphy features.

Chapter 2

Background

The main aim of this chapter is to provide the reader with a very brief introduction to the topic of bipolar affective disorder, as expounded in section 2.1. It further delves into the clinical diagnostics, standard treatment and clinical state assessment associated with the disorder, as detailed in section 2.2, and continues with long-term monitoring using self-assessment, digital and actigraphic monitoring, as elucidated in section 2.3.

2.1 Bipolar Affective Disorder

BD, formerly known as manic-depressive disorder, is a group of severe mental illnesses characterised by periods of extreme mood swings. These fluctuations in mental health include relapses, episodes of emotional highs known as mania or hypomania (less extreme than mania), lows known as depression or mixed states, with remissions (absence or minimal symptoms of both mania and depression). The alternating episodes mentioned above can occur infrequently or several times a year and are more severe than the normal ups and downs experienced by the normal population. They can affect many areas of life, such as sleep, energy levels, behaviour and even clear thinking. In 2019, almost 40 million people worldwide were affected by this severe chronic mood disorder [2].

According to the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5), BD is experienced in three forms: bipolar 1, bipolar 2 and cyclothymic disorder. All types of BD are characterised by episodes of extreme alternating moods. The main difference between bipolar 1 and bipolar 2 disorders is the intensity of the manic episodes. Bipolar 1 disorder cause full manic episode, while bipolar 2 causes less severe hypomania. Another noticeable difference is the prevalence of major depressive episodes. A person with bipolar 1 may or may not experience a major depressive episode whereas person with bipolar 2 will experience at least one major depressive episode. The critical feature of cyclothymic disorder (cyclothymia), a milder form of the bipolar disorder, represents episodes of hypomania alternating with mild depression for at least two years without no symptom-free period lasting longer then 8 weeks. [3]

BD affects men and women equally and the first episode can be any, mania, depression, hypomania. Manic episodes tend to occur more frequently in men, but depression is the more

common first episode for both men and women. [4]

The factors contributing to relapse in BD are not fully known. Recent findings show that there could be association with circadian rhythm dysregulation [5], [6] and sleep disturbance [7]–[10]. Also the pathogenesis of BD is not clearly understood. According to some studies the pathogenesis is associated with mitochondrial dysfunctions [11], [12].

Almost 800,000 people worldwide commit suicide each year, which is about one death every 40 seconds [13]. People who live with BD are at an increased risk of suicide. Among all mental health disorders, BD has the highest suicide rate. This rate is about 10 to 30 times higher in people living with BD than in the general population [14]. Approximately 3–14 % of all suicide deaths are linked to BD [15]. Various studies indicate that 4-19 % of patients with BD eventually commit suicide, and 20-60 % attempt suicide at least once in lifetime [16].

2.2 Clinical Practice

2.2.1 Diagnostics

The diagnosis of BD is typically made through a detailed clinical evaluation, which may include evidence from family members or other sources of third-party information [17]. However, there is currently no laboratory, imaging nor psychological examination or biomarker that can confirm this disease definitively. A manic episode is also crucial for the diagnostic process, but during this period patients do not typically seek medical help, unlike in the depressed state. Therefore, mania often remains unnoticed, and the patient is treated for a different disease than he should. Therefore, as it may be challenging to make an accurate diagnosis, doctors may need to do ongoing evaluations, including tracking mood patterns over time (see section 2.3).

The clinical treatment includes using medication such as mood stabilisers, antidepressants, and antipsychotics, psychological therapy, and electroconvulsive therapy [17]. The initial pharmacological therapy is usually monotherapy by mood stabilisers or antipsychotics. Lithium is the oldest and a commonly used mood stabiliser, but it is only effective in about one-third of patients [18] and requires not only close monitoring of dosage levels and potential side effects but also periodical testing with respect to the renal function, since the effective dosage is only slightly lower than the toxic levels. Other options for monotherapy include valproate or antipsychotics. During acute episodes, these medications may be used in combination with other mood stabilisers and it usually takes longer period of time to adjust the optimal dosage for these medication combinations. Antidepressants are typically not recommended treatment of BD as they may cause rapid cycling or manic episodes [4].

2.2.2 Standard Treatment

The standard treatment for bipolar disorder typically includes a combination of medication and psychotherapy. Medications commonly used to treat bipolar disorder include mood stabilisers such as lithium and valproic acid, as well as atypical antipsychotics such as olanzapine and

quetiapine [3]. For depressive episodes, antidepressants may also be prescribed in combination with mood stabilisers, but they must be used with caution as they can trigger manic episodes in some people with bipolar disorder. It is an evidence-based treatment that has been found to be effective for a variety of mental health conditions.

Psychological intervention is also an important component of treatment for BD. In most cases, they are focused on educating the patient on how to cope with their illness. One of these methods is Cognitive Behavioral Therapy (CBT), which, as stated by the American Psychological Association (APA), is a type of psychotherapy that helps individuals develop coping strategies and problem-solving skills to deal with difficult situations and emotions [3]. CBT is based on the idea that our thoughts, feelings and behaviours are interconnected, and that negative patterns in one area can affect the other [19]. CBT is widely used to treat a variety of mental health conditions, including depression, anxiety disorders, and BD, as well as other conditions, such as chronic pain and insomnia. It is considered a short-term, goal-oriented therapy that takes an average of 20 sessions followed by enhancement one [20].

2.2.3 Clinical State Assessment

Regular clinical examinations are commonly used for a comprehensive assessment of the patient's mental state. Careful assessment of the patient's condition by an expert is an important part of evidence-based practice. Clinical scales should be the most objective approach of evaluation of mental disorders. Manic and depressive symptoms of BD are mostly evaluated separately. The Montgomery-Åsberg Depression Rating Scale (MADRS) [21], the Hamilton Rating Scale for Depression [22], the Inventory for Depressive Symptomatology [23], or the Bipolar Depression Rating Scale [24] are the symptom rating scales used for evaluating depressive symptoms of BD. Manic symptoms may be assessed by the Young Mania Rating Scale (YMRS) [25], the Bech-Rafaelsen Mania Rating Scale [26] , the Clinical-Administered Rating Scale for Mania [27], or the Observer-Rated Scale for Mania [28]. Scales such as the National Institute of Mental Health's Prospective Life Chart Method [29], Clinician Monitoring Form [30], Brief Bipolar Disorder Symptom Scale [31], and Bipolar Inventory of Symptoms Scale [32] evaluate both manic and depressive symptoms together.

2.3 Long-Term Monitoring

Long-term Monitoring (LTM) systems were developed as a result of the need for a finer progression monitoring of patient's disease states over time. These systems may be categorised as subjective self-assessments and objective activity monitoring and behavioural analyses. The optimal LTM system would be a combination of self-assessment and at least one of the two objective measures.

2.3.1 Self-assessment

Self-assessment questionnaires represent a subjective disease progression monitoring approach, through which patients report their current mental state. In practice, this approach is already used to monitor the progress of illness between medical check-ups and also for early warning of an upcoming relapse.

2.3.2 Digital monitoring and actigraphy

Activity monitoring has a long history in sleep studies [33] and in medicine in general [34]. Actigraphy refers to monitoring and collecting data generated by the movements of body parts using wristwatch-like devices called an actigraph [35] that measure acceleration of the body part they are attached to and thus detect movement. It is a non-invasive approach that records and integrates the occurrence and degree of limb movement activity and rest [36] and thus circadian rhythm over time.

The main concept behind these devices is to make them as simple, non-invasive and long battery life as possible to ensure that the resulting monitoring is as continuous as possible without the need for frequent recharging. Another important feature is water resistance, as patients often forget to put the wearable device back on after hygiene or water activities. Patients suffering from psychological disorders face daily not only the symptoms of their illness, but also the discrimination caused by social stigma [37]. The growing popularity of fitness activity trackers and smartwatches in general makes it easier to use these devices discreetly, thereby reducing the potential risk of stigmatisation of mental disorders.

In most studies the actigraphy unit is placed on the wrist of the non-dominant hand. However it can be worn on the wrist, ankle, waist or even finger as a ring. Some authors have reported that the placement of the actigraph on the dominant or non-dominant hand wrist does not have influence on the results of the sleep-wake scoring algorithm in spite of significant differences in activity levels between the dominant and non-dominant hands [38]. Contrastingly, assessment of actigraph placement by Violani and colleagues has indicated differences of motor activity across the night between the dominant and non-dominant writs [39]. Therefore cross-study comparisons would be more accurate if placement was standardised.

Another approach to monitoring psychiatric patients is behavioural analyses examining the impact of smartphone use on disease development and status. Most of these methods are still being studied and their practical use would be severely limited by regulations due to privacy issues. [40]

2.3.3 Activity monitoring in BD

As the manic phase in BD patients is associated with an increase in psychomotor activity and, on the contrary, in BD depression, longitudinal actigraphy is a very promising approach to monitor phase shifts and clinical changes. Previous studies [41]–[43] have found distinctly different patterns of activity in patients with bipolar disorder during episodes of mania and

depression, as well as clear differences between patients and healthy control subjects, as assessed by machine-learning actigraphy-based (linear as well as nonlinear mathematical) models. There is also sufficient evidence of an association between BD and activity, both in remission and relapse episodes [44].

Existing literature indicates that individuals with BD tend to have lower overall motor activity and reduced peak activity compared to healthy controls [45], [46]. Mood instability in BD is associated with increased fragmentation of activity profiles within and across days, resulting increased variability in actigraphy parameters [47]. Sleep quality reduction in BD is is reflected by increased motor activity and prolonged wakefulness during nighttime sleep [48]. The assumption of longer and more variable sleep duration in BD is supported by some studies [7], [49], although conflicting findings exist [50], [51].Longer sleep latency in patients with BD may be manifested by reduced pre-sleep activity and greater activity after sleep onset, leading to greater variability in both sleep latency and restlessness [7], [8], [49].

Chapter 3

Methods

In this chapter we describe the methods used in this thesis. For a better overview, all methods are shown in the following map:

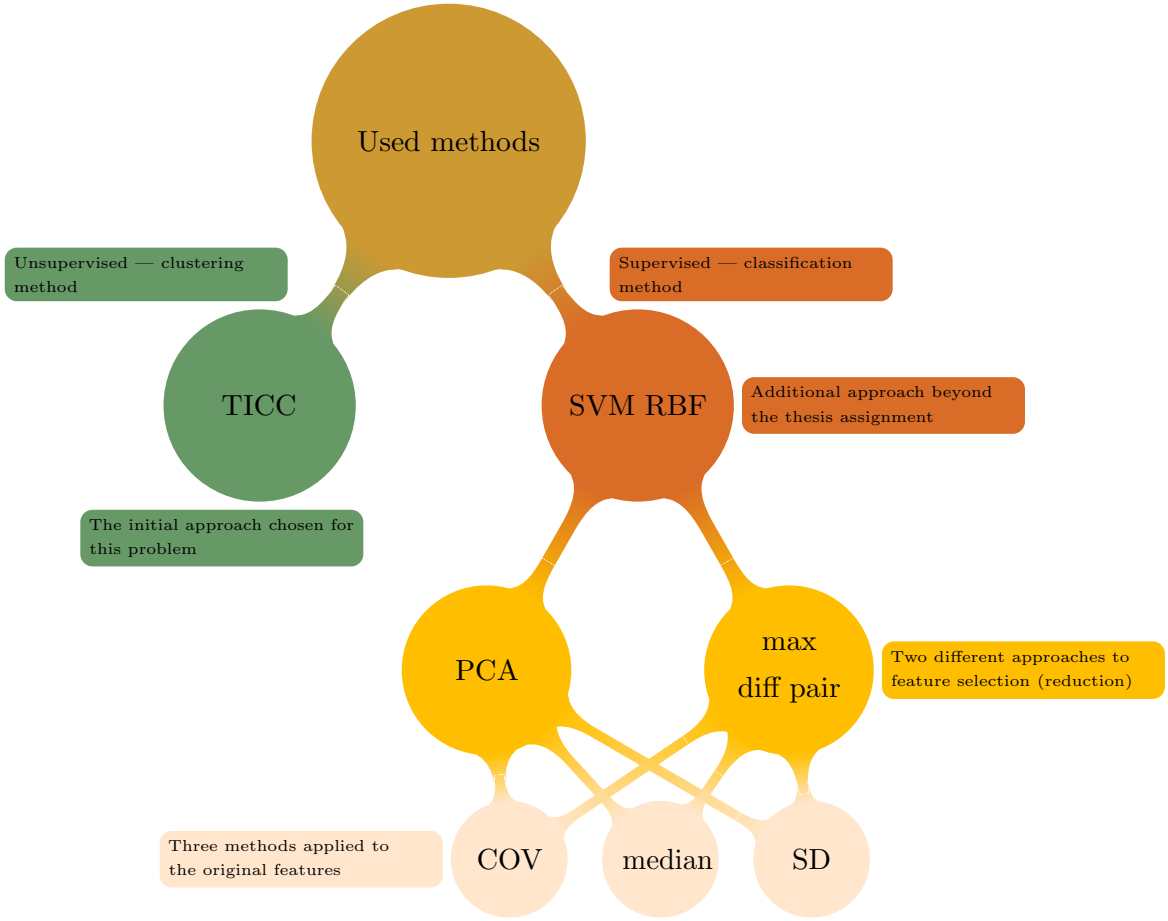


Figure 3.1: Overview of used methods

The original idea was to use the existing unsupervised technique and novel method based on the multivariate series of daily actigraphy features: Toeplitz Inverse Covariance-Based Clustering (TICC) algorithm to identify clinical episodes in BD patients. This method, introduced in section 3.1, was evaluated on simulated data for testing and deeper investigation. Although this unsupervised method appeared to be very promising, it did not suit for our real data. Since this unsupervised approach was applied to a supervised learning problem, we decided to add a supervised classification using Radial Basis Function (RBF) kernel SVM classifier, detailed in section 3.2.

For the case of supervised approach, we again used either covariance matrices as in TICC as input to the classifier, or feature values of the actigraphy data — median and SD. For simplicity, we tried to identify only remission-mania or remission-depression separately (data annotation and data preprocessing are described in section 4).

For supervised approaches, a feature selection was necessary. This was achieved by two methods, described in subsection 3.2.1 — using the Principal Component Analysis (PCA) and our Maximal-Difference Method (max-diff) method based on the maximal difference of feature pairs in the COV (or median or SD) between the remission and corresponding episode (mania or depression).

3.1 Unsupervised Approach

This section is dedicated to the unsupervised method: TICC — described in section 3.1.1. This algorithm appeared to be a good option to apply to this multivariate time-series data. However, it was necessary to determine if it was suitable for this data due to sampling and dimensionality (hence we performed experiments on simulated data). And secondly to come up with a reasonable cluster assignment approach (see section 3.1.1.2) to convert the clustering results to labels, since we apply this unsupervised method to a supervised learning problem. In the results section 5.1.1, a comparison of how different cluster evaluation methods behave for different cases of results is available before evaluating the unsupervised method itself.

3.1.1 Toeplitz Inverse Covariance-Based Clustering

The Toeplitz inverse covariance-based clustering approach proposed by Hallac et al. [1] is a promising method for clustering multivariate time series data. Their method involves estimating the inverse covariance matrix for each time series, and then calculating a Toeplitz version of this matrix that captures the temporal dependencies between the variables. This Toeplitz inverse covariance matrix is used as a distance metric to cluster the time series data using a spectral clustering algorithm. The authors validated their approach by comparing TICC to several state-of-the-art baselines in a series of synthetic experiments and then demonstrated on a dataset of automotive sensors how TICC algorithm can be used in real-world scenarios. This approach has the potential to advance the field of time series clustering by enabling the identification of more complex and nuanced patterns in multivariate time series data.

TICC approach divides each time serie into several clusters, each characterised by a correlation network or Markov Random Field (MRF) defined in a short time window of size w. This MRF manages the time-invariant partial correlation structure of any window on the side of a segment belonging to this cluster. TICC learns both the cluster MRF and the time series segmentation using a correlation network. [1]

TICC algorithm could be also applied to the features of actigraphic data for the recognition of clinical episodes in bipolar disorder. By identifying clusters of individuals with similar patterns of rest-activity, it may be possible to identify individuals who are at increased risk for clinical episodes or who may respond differently to treatment.This could potentially lead to more personalised and effective treatments for bipolar disorder.

TICC involves several steps, including the estimation of inverse covariance matrices, the transformation of these matrices into Toeplitz matrices, and the use of spectral clustering to group time series data based on their distance in the Toeplitz inverse covariance-based matrix.

A straightforward visualisation of this method is given below, where n is the number of sensors (number of parallel time series of length T), A, B and C are the clusters that TICC segments into a sequence of states. Each cluster is defined by a correlation network or MRF in a short time window of length w.

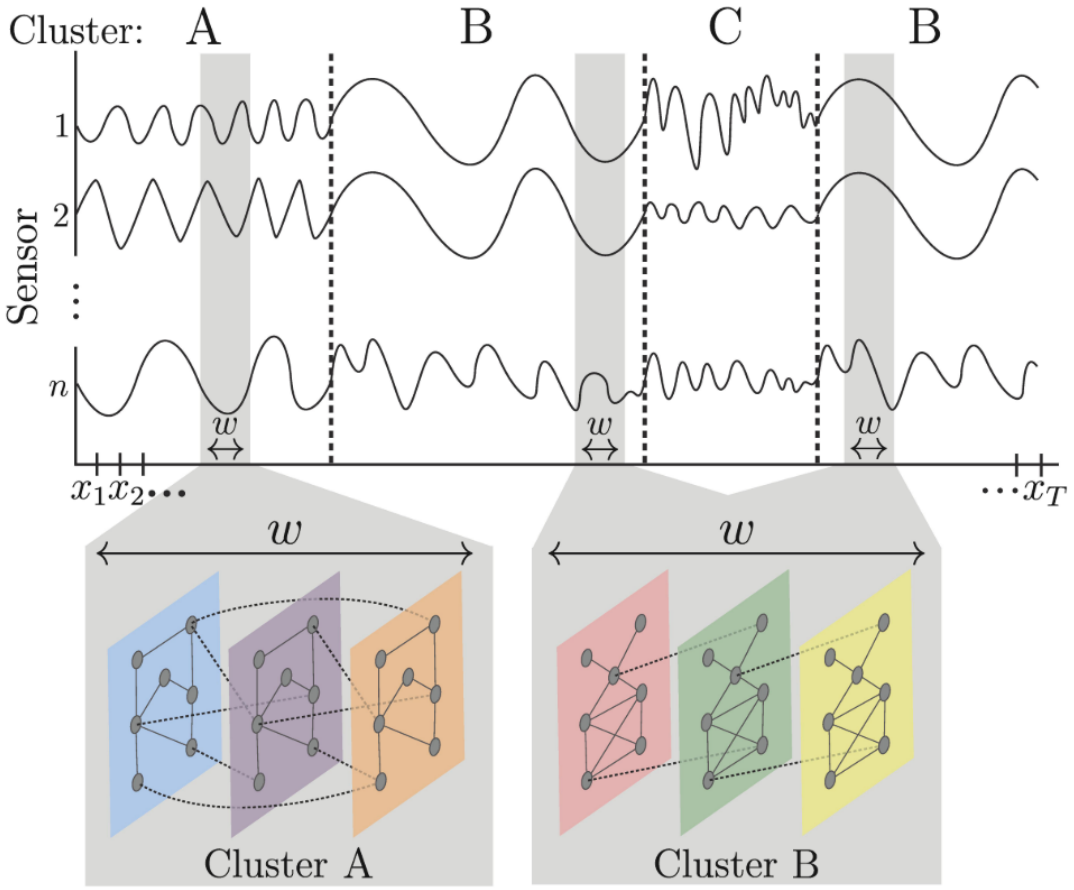


Figure 3.2: Visualisation of Toeplitz Inverse Covariance-based Clustering, reprinted from [1]

The algorithm for TICC can be summarised as follows:

- **Input:** Multivariate time series data consisting of n time series with T observations each.
- Estimates the inverse covariance matrix for each subset of data (defined by the time window w) $X_t = x_{t-w+1}, \ldots, x_t$ using the graphical lasso algorithm.
- Constructs a Toeplitz inverse covariance matrix for each subset of data (defined by the time window w) by replicating the inverse covariance matrix along the diagonal and sub-diagonals. This captures the temporal dependencies between variables.
- Computes the pairwise distances between the Toeplitz inverse covariance matrices for all pairs of time series using the Frobenius norm.
- Defines each cluster using a multi-layer MRF.
- Updates the MRF clusters by solving the Toeplitz graphical lasso algorithm and finds the optimal solution.
- **Output:** Cluster labels for each time series.

The computational complexity of the algorithm is dominated by the inverse covariance estimation step, which has a complexity of $\mathcal{O}(nT^3)$.

The TICC problem can be expressed by the following equation:

$$\underset{\Theta \in \mathcal{T}, \mathrm{P}}{\operatorname{argmin}} \sum_{i=1}^{K} [\overbrace{\left\|\lambda \circ \Theta_i\right\|_1}^{\text{sparsity}} + \sum_{X_t \in P_i} (\overbrace{-\ell\ell\left(X_t, \Theta_i\right)}^{\text{log likelihood}} + \overbrace{\beta \mathbb{1}\left\{X_{t-1} \notin P_i\right\}}^{\text{temporal consistency}})], \tag{3.1}$$

$$\ell\ell\left(X_t, \Theta_i\right) = -\frac{1}{2}\left(X_t - \mu_i\right)^T \Theta_i \left(X_t - \mu_i\right) + \frac{1}{2} \log\det \Theta_i - \frac{n}{2} \log(2\pi),$$

where T is the set of symmetric block Toeplitz matrices $nw \times nw$ and $\left\|\lambda \circ \Theta_i\right\|_1$ is the ℓ_1-norm penalisation of the element-wise (Hadamard) product to instigate a sparse inverse covariance (where $\lambda \in \mathbf{R}^{nwnw}$ is a regularisation parameter). Moreover, $\ell\ell(X_t, \Theta_i)$ is the log-likelihood that X_t comes from cluster i, where $_i$ is the empirical mean of cluster i. β is the parameter that enforces time consistency, and $\mathbb{1}\left\{X_{t-1} \notin P_i\right\}$ is the indicator function that checks whether neighbouring points are assigned to the same cluster. [1]

Since TICC is essentially an Expectation-Maximisation algorithm, we can also express it as follows [1]:

Algorithm 1: Toeplitz Inverse Covariance-Based Clustering

initialize *Cluster parameters* Θ*; cluster assignments* P*.*
repeat *E-step: Assign points to clusters* $\rightarrow P$*,*
M-step: Update cluster parameters $\rightarrow \Theta$*.*
until *Stationarity*
return *(*Θ*,* P*)*

3.1.1.1 Regularisation Parameters Setting

To regularise the TICC behaviour the following adjustable parameters are used:

- λ — sparsity,
- β — temporal consistency,
- K — number of expected clusters,
- w — window length,

To set the most appropriate parameters, we used grid-search for the parameters λ and β to find the best combination as follows:

$$\lambda = [0.01,\ 0.1], \qquad \beta = [10,\ 100,\ 500]. \tag{3.2}$$

The parameters K and w were set manually according to known input data.

3.1.1.2 Cluster Assignment Evaluation

Cluster evaluation measures can be divided into two types: external measures and internal measures. External measures compare the expected or known ground truth result with the actual result, while internal measures evaluate the quality of the result based on other characteristics such as the compactness or separation of the clusters. The key difference between the two types is that external measures rely on a qualitative comparison, while internal measures do not have access to the expected result.

The expected or known ground truth results can be class labels which are often created (semi)manually by experts and can be regarded as a gold standard (classes considered to be 'correct') [52]. The clustering approaches are not always able to find the correct number of clusters as in gold standard. Algorithms can both overestimate or underestimate the number of clusters. In such scenario, the standard criteria used to classify results cannot be used.

One of the most popular information theory measures for evaluating clustering results is *Normalized Mutual Information (NMI)* [53]. *NMI* is *Mutual Information (MI)* normalised to scale the results between 0 and 1 (1 indicating complete agreement between the sets).

Most common evaluation criteria according to some previous data binning studies [54]–[56] that could be also used in our case are *precision* (also known as *Positive Predictive Value (PPV)*), *recall* (also known as *sensitivity*), *F1-score* and *Adjusted Rand Index (ARI)*.

The authors of the TICC algorithm [1] evaluated the clustering accuracy by measuring the *macro-F1 score*, which is basically the arithmetic mean of the *F1-scores* for all the clusters as mentioned above.

To evaluate the TICC algortihm, we applied the four common criteria: *Precision*, *Recall*, *F1-score* and *ARI*. Their detailed description is given below.

Assume there are N classes present in the manual evaluation by expert and M clusters predicted by the TICC algorithm. Let R_{ij} be the total number of (time series) samples clustered to the i^{th} cluster and belongs to the j^{th} class in our true labels. This can be represented using contingency table as follows:

	Classes j				
Clusters i	1	2	3	...	N
1	R_{11}	R_{12}	R_{13}	...	R_{1N}
2	R_{21}	R_{22}	R_{23}	...	R_{2N}
3	R_{31}	R_{32}	R_{33}	...	R_{3N}
⋮	⋮	⋮	⋮	⋱	⋮
M	R_{M1}	R_{M2}	R_{M3}	...	R_{MN}

Table 3.1: Contingency table of the clustering results

Precision

Precision (also known as *PPV*) give us information about how many of all positive predictions are really positive. That can be describe using following formula:

$$Precision = \frac{TP}{TP + FP}, \tag{3.3}$$

where TP represents true positives and FP false positives.

In the clustering task, each cluster is assigned a true label class with the maximum number of corresponding samples (represented by the R_{ij} matrix). Then the precision can be expressed as the sum of the maximum number of samples for each cluster divided by the total number of clustered samples.

$$Precision = \frac{\sum_{i=1}^{M} \max_{j} R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij}}. \tag{3.4}$$

Recall

Recall (also known as *sensitivity*) represents how many of all real positive cases are predicted positively. That can be again describe using following formula:

$$Recall = \frac{TP}{TP + FN}, \tag{3.5}$$

where TP represents true positives and FN false negatives.

In the clustering task, every true labels class is assigned an obtained cluster with the maximum number of corresponding samples (represented by the R_{ij} matrix). Then the precision can be expressed as the sum of the maximum number of samples for each class divided by the total number of clustered and unclustered [1] samples.

$$Recall = \frac{\sum_{j=1}^{N} \max_{i} R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij} + number\,of\,uncluslustered\,samples}. \tag{3.6}$$

F1-score

F1-score is the harmonic mean of *precision* and *recall* and is defined as:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3.7}$$

macro-F1-score

macro-F1-score is the arithmetic mean of all per-class *F1-scores* defined as:

$$macro\text{-}F1\text{-}score = \frac{F1\text{-}score}{number\,of\,classes}. \tag{3.8}$$

ARI

The *Rand Index (RI)* measures how similar the clustering results are to the reference classes. The formula of RI is:

$$RI = \frac{number\,of\,agreeing\,clusters}{number\,of\,pairs} \tag{3.9}$$

and range from 0 to 1, where 1 represents perfect match.

In mathematical terms, RI is related to accuracy, but it can also be applied when class labels are not used. The RI would be non-zero just by chance, assuming random clustering. Therefore the RI can be "adjusted for chance" into the ARI using the following formula:

$$ARI = \frac{RI - RI_{exp}}{max(RI) - RI_{exp}}, \tag{3.10}$$

[1] In our case, when using the TICC algorithm, we will always have the number of unclustered samples equal to zero.

where RI_{exp} is the expected RI.

By modifying this, we can again express this formula using the R_{ij} matrix as follows:

$$ARI = \frac{\sum_{i,j}\binom{a_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{3.11}$$

$$t_1 = \sum_i \binom{\sum_j a_{ij}}{2}, \qquad t_2 = \sum_j \binom{\sum_i a_{ij}}{2}, \qquad t_3 = \frac{2t_1 t_2}{\binom{\sum_{i=1}^{M}\sum_{j=1}^{N} R_{ij}}{2}}$$

NMI

Let Y be the set of ground truth labels and C the set of computed cluster labels. Then NMI between these two sets is:

$$NMI(Y,C) = \frac{2 \times MI(Y,C)}{H(Y) + H(C)} = \frac{2 \times (H(Y) + H(C) - H(Y,C))}{H(Y) + H(C)},$$

where $H(Y)$ and $H(C)$ is the individual entropy, $H(Y,C)$ is the joint entropy and $MI(Y,C)$ is the mutual information.

This can be expressed using probabilities as follows:

$$NMI(Y,C) = \frac{2 \times \sum_{y\in Y}\sum_{c\in C} P_{Y,C}(y,c)\log_2\frac{P_{Y,C}(y,c)}{P_Y(y)P_C(c)}}{\sum_{y\in Y} P_Y(y)\log_2 P_Y(y) + \sum_{c\in C} P_C(c)\log_2 P_C(c)},$$

where $P_{Y,C}$ is the joint probability mass function and P_Y and P_C are the marginal probability mass functions.

3.2 Supervised Approach

As an additional approach beyond the thesis assignment, we chose SVM [57] with RBF kernel — one of the most generalised form of kernalisation, primarily because of its simplicity and efficiency to deal with non-linear relationships in data. SVM is essentially a linear supervised learning algorithm, that can be used for classification and regression tasks. As input to this classifier we used 3 approaches

- COV — as a supervised parallel approach of the TICC method
- median
- SD

Each of these methods was applied to the input 90 actigraphic features over time, depending on the one of three validation scenarios, either to all data belonging to a given state (mania, depression or remission) method or within a 7-day overlapping window for each day as follows:

- Leave-One-Patient-Out Cross-Validation Leave-One-Person-Out Cross-Validation (LOOCV)
  - Each method was calculated over all days for a given clinical state for all patients except one and then tested on this left-out patient subsequently.
- 7-day Window Across All Patients
  - In this case, the methods were computed in a 7-day window of training data and then tested on the testing set again in a 7-day window for different dataset types across all patients[2].
- 7-day Window for Single Patient
  - In this case, the methods were computed in a 7-day window of training data and then tested on the testing set again in a 7-day window for different types of datasets, each time for a single patient[2].

3.2.1 Feature selection

Since we had $N = 90$ actigraphic features for each day of preprocessed data (see section 4.2.3) the resulting COV matrix for each day respectively the upper triangular matrix without diagonal contained $N(N-1)/2 = 4005$ values. For median and std methods 90 for each day. Therefore, it was appropriate to reduce this number of features. Each time we used feature extraction using PCA on the COV (or one of the feature value methods) of the original features, defined in section 3.2.1.1. Eigenvectors were obtained from COV (or one of the methods) computed for each clinical episode for each patient on training dataset (i.e., two matrices for each patient)

[2] All used types of dataset splitting are described in section 4.2.4.

and then applied to correlation matrices computed from a 7-day overlapping window for each day. In the case of the LOOCV method, COV, median, and SD were calculated on all days belonging to a given state (mania, depression or remission) for the remaining patients. As a second approach to reduce the number of features we chose feature selection using the approach based on maximal difference of feature COV, median, or SD values between the two clinical episodes (remission-mania or remission-depression) to preserve only the most distinctive features, described in section 3.2.1.2.

3.2.1.1 PCA Method

Principal component analysis (PCA) is a data analysis technique for a dimensionality reduction that aims to find a linear projection of data points onto a lower dimensional subspace while minimizing information loss. To find these new uncorrelated orthogonal variables, which are called principal components, the eigenvalue/eigenvector problem is solved and the new variables are defined based on the input data set. [58]–[60]

Consider a set $\mathbf{X} \in \mathbb{R}^D$ and a observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The goal of PCA is to find a projection onto the space of dimension $M < D$ that maximizes the variance of the projected data. For simplicity, let us seek a projection onto a one-dimensional subspace $M = 1$. We define by an unit vector $\mathbf{u}_1 \in \mathbb{R}^D$. Then each data point $\mathbf{x}_n$ is projected onto the this new space (scalar value) $\mathbf{u}_1^T \mathbf{x}_n$. With the assumption of centred data $x_n = x_n - \overline{x}$, the variance of the projected data can be expressed as

$$\frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{x}_n)(\mathbf{x}_n^T\mathbf{u}_1) = \mathbf{u}_1^T\mathbf{\Sigma}\mathbf{u}_1, \tag{3.12}$$

where $\mathbf{\Sigma}$ is the covariance matrix of the observed data in the original high dimensional space

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T. \tag{3.13}$$

Since $\mathbf{u}_1$ is a unit vector $\mathbf{u}_1^T\mathbf{u}_1 = 1$, we can maximise the variance using Lagrange multiplier λ_1 as follows

$$\mathcal{L}(\mathbf{u}_1, \lambda_1) = \mathbf{u}_1^T\mathbf{\Sigma}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1)) \tag{3.14}$$

Setting the derivative by $\mathbf{u}_1$ of equation 3.14 to 0

$$\frac{\partial\mathcal{L}(\mathbf{u}_1)}{\partial\mathbf{u}_1} = 2\mathbf{\Sigma}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1 = 0, \tag{3.15}$$

we get stationary point

$$\mathbf{\Sigma}\mathbf{u}_1 = \lambda_1\mathbf{u}_1. \tag{3.16}$$

It follows that at the stationary point, $\mathbf{u}_1$ must be the eigenvector of $\mathbf{\Sigma}$ and λ_1 the corre-

sponding eigenvalue. If we left-multiply the equation with the $\mathbf{u}_1^T$ we see that the maximal variance is equal to the eigenvalue λ_1 .

$$\lambda_1 = \mathbf{u}_1^T \mathbf{\Sigma} \mathbf{u}_1 \tag{3.17}$$

Thus, the variance will be maximal if the vector $\mathbf{u}_1$ is equal to the eigenvector corresponding to the largest eigenvalue - first principal component. Then for a general M-dimensional space, it can be proven that the optimal projection preserving the largest amount of variance of the projected data is defined by M eigenvectors$\{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$ of the covariance matrix $\mathbf{\Sigma}$ belonging to the corresponding M largest eigenvalues $\{\lambda_1, \ldots, \lambda_M\}$. [58], [59]

In this thesis,PCA method was used to feature extraction from the original high-dimensional space. In each application of this method, only the first 15 components (out of 4005 for the COV method and 90 for the median and SD method) were used, or fewer if their explained variance of the original data was greater than or equal to 99%.

3.2.1.2 Maximal Difference Method

In this section we describe the feature selection using the max-diff between the clinical states. On the training data, we calculated the correlation matrix of all actigraphic features, separately for remission $\mathbf{r}_{\text{rem}}$ and for mania $\mathbf{r}_{\text{episode}}$ (respectively depression), and we sorted all pairs of features in descending order according to the absolute difference between these episodes (diff pairs). However, within these pairs there were highly correlated features that were paired with the same feature. To mitigate this issue, highly correlated feature pairs with a correlation coefficient greater than 0.6 (r_{rem} or $r_{episode} > 0.6$), which had the same feature within the diff pair, were firstly removed. Respectively, only the pair with the highest $\mathbf{r}_{\text{diff}}$ value was retained in the difference correlation matrix. Subsequently, this selected pair was used to create a mask, which was represented as a matrix with dimensions of n x n, where n corresponds to the number of original actigraphic elements. This mask facilitated a straightforward selection of feature pairs from the correlation matrices within the designated time window. The algorithm presented below (Algorithm 2) provides a comprehensive representation of the step-by-step procedure for this feature selection.

Algorithm 2: Feature selection based on maximal difference between clinical states

```
input Correlation matrices r_rem, r_episode upper triangular matrix without diagonal
select r_rem, r_episode >= 0.6
r_high = join (elements of r_rem, r_episode)
sort elements of r_high in descending order
get names of high correlated feature pairs corresponding to sorted elements of r_high:
        HighCorrFeaturePairs[:, 1],
        HighCorrFeaturePairs[:, 2]
sort elements of r_diff = |r_rem - r_episode| in descending order
get names of sorted feature pairs corresponding to sorted elements of r_diff:
        MaxDiffFeaturePairs[:, 1],
        MaxDiffFeaturePairs[:, 2]
for i = 1, ..., number of HighCorrFeaturePairs do
    if both feature names HighCorrFeaturePairs in MaxDiffFeaturePairs then
        if pairs in MaxDiffFeaturePairs are same features then
            keep only the pair with higher diff value in MaxDiffFeaturePairs
            update MaxDiffFeaturePairs
        end
    end
end
SelectedFeaturePairs = MaxDiffFeaturePairs
return SelectedFeaturePairs.
```

For the case of the median and SD method, this was a simpler procedure. Since the computed values were not matrices describing the relationships between pairs of features, only 90 values for each day, it was only necessary to find the maximal difference between the two states without any removal of dependent features.

3.2.2 Support Vector Machine

Support Vector Machines (SVMs) are a powerful class of supervised machine learning algorithms used for classification and regression tasks. SVM models, developed by Corinna Cortes and Vladimir Vapnik [57] in the 1990s, excel at handling complex datasets by determining optimal hyper-planes that separate different classes. These hyper-planes, defined by support vectors, maximise the spread between classes, making SVMs particularly effective in high-dimensional spaces. [59], [60]

Let us assume a linear classification model for simplicity. Consider a training set $\mathbf{X} \in \mathbb{R}^D$ and a observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and their corresponding classes (labels) $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $y \in \{-1, 1\}$ then the discrimination function for the binary classification task can be written as follows

$$\mathbf{g}(\mathbf{x}_n) = \mathbf{w}^T \mathbf{x}_n + b. \tag{3.18}$$

In general, $\mathbf{g}(\mathbf{x}_n)$ is a hyper-plane in N-dimensional feature space.

The assigned class $\hat{\mathbf{g}}(\mathbf{x}_n)$ for the points from the recognition area $\mathbf{X}$ is defined as

$$\hat{\mathbf{g}}(\mathbf{x}_n) = sign(\mathbf{g}(\mathbf{x}_n)). \tag{3.19}$$

SVM searches for an optimal separating hyper-plane satisfying the criterion:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0 \tag{3.20}$$

Then for the case of a linearly separable training dataset, there is at least one combination of parameters $\mathbf{w}_n$ and b satisfying this condition. If there are multiple pairs of parameters $\mathbf{w}_n$ and b satisfying this condition, SVM finds the optimal one by maximising the minimum distance separating the hyper-planes from all elements of the training set — the so-called margin. This is because the larger the margin, the better the generalisation to the test. [59], [60]

The orthogonal distance of any point $\mathbf{x}_n$ from the decision hyper-plane is given by the following expression $\frac{y_n(\mathbf{w}^T\mathbf{x}_n+b)}{\|\mathbf{w}_n\|}$. Margin is then defined as the orthogonal distance of the decision hyper-plane to the nearest point $\mathbf{x}_n$ from the dataset. Since parameter scaling $\mathbf{w} \to \kappa\mathbf{w}$ and $b \to \kappa b$ does not affect the distance of a point from the decision hyper-plane, we can set For the nearest point $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$, thus each point $\mathbf{x}_n$ meets the condition

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1. \tag{3.21}$$

Then the task of finding the optimal pair of parameters $\mathbf{w}_n$ and b can be formulated as

$$(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w}^*, b^*} \frac{1}{2}\|\mathbf{w}\|^2, \text{ subject to } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1. \tag{3.22}$$

This problem can be further formulated as a dual problem using Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)_T$ as follows

$$\mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} \alpha_n \left\{ y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \right\}, \text{ where } \alpha_n \geq 0, n = 1, \ldots, N. \tag{3.23}$$

By solving this problem, all non-zero α_n will belong to the vectors $\mathbf{x}_n$ that are closest to the separating hyper-plane. These vectors will be reflected in the learning result of the classifier and are called, as already the name of the classifier suggests, support vectors. [59], [60]

If we set the derivatives of 3.23 with respect to $\mathbf{w}_n$ and b to be zero, we obtain the following two conditions

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \text{ and } 0 = \sum_{n=1}^{N} \alpha_n y_n. \tag{3.24}$$

Substitution of these conditions 3.24 into equation 3.23 to eliminate the variables $\mathbf{w}_n$ and b leads to a dual formulation of the problem

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}^*} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x_n}^{\mathrm{T}} \mathbf{x_m}, \tag{3.25}$$

where

$$\alpha_n \geq 0, \;\; n = 1, \ldots, N \;\text{ and }\; \sum_{n=1}^{N} \alpha_n y_n. \tag{3.26}$$

This constrained optimization of this form satisfies the Karush-Kuhn-Tucker condition, which in this case requires the following three properties to hold

$$\begin{aligned} \alpha_n &\geq 0 \\ y_n y\left(\mathbf{x}_n\right) - 1 &\geq 0 \\ \alpha_n \left\{ y_n y\left(\mathbf{x}_n\right) = 1 \right\} &= 0 \end{aligned} \tag{3.27}$$

Thus, for each data point, either $\alpha_n = 0$ or tny(xn)=1. Any data point for which an =0 does not appear in the summation in (7.13), and therefore plays no role in making predictions for new data points. The remaining data points are called support vectors, and since they satisfy the condition tny(xn)=1, they correspond to points that lie on the hyper-planes of maximal margin in the feature space. This property is crucial for the practical applicability of support vector machines. [59], [60]

If the data are not linearly separable, condition will not be satisfied and the classification of the training data will lead to poor generalization on the test set. The condition can be mitigated by introducing an additional variable $\xi_n \geq 0$ (called slack variable) for each data point, where $\xi_n = 0$ for correctly classified points. These allow relaxation, i.e. the possibility of breaking some of the inequalities. Then condition can be rewritten in the following form

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \quad n = 1, \ldots N. \tag{3.28}$$

The points with non-zero weak variable, are located either in the region defined by the margin ($0 < \xi \leq 1$) or on the wrong side of the decision hyper-plane ($\xi \leq 1$). The addition of slack variable leads to a new formulation of the optimization problem

$$\mathbf{w}^*, b^*, \xi_n^* = \arg\min_{\mathbf{w}, b, \xi_n} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \xi_n \tag{3.29}$$

$$t_n \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n + b \right) \geq 1 - \xi_n, \quad \xi_n \geq 0, \qquad n = 1, \ldots, N. \tag{3.30}$$

The regularization parameter C represents the trade-off between margin size and penalty for misclassified points - or the trade-off between the error on the training set and the complexity of the model. [59], [60]

For a classification problem whose solution is a nonlinear decision boundary, the kernel trick can be used. The main idea is to map the data into a higher dimensional space (dimension lifting)

using nonlinear transformations in which the linear unseparable data are separable. The kernel trick does not require knowledge of the exact mapping $\phi(\mathbf{x})$ of points, only the corresponding kernel function $\kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}} \phi(\mathbf{x}')$. The kernel function can be any positive semidefinite function. As a core functions are most commonly used:

- Linear function (identity): $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i}^T \boldsymbol{x_j}$.
- Polynomial function: $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(1 + \boldsymbol{x_i}^T \boldsymbol{x}_j\right)^k$, where k is the order of the polynom.
- Radial basis function (Gaussian): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$, where σ denotes the standard deviation in the Gaussian distribution.
- Two-layer perceptron (neural network) $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh\left(a \boldsymbol{x}_i^T \boldsymbol{x}_j + b\right)$, where a becomes positive and b negative.

3.2.3 Hyperparameter Setting and Cross-Validation

We used grid search to set the classifier hyperparameters. Similar to the approach taken by Chapelle and Zien [61], we applied a heuristic approach to select the appropriate search range for standard deviation in the RBF kernel σ and regularization parameter C in SVM.

C-parameter controls the smoothness of decision boundary as follows:

- $C \to 0$ — large margin (smooth decision boundary),
- $C \to \infty$ — narrow margin (convoluted decision boundary).

σ-parameter represents inverse of the radius of influence of samples selected by the model as support vectors:

- $\sigma \to 0$ — decision boundary tends to be too flexible (hazard of overfitting),
- $\sigma \to \infty$ — the decision boundary tends to be limited and fails to capture the complexity or shape of the data (it is affected by the entire training set and behaves similarly to a linear model and is prone to misclassification in prediction, but avoids the danger of overfitting).

First, we performed a 5-fold cross-validation on the training set with a coarse grid search:

$$C = 10^{(-2:1:4)}, \quad \sigma = 10^{(-2:1:4)} \tag{3.31}$$

We selected the best pair of these parameters C^* and σ^* based on the minimum cross-validated classification error over all 5 folds and repeated this step with a finer grid:

$$C = C^* \cdot 2^{(-3:1:3)}, \quad \sigma = \sigma^* \cdot 2^{(-3:1:3)} \tag{3.32}$$

Again, we selected the best pair of hyperparameters by following the procedure in the preceding step. These selected parameters were then used to re-train the model on the whole

training set. Subsequent evaluation involved assessing its performance on an independent testing set.

The hypermeter settings was tested in three validation scenarios

- Leave-One-Patient-Out Cross-Validation LOOCV
  - In this case, a 5-fold validation with double grid search was performed within each patient iteration. Afterwards, the most optimal parameters were used to re-train this model on the entire training dataset — i.e., on all but one patient. The behaviour of the new model was then evaluated on this left-out patient.
- 7-day Window Across All Patients,
  - Again, a 5-fold validation with double grid search was always performed on the corresponding training dataset, followed by retraining the model with the best parameters on the whole training set (see section 4.2.4). With the difference that in this case it was a classification of individual days across all patients not just clinical states.
- 7-day Window for Single Patient.
  - This validation scenario is analogous to the previous one, only with classification of each patient separately not across all.

3.2.4 Evaluation Metrics for Classification Model

Since the exact label assignment is known for the classification task, it is not necessary to use more complex metrics for evaluation as in the case of clustering. Therefore, we chose the classical metrics sensitivity (SEN), specificity (SPEC) and accuracy (ACC). There are several key terms that are commonly used to describe these metrics:

Classification outcome

- TP ..... real positive and classified as positive (True Positive),
- TN ..... real negative and classified as negative (True Negative),
- FP ..... real negative and classified as positive (False Positive),
- FN ..... real positive and classified as negative (False Negative).

Reality

- P ..... number of real positive cases,
- N ..... number of real negative cases.

SEN or TPR (True Positive Rate) is a conditional probability

$$SEN = P(TP|P) = \frac{TP}{TP + FN} = \frac{number\ of\ true\ positive\ assessments}{number\ of\ all\ positive\ assessments},$$

indicating the relative frequency of correctly classified positive cases.

SPEC or TNR (True Negative Rate) is a conditional probability

$$SPEC = P(TN|N) = \frac{TN}{TN + FP} = \frac{number\ of\ true\ negative\ assessments}{number\ of\ all\ negative\ assessments},$$

showing the relative frequency of correctly classified negative cases.

ACC is the conditional probability

$$SPEC = P(TN|N) = \frac{TP + TN}{TP + TN + FP + FN} = \frac{number\ of\ correct\ assessments}{number\ of\ all\ assessments},$$

that the classifier correctly evaluates the case.

Chapter 4

Datasets

This chapter describes all the data processed in this thesis. The first section 4.1 describes the simulated data used to evaluate and to further explore the TICC method. The second section is devoted to presenting real longitudinally acquired actigraphy data from BD patients, their annotation and prepocessiong datasets.

4.1 Simulated data

In order to evaluate the effect of the set parameter values on the identification accuracy using TICC algorithms, we generated multivariate signals with the require covariance structures. To do so, we used the Matlab function R = mvnrnd(mu,Sigma,n) which returns a matrix **R** of n random vectors selected from the same multivariate normal distribution with mean vector **mu** and covariance matrix **Sigma** [62].

```
% Parameters
s1 = 1; s2 = 5; % change values of high covariance
N = 5;  % Number of simulated features
block_sizes = [10, 10, 10]; % Sizes of the low, high, and low
   covariance clusters
num_samples = sum(block_sizes);  % Total number of samples

% Create covariance matrix
mu1 = 1*rand(1,N);        % mean
Sigma1 = s1*rand(N);      % sigma
Sigma1=(Sigma1'*Sigma1); % positive semidefinite matrix

mu2 = 1*rand(1,N);        % mean
Sigma2 = s2*rand(N);      % sigma
Sigma2=(Sigma2'*Sigma2); % positive semidefinite matrix
```

```
data = zeros(num_samples, N);

% Generate signals
data = [mvnrnd(mu1, Sigma1, block_sizes(1));
        mvnrnd(mu2, Sigma2, block_sizes(2));
        mvnrnd(mu1, Sigma1, block_sizes(3))];

% Compute covariances of generated signals
bs = [0 block_sizes];
for i = 1:numel(block_sizes)
    COV{i} = cov(data(1+sum(bs(1:i)):sum(bs(1:i+1)),:));
end
```

Since the mean value is not important for this task, we left it unchanged and modified only the covariance values. Below are visualisations of the required covariances and covariances calculated from the resulting signal.

Below are visualisations of the desired and calculated covariances from the resulting signal. The first and third clusters are generated from the same desired covariance matrix and the second cluster is generated from significantly higher covariance values. For comparison, a short (10 samples for each class) and a long (10 000 samples for each class) signal are shown. In both cases, the posterior covariances were the same, with an average absolute difference between the covariance for the first (or third) and second class being 33.

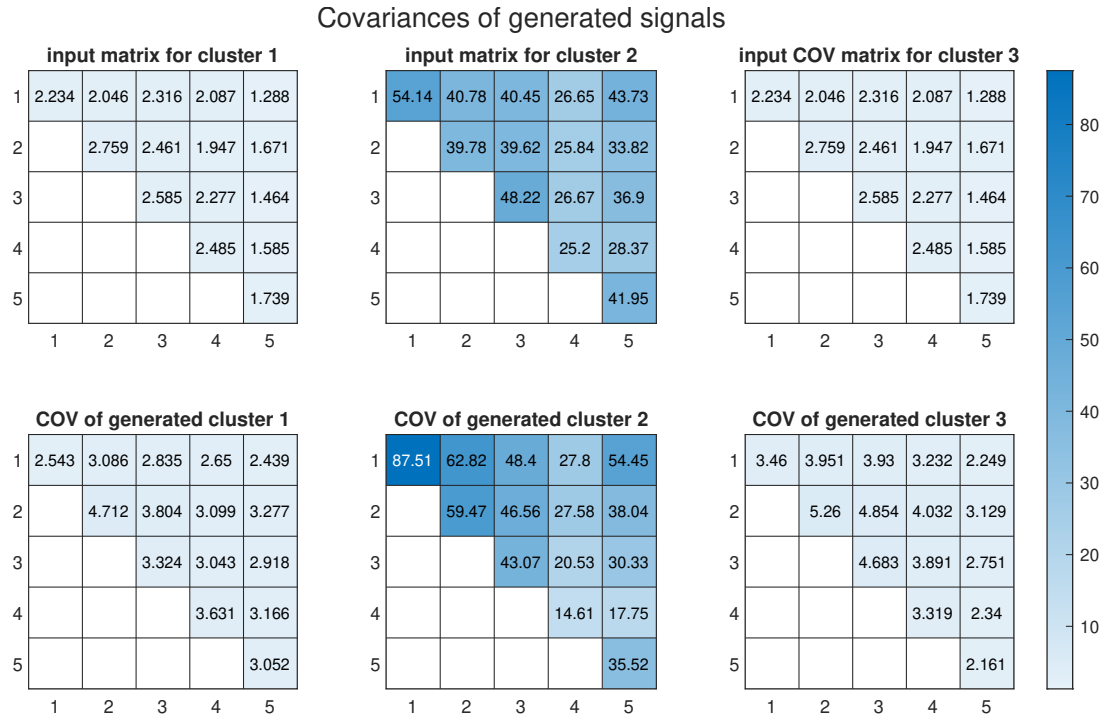


Figure 4.1: Required and obtained covariances of the generated short signal

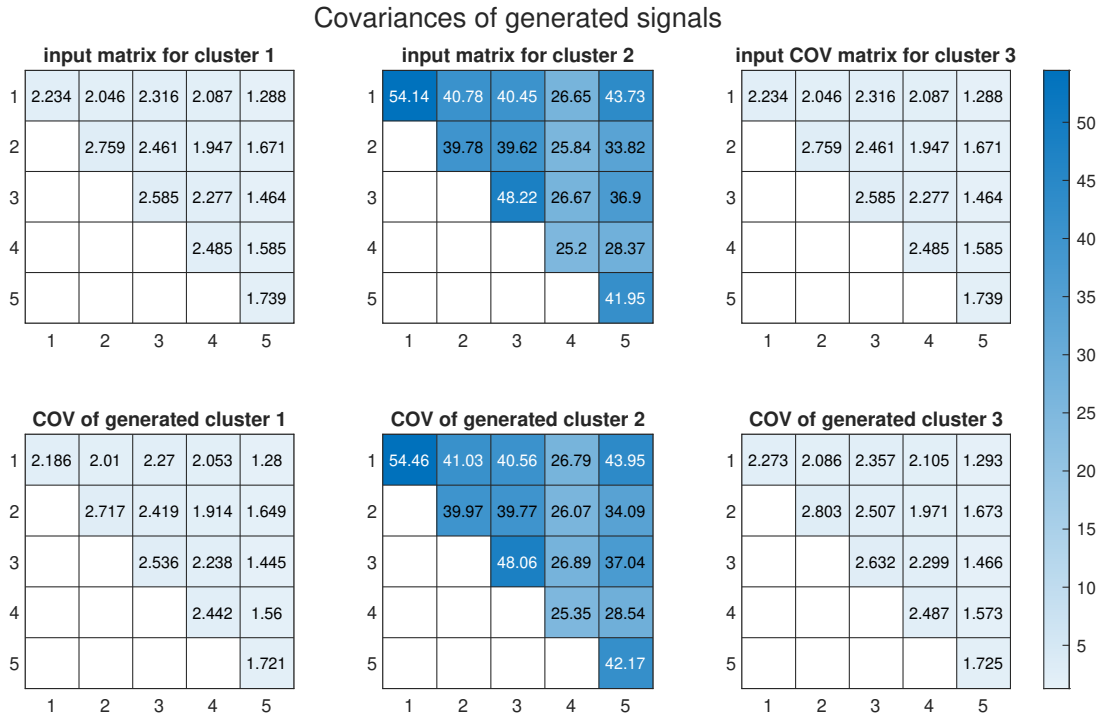


Figure 4.2: Required and obtained covariances of the generated long signal

We can note the difference in the similarity of the required and obtained covariances caused by the Central Limit Theorem . Therefore, it is appropriate to take into account the lower reliability and accuracy of the shorter generated signals.

4.2 Real data

As part of this thesis, I have adopted and incorporated an extensive data set developed by Ing. Jakub Schneider Ph.D. as part of his doctoral thesis [42]. The dataset (introduced in section 4.2.1) consists of a 90 calculated features for each day from the long-term actigraphy monitoring of BD patients within the AKTIBIPO400 clinical study (organized by Mindpax in cooperation with National Institute of Mental Health (NIMH) in Klecany), including 369 BD patients, out of which 115 also underwent MADRS and YMRS clinical scales once a month. The follow-up period was 18 months with a possibility of extension. They were also asked to complete the weekly Aktibipo Self-rating EMA (ASERT) questionnaire (described in section 4.2.2) through the provided Mindpax mobile application.

4.2.1 Actigraphic feature dataset

The following list is a brief description of all the actigraphic features developed by Ing. Jakub Schneider Ph.D. [42].

- **date** – dates where there were recorded data
- **free** – [0 or 1] for each represents whether it was a free day (holidays and weekends)

- **Cosine analysis** – cosine analyses parameters with variable window length (generally 7 or 14 days marked at the end of each parameter name)
  - **amplit_X** – amplitude of fitted cosine function one for each day (based on the window)
  - **phi_X** – phase shift of activity fitted cosine function one for each day (based on the window)
  - **mesor_X** – mean value of activity fitted cosine function one for each day (based on the window)
  - **DARs_X** – daily activity rhythm calculated as amplit_X/mesor_X (also named circadian quotient CQ)
  - **GoF_X** – Goodness of Fit represented as Mean Square of Errors (MSE)
  - **GoF_expprc_X** – Goodness of Fit represented as a percentage of explained variability $100 \times (\mathrm{TSS} - \mathrm{RSS})/\mathrm{TSS}$ where TSS is the total sum of squares and RSS is the residual sum of squares
- **Nonparametric analyses** – statical parameters calculated for a window (X at the end of the name marks the length of the window) or for each day
  - **IV_X** – Intradaily variability for the window of X days ignoring days with more than 20% unknown values, sampling period set to 20 min
  - **IS_X** – Interdaily stability for the window of X days, ignoring days with more than 20% unknown values, sampling period set to 20 min
  - **IV60_X** and **IS60_X** – the same as previous but based on the original definition with the sampling period of 1 hour
  - **M10** – activity within 10 most active hours of the day, midnight to midnight for each day (calculation may use data from the previous and following day)
  - **M10_time** – mid-time of the 10 most active hours window
  - **L5** – activity in the least active 5 hours
  - **L5_time** – mid-time of the 5 least active hours window
  - **RA** – relative amplitude between M10 and L5 parameters
  - **M10_RMSSD** – root mean square of successive differences (RMSSD) of activity signal in the M10 window
  - **M10_SD** – Standard deviation (SD) of activity signal in the M10 window
  - **M10_X**, **M10_time_X**, **L5_X**, **L5_time_X**, **RA_X** – the same as previous but calculated from an average day in the X-day-long window
- **Activity profile** – mean activity within the days
  - **daily_act** – mean activity midnight to midnight

  - **midn2morn** – mean activity between 0:00-6:00 local time
  - **morn2noon** – mean activity between 6:00-12:00 local time
  - **noon2even** – mean activity between 12:00-18:00 local time
  - **even2midn** – mean activity between 18:00-24:00 local time
  - **daily_acti_points** – sum of actipoints provided for each 5-min segment based on physical activity levels associated with different levels of physical activity/exercise
  - **daily_acti_points_sleep_corr** – average actipoint score for active (non-sleep) part of day

- **Relative activity levels**
  - **dayAct_high** – a percentage of activity that for a specific day is higher than 75% quantile of the whole user's activity
  - **dayAct_moderate** – a percentage of activity in the range between 50% and 75% quantiles of the whole user's activity
  - **dayAct_sedentary** – a percentage of activity in the range between 25% and 50% quantiles of the whole user's activity
  - **dayAct_low** – a percentage of activity that for a specific day is lower than 25% quantile of the whole user's activity
  - **RMSSD_daily** – RMSSD of activity from midnight to midnight

- **Sleep analyses** –results of sleep analyses based on main daily sleep
  - **Sleep_duration** – sleep duration for each day in hours (main - the longest sleep)
  - **Sleep_dur_daily18** – a sum of sleep durations of all sleeps longer than 5 minutes that occur between the previous day 18 o'clock to the associated day 18 o'clock
  - **Sleep_dur_daily** – a sum of sleep durations of all sleeps longer than 5 minutes that occur between in a calendar day (midnight to midnight)
  - **Sleep_dur_daily** – sleep duration for each day (sum of all periods of sleep)
  - **Miss_dur_daily** – amount (hours) of missing data for each day
  - **Sleep_midtime** – the middle of sleep
  - **Sleep_imobile** – 0-1 (0 - 100%) part of sleep detected as immobile
  - **Sleep_active** – 0-1 part of sleep detected as active
    - **Sleep_on** – time of the main daily sleep onset
    - **Sleep_off** – time of the main daily sleep offset
    - **sle_imobile_BefAftMS_1** – the same as Sleep_imobile, but this value represents only the part before midsleep

  - **sle_imobile_BefAftMS_2** – the same as Sleep_imobile, but this value represents only the part after midsleep
  - **sle_active_BefAftMS_1** – the same as Sleep_active, but this value represents only the part before midsleep
  - **sle_active_BefAftMS_2** – the same as Sleep_active, but this value represents only the part after midsleep
  - **RMSSD_sleep** – RMSSD of activity during the main daily sleep
  - **RMSSD_sleepf** – same as previous, but the sleep actigraph is filtered by a median filter with a dimension of 10
  - **WASO** – wake after sleep onset, a sum of minutes during a night that are not detected as sleep (based on sleep classifier) (it may be affected by wrongly detected wearable removal)
- **bef_fall** – activity in 2 hours before the main daily sleep begins
- **aft_fall** – activity in 2 hours after the main daily sleep begins
- **bef_wake** – activity in 2 hours before the main daily sleep ends
- **aft_wake** – activity in 2 hours after the main daily sleep ends
- **bef_fall_std** – variability in activity 2 hours before sleep onset (Sleep_on)
- **aft_fall_std** – variability in activity 2 hours after sleep onset (Sleep_on)
- **bef_wake_std** – variability in activity 2 hours before wake up (Sleep_off)
- **aft_wake_std** – variability in activity 2 hours after wake up (Sleep_off)
- **sl_fall_step** – step in average activity in 1 hour before and after sleep onset
- **sl_wake_step** – step in average activity in 1 hour before and after wake up
- **RMSSD** – root mean square of successive difference (RMSSD) of sleep activity data

- **Fractal and other measures of complexity**
  - **SampEnt** – Sampling entropy based on physio.net codes parameters set to m = 2, r = 0.2, window 7 days, sample period one hour
    - **SlopeEntr_M10t** – slope entropy for daily M10 window
    - **min_lag_aft_wake** – correlation of activity after wake up with itself shifted by 1 minute
    - **lag_act_day** – correlation of activity of the active part of the day (no-night-sleep) and activity shifted by 5 minutes
    - **RMSSD_SD_act_day** – ratio of RMSSD and SD for a calendar day
    - **RMSSD_rel_act_day** – RMSSD for a calendar day normalised by average activity during that day
    - **SD_rel_act_day** – SD for a calendar day normalised by average activity during that day

- **Technical monitoring**

  - **Miss_dur_daily** – sum of durations when there are no valid data from the given wearable, both missing and off

  - **Miss_dur_daily18** – same as previous, but the daybreak is set at 18 o'clock

  - **Offs_dur_daily** – duration of off periods, the time when the wearable was most probably removed in a given day

  - **Offs_dur_n_seg** – number of separate off periods, how many times the wearable was removed during a day

  - **miss_day_part** – missing data part of the day, ranging from 0 to 1 (considering daylight saving time changes)

  - **rec_day_part** – valid data ranging from 0 to 1 for a given day (considering daylight saving time changes)

  - **miss_7** – missing data for a 7-day window

  - **miss_14** – missing data for a 14-day window

  - **nan_ratio_M10** – missing data in the M10 daily segment

4.2.2 Annotation

Manually created labels based on ASERT questionnaires and MADRS and YMRS clinical scales (see section 2.2.3) were used as ground truth labels.

The Aktibipo Self-rating ecological momentary assessment (ASERT) is a self-report mood questionnaire designed for BD patients invented in NIMH [63]. It is a 10-item mobile app-based questionnaire consisting of depression (4), mania (4), and nonspecific (2) symptom items, each with 5 possible response levels.

group	question
Depressive	
1	I feel sad, downhearted
2	I do not enjoy anything, and nothing pleases me
3	I have no energy
4	I feel gloomy and pessimistic about the future
Manic	
5	I feel unusually great, optimistic
6	I have excess energy
7	My thinking is very fast, others cannot keep up with me
8	I need to sleep less than usual
Nonspecific	
9	I feel restless, tense
10	I cannot focus
Reply options:	
	0 = I do not agree; **1** = more likely I do not agree;
	2 = I probably agree; **3** = I agree; **4** = I completely agree

Table 4.1: Aktibipo Self-rating questionnaire (ASERT)

Based on these self-assessments and clinical scales, each patient's day was labelled by one of the following states:

- remission
- mania-onset
- mania
- mania-offset
- depression-onset
- depression
- depression-offset
- unknown

After aligning the provided labels with the corresponding actigraphic data, a dataset was obtained. A summary of the lengths of the individual states using descriptive measures—25th percentile (p25), median and 75th percentile (p75) is given in the following table 4.2.

	labelled states								
data length [days]	remission	mania onset	mania	mania offset	depression onset	depression	depression offset	unknown	**all states**
p25	21	0	0	0	0	0	0	0	95
p50	71	0	0	0	9	0	9	0	176.5
p75	172	14	0	9.5	34	31	32	0	324.5
total	10803	991	506	824	2412	2164	2139	261	20100

Table 4.2: Original dataset annotation overview (all 92 patients)

4.2.3 Dataset Preprocessing

As part of the data preprocessing, an analysis of annotated actigraphy data from 92 patients with BD was performed. The aim was to simplify the task of identifying states and to optimise the dataset for subsequent scientific analysis. In a first step, the data were filtered based on the presence of the labels "onset", "offset" or "unknown". This removed any days that did not serve to clearly categorise the states, which resulted in narrowing the dataset to relevant information. Subsequently, a distinction was made between dates labelled as "remission" and "depression" or "remission" and "mania" for the purpose of two separate tasks of identifying relevant states. This step achieved better segmentation and specialisation of the data for specific analytical tasks. In order to achieve a more balanced dataset, a reduction of days in remission was performed. This reduction was performed by selecting blocks of consecutive days of mania or depression and adding an equal number of days in remission (or the maximal days possible). Specifically, half of the days before a given block and half of the days after a given block were selected. Overall, the modifications made to the dataset were aimed at eliminating ambiguity and bias in the data, while creating a balanced and simplified dataset for subsequent analysis of actigraphic features structures in BD.

An overview of the dataset before and after preprocessing is shown in the following three tables using the p25, p50 and p75 of data length in days:

	labelled states		
data length [days]	remission	mania	**all selected states**
p25	19	20	39
p50	22	21	43
p75	38	37	75
total	259	259	518

Table 4.3: Mania-remission dataset annotation overview (9 patients)

	labelled states		
data length [days]	remission	depression	**all selected states**
p25	20	21	42
p50	25	25	50
p75	43	41	84
total	558	53	1111

Table 4.4: Depression-remission dataset annotation overview (18 patients)

It should be noted that out of the original dataset of 92 patients with approximately 2 years of monitoring, only 9 patients with manic states and 18 patients with depressive states remained after this data preprocessing.

4.2.4 Training and Testing Sets Split

From these modified data, described in section 4.2.3, we created 3 different train-test sets splits to be used for our analyses[1]:

4.2.4.1 Patient-based split dataset

The first approximately 70% of patients were used for the training and cross-validation set, the remaining 30% as a test set. This approach aims to evaluate the model's ability to generalize across a diverse patient population. Moderate to high accuracy on the test set is expected, indicating the model's capacity to generalise to new patients.

[1]Clearly, only the state-based split and time-based split were used to identify the episode for each patient separately.

4.2.4.2 State-based spilt dataset

Independently of the patients, each episode (mania or depression) and remission was divided in a ratio 7:3 and the longer part was used for the training and cross-validation set and the shorter part for the test set. This method assesses the model's capability to discern patterns within distinct states of the patients' conditions. Moderate to high accuracy is expected, especially if the model can effectively capture state-specific features. However, generalisability might be lower if states vary significantly between patients.

4.2.4.3 Time-based split dataset

In order to generate the most dependent training and testing data possible, we used every third day (again independently of the patients) for the test set and the remaining days for the training and cross-validation set. Potentially high accuracy is expected if the model effectively captures temporal patterns. However, there might be a risk of overfitting to specific temporal intervals, affecting generalisability to different time-frames. Such a dependent distribution was developed primarily for the purpose of testing the model itself with the least possible influence of longitudinal changes in the data.

Chapter 5

Results

This chapter presents the results of all experiments. It is divided into two sections based on the approach: unsupervised 5.1 and supervised 5.2.

5.1 Unsupervised Approach

First, we will focus on selected methods of cluster assignment evaluation in section 5.1.1, followed by the actual evaluation of the TICC method on simulated data in section 5.1.2.

5.1.1 Cluster Assignment Evaluation

In order to compare the characteristics of selected clustering evaluation criteria for all possible cases of clustering results, we investigated these approaches (see section 3.1.1.2) on simulated data with 137 samples. If M is the number of clusters and N is the number of classes in the gold standard, the possible clustering results can be:

- $M = N$ ... number of clusters is same as the number of classes in the gold standard,
- $M < N$ ... number of clusters is less than the number of classes in the gold standard,
- $M > N$ ... number of clusters is greater than the number of classes in the gold standard.

Case $M = N$

Let us assume the first case where we have 5 clusters and 5 classes in the gold standard, data with 137 samples, 127 clustered samples and 10 unclustered samples. We get the 5×5 contingency table:

The results of clustering evaluation criteria computed using the formulas described in the the section 3.1.1.2 are:

$$Precision = 91.34\%, \qquad Recall = 84.67\%, \qquad F1\text{-}score = 87.88\%, \qquad ARI = 77.73\%.$$

	Classes j				
Clusters i	1	2	3	4	5
1	26	1	0	3	0
2	0	13	1	2	0
3	0	0	19	0	0
4	4	0	1	35	0
5	0	0	0	0	23

Table 5.1: Contingency table of the clustering results with same number of clusters and classes

Case $M < N$

Let us assume the second case where we have 4 clusters and 5 classes in the gold standard, data with 137 samples, 127 clustered samples and 10 unclustered samples. We obtain the 4×5 contingency table:

	Classes j				
Clusters i	1	2	3	4	5
1	26	1	0	3	0
2	0	13	1	2	0
3	0	0	19	0	0
4	4	0	0	35	23

Table 5.2: Contingency table of the clustering results with the number of clusters less than the number of classes

The results of clustering evaluation criteria computed using the formulas described in the the section 3.1.1.2 are:

$$Precision = 73.23\%, \qquad Recall = 84.67\%, \qquad F1\text{-}score = 78.54\%, \qquad ARI = 53.38\%.$$

Case $M > N$

Let us assume the last case where we have 6 clusters and 5 classes in the gold standard, data with 137 samples, 127 clustered samples and 10 unclustered samples. We can get the 6×5 contingency table:

The results of clustering evaluation criteria computed using the formulas described in the the section 3.1.1.2 are:

$$Precision = 91.34\%, \qquad Recall = 77.37\%, \qquad F1\text{-}score = 83.78\%, \qquad ARI = 72.20\%.$$

	Classes j				
Clusters i	1	2	3	4	5
1	26	1	0	3	0
2	0	13	1	2	0
3	0	0	19	0	0
4	4	0	0	35	0
5	0	0	0	0	13
6	0	0	0	0	10

Table 5.3: Contingency table of the clustering results with the number of clusters greater than the number of classes

Overview

Let us compare all of the obtained values for each of the example cases. First, the case with a nonzero number of unclustered samples as above:

	Cases		
evaluation criteria	$M = N$	$M < N$	$M > N$
Precision	91.34%	$\searrow$ 73.23%	$\rightarrow$ 91.34%
Recall	84.67%	$\rightarrow$ 84.67%	$\searrow$ 77.37%
F1-score	87.88%	$\searrow$ 78.54%	$\searrow$ 83.78%
ARI	77.73%	$\searrow$ 53.38%	$\searrow$ 72.20%

Table 5.4: Precision, Recall, F1-score and ARI values for each of the example cases with nonzero number of unclustered samples

As shown in table 5.4, if the clustering method under-estimates the number of clusters (case $M < N$), in other words, each cluster can be assigned to multiple gold standards, the recall value will remain the same, and the ARI, precision and thus also F1-score and will decrease compared to the $M = N$ case. Whereas, if the clustering method over-estimate the number of clusters (case $M > N$), in other words, each gold standard can be represented by multiple clusters, the precision will remain the same, and the ARI, recall and thus also F1-score will again be lower compared to the M=N case.

If we were to consider the case where all samples are clustered, i.e. the number of unclustered samples is zero and the number of clustered samples is 127, we would get a similar result. Only in the case M=N we would get the same value for precision, recall and F1-score as shown in the following table.

	Cases		
evaluation criteria	$M = N$	$M < N$	$M > N$
Precision	91.34%	$\searrow$ 73.23%	$\rightarrow$ 91.34%
Recall	91.34%	$\rightarrow$ 91.34%	$\searrow$ 83.46%
F1-score	91.34%	$\searrow$ 81.29%	$\searrow$ 87.22%
ARI	77.73%	$\searrow$ 53.38%	$\searrow$ 72.20%

Table 5.5: Precision, Recall, F1-score and ARI values for each of the example cases with zero number of unclustered samples

From the table 5.5, we can see that the number of unclustered samples does not have impact on the precision and ARI result value (case $M = N$). On the contrary, recall and F1-score will increase in the case of zero number of unclustered samples. If the clustering method underestimates the number of clusters (case $M < N$), some ground truth classes may be merged into a single cluster, resulting in a lower recall value because not all true classes are identified. However, the precision may increase since merging classes can reduce the number of false positives. Conversely, if the number of clusters generated by the algorithm is greater than the number of ground truth classes (case $M > N$), the precision and recall values will also be affected. In this case, the algorithm may split some true classes into multiple clusters, resulting in a lower precision value because of the increased number of false positives. However, recall may increase since each true class is more likely to be identified in one of the multiple clusters.

Again, here is the criteria evaluation for the case of zero unclustered samples, this time over 10000-iteration in (mean value of criteria for 10000 different datasets):

	Cases		
evaluation criteria	$M = N$	$M < N$	$M > N$
Precision	76.1%	$\searrow$ 68.9%	$\rightarrow$ 76.1%
Recall	76.4%	$\rightarrow$ 77.1%	$\searrow$ 63.2%
F1-score	76.2%	$\searrow$ 72.6%	$\searrow$ 68.8%
ARI	54.2%	$\searrow$ 42.7%	$\searrow$ 44.3%

Table 5.6: Precision, Recall, F1-score and ARI values for each of the example cases with zero number of unclustered samples over 10000-iterations

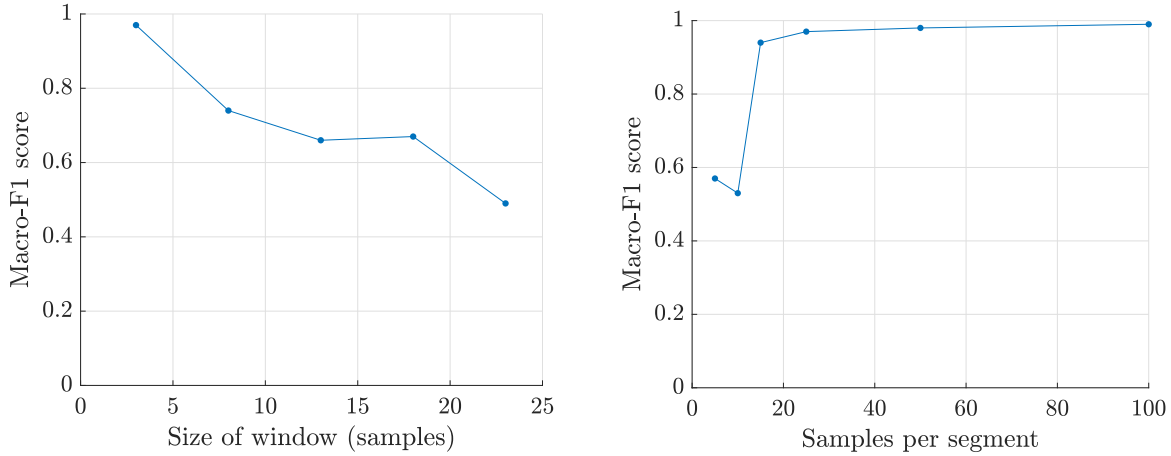
From this table 5.6 it s evident that the above observation is valid even if 10000 different datasets are used.

5.1.2 TICC Method Evaluation

Based on evaluation using simulated data and parameter tuning using grid-search, we found that TICC is very robust to the sparsity parameter λ selection and, relatively, to wisely chosen

window length w. Thus, the key parameters are the one controlling temporal consistency β and the number of clusters K, which in our case is known in advance.

The following figures show the effect of window length (parameter w) and number of samples per segment for one selected signal from our simulated dataset, described in section 4.1. This signal was generated by two repeating classes (class 1 and class 2) as follows: 121212. In the case of testing the window length (a), the optimal combination of the λ and β parameters was chosen using a grid search (see section 3.1.1.1) with equal covariance difference between the classes and 25 samples per each segment. In the case of testing the number of samples per segment, the window size was fixed at $w = 3$, the covariance difference between the classes were equal and again the best combination of the λ and β parameters was chosen. Since all selected methods were correlated, we only displayed the macro-f1-score criterion on the y-axis, since it has the best predictive value due to its definition (see definition 3.8).



(a) Effect of window size on clustering accuracy

(b) Effect of number of samples per segment on clustering accuracy

Figure 5.1: Evaluation of TICC behaviour on simulated data with 25 samples per segment for evaluation window size (a) and fixed window length w = 3 for segment size evaluation (b)

As you can see from the figure 5.3 (a) the window size must be much smaller than the length of the cluster, because as the window size approaches the number of samples per segment the accuracy decreases. Similar information is given by the figure (b), which shows the decrease in accuracy with smaller samples per segment.

Although TICC is very robust to even small differences in covariances between classes, it requires much larger sampling than our real data. The authors of this method presented in their work a large decrease in accuracy of the method even at 100 samples per segment [1]. i In our evaluation, we were able to successfully cluster even much shorter segments (20 or more samples per segment) on suitably simulated data. However, segments with less than 10 samples typically appear in our real data, therefore this method could not be successfully applied.

5.2 Supervised Approach

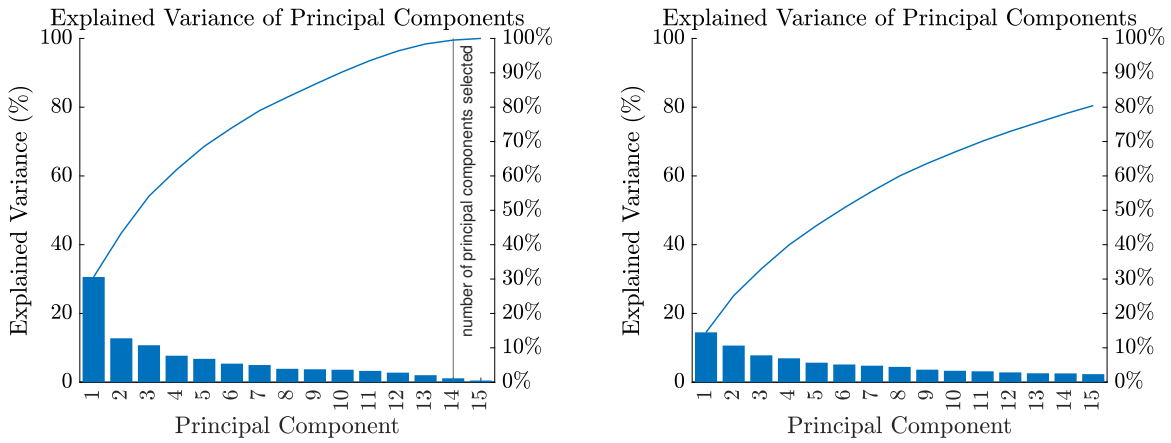
In this section, we present the results of supervised classification of BD states (mania-remission or depression-remission) using different methods of splitting training and testing datasets. We employed three distinct approaches: patient-based split, state-based split, and time-based split (see section 4.2.4). For each dataset, we computed correlation matrices COV from actigraphic features and their values (median and SD) and then reduced dimensionality using the PCA or the max-diff method (see results in section 5.2.1) and used the resulting data as input for the RBF SVM algorithm in three different validation scenarios (described in section 3.2).

5.2.1 Feature selection

In this section, we elucidate the outcomes obtained from by applying feature dimension reduction techniques. The initial subsection delineates the observed behavior of the data subsequent to the application of PCA on the COV, median, and SD methods. Following this, the subsequent segment expounds upon the data characterisation ensuing the implementation of the max-diff method.

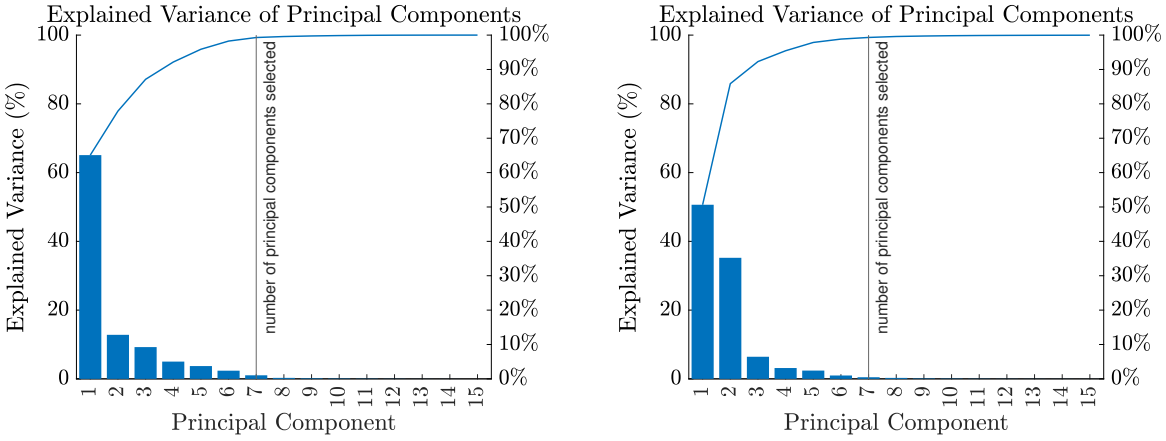
5.2.1.1 PCA method

The following figures show the explained variance of the principal components on the COV as well as on the median and SD values.



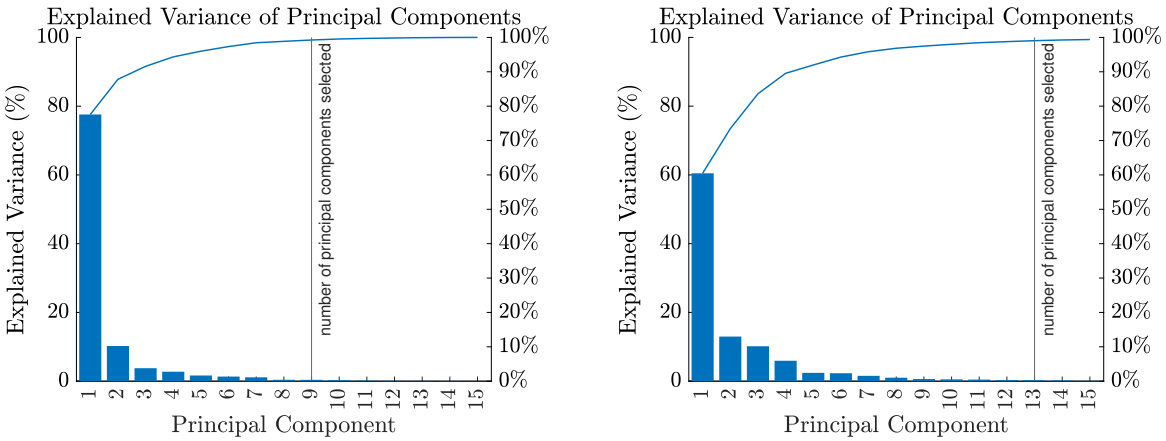
(a) mania-remission (b) depression-remission

Figure 5.2: Explained Variance of Principal Components on COV Structures



(a) mania-remission (b) depression-remission

Figure 5.3: Explained Variance of Principal Components on Median Values



(a) mania-remission (b) depression-remission

Figure 5.4: Explained Variance of Principal Components on SD Values

In most cases, 99% of the variance was explained by less than 15 components — a hand-picked threshold. In the case of COV on depression-remission states, this variance was the lowest. The first 15 largest principal components captured only 80% of the variance which is still sufficient.

5.2.1.2 Maximal Difference Pair Method

The optimal strategy involves training the classifier exclusively on a subset of individuals and then evaluating the performance of the model on different patients. Unfortunately, splitting the training and testing data patient-based may yield inaccurate results due to the considerable variability of actigraphy data between different subjects. This phenomenon is clearly illustrated in the following figures. All following visualisations are created from the original full dataset (without remission reduction).

In the figure 5.5, which shows the comparison of the absolute difference between the correlation coefficients for the individual pairs of states (remission–mania, remission–depression, mania–

depression) on the whole dataset, we can see that the largest absolute difference of the correlation matrices is between the states of remission–mania, followed by mania–depression, and on the contrary the smallest difference is achieved by the states of remission–depression. From the analogously obtained figure 5.6 and the figure 5.7 that shows the values of the correlation coefficients for the individual features in the states between the patient-based split training and testing data, we can observe a large difference in the values between the sets – i.e. the patients. Similarly, from the figures 5.8 and 5.9 for the state-split datasets, we can again see the significant difference in values between the training and testing sets. Whereas for the remission state, the values are very similar. The most comparable correlation values between the training and testing sets are observed for the case of the time-split dataset, shown in the figure 5.10 and figure 5.11, which are highly dependent. Here we can observe remarkably close values for all states.

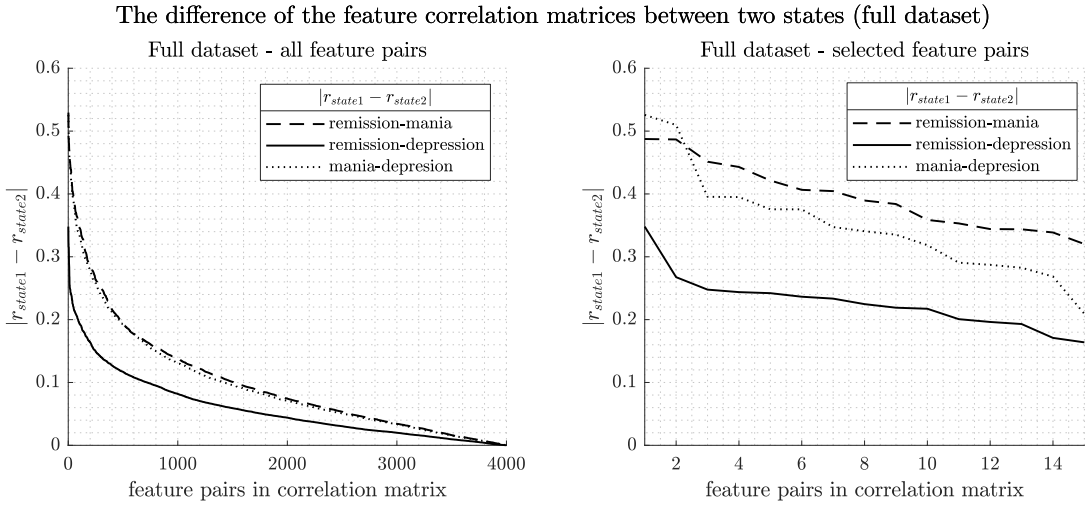


Figure 5.5: The difference of the feature correlation matrices between two states on full dataset

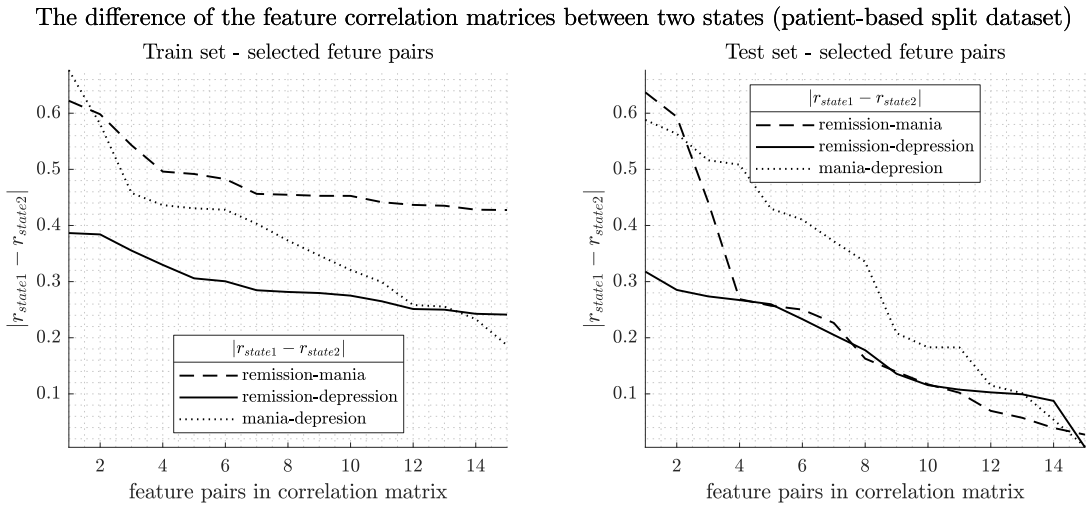


Figure 5.6: The difference of the feature correlation matrices between two states on patient-based dataset

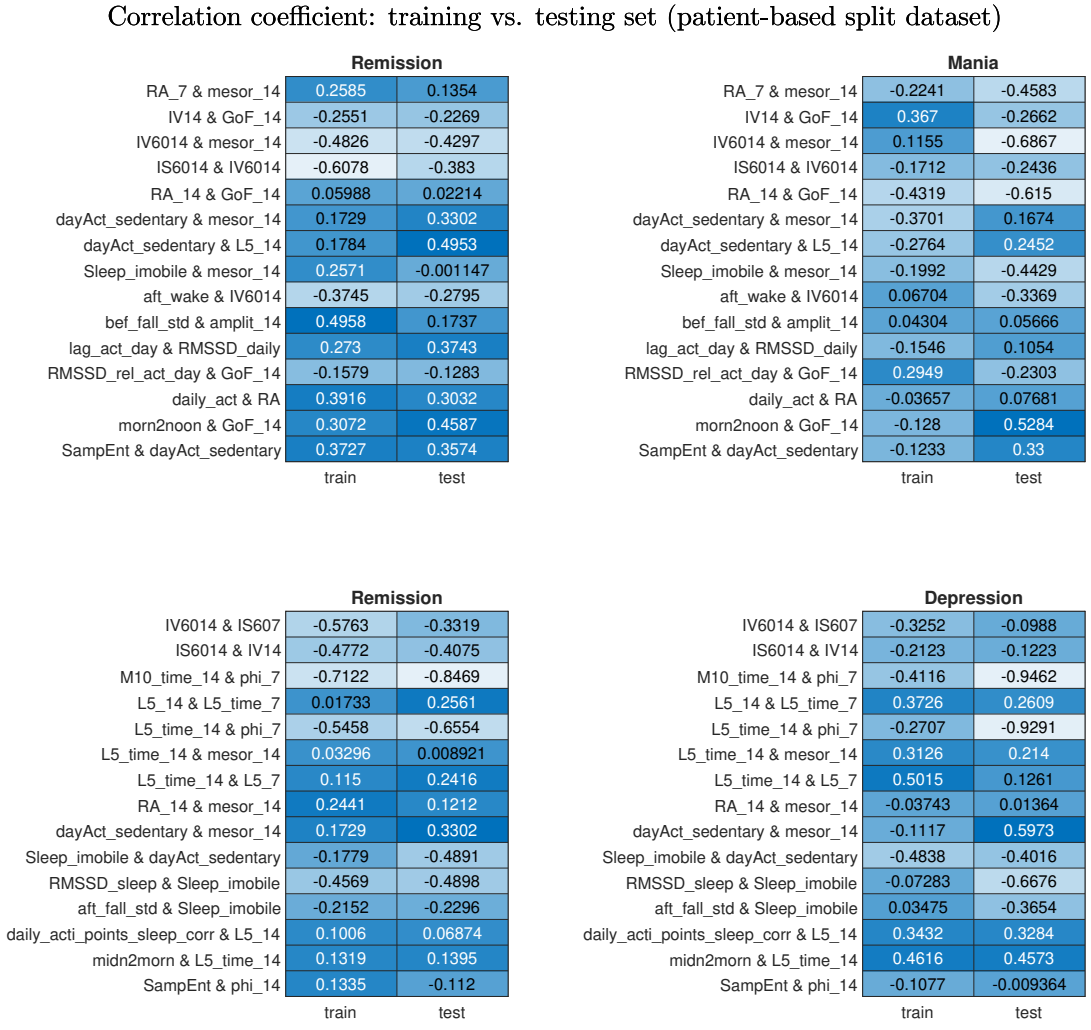


Figure 5.7: Correlation coefficient: training vs. testing set for patient-based split dataset.

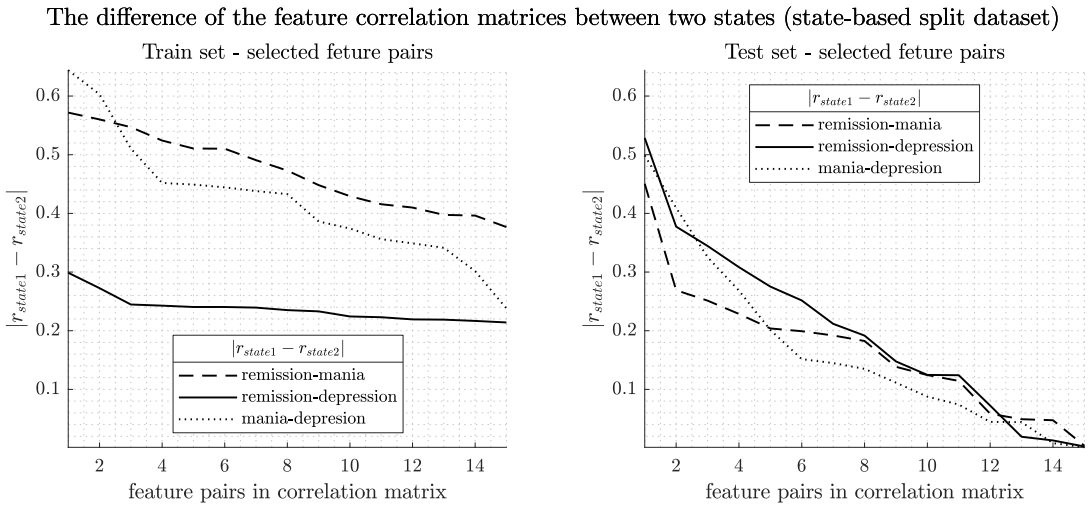


Figure 5.8: The difference of the feature correlation matrices between two states on state-based dataset

Correlation coefficient: training vs. testing set (state-based split dataset)

	Remission	
	train	test
GoF_14 & amplit_7	0.5178	0.525
IV607 & mesor_14	-0.4373	-0.4149
IV6014 & IS14	-0.6553	-0.6545
M10_14 & IV6014	-0.6169	-0.617
dayAct_sedentary & mesor_14	0.2083	0.1926
dayAct_sedentary & L5	0.2463	0.2421
RMSSD_daily & mesor_7	0.6899	0.6593
Sleep_imobile & GoF_14	0.1458	0.1255
bef_fall_std & Sleep_imobile	0.2126	0.2015
lag_act_day & RMSSD_daily	0.2985	0.2751
miss_day_part & Miss_dur_daily	0.8729	0.8851
daily_acti_points & sle_imobile_BefAftMS_1	0.1767	0.1754
daily_acti_points_sleep_corr & bef_fall_std	0.5198	0.5044
daily_act & RA	0.3734	0.388
SampEnt & dayAct_sedentary	0.3681	0.3705

	Mania	
	train	test
GoF_14 & amplit_7	-0.04233	0.3258
IV607 & mesor_14	-0.007981	-0.3674
IV6014 & IS14	-0.1083	-0.6585
M10_14 & IV6014	-0.04517	-0.4251
dayAct_sedentary & mesor_14	-0.3023	-0.07642
dayAct_sedentary & L5	-0.1512	0.1176
RMSSD_daily & mesor_7	0.217	0.7085
Sleep_imobile & GoF_14	-0.3784	-0.07839
bef_fall_std & Sleep_imobile	-0.2356	0.06297
lag_act_day & RMSSD_daily	-0.2118	0.2166
miss_day_part & Miss_dur_daily	0.4962	0.9996
daily_acti_points & sle_imobile_BefAftMS_1	-0.2332	-0.05325
daily_acti_points_sleep_corr & bef_fall_std	0.1235	0.05388
daily_act & RA	-0.04221	0.1366
SampEnt & dayAct_sedentary	-0.1226	0.188

	Remission	
	train	test
L5_7 & mesor_14	0.1868	0.1836
IS6014 & IV14	-0.4755	-0.4314
L5_14 & M10_7	0.01216	0.0167
L5_time_14 & GoF_7	0.1087	0.1113
RA_14 & GoF_14	0.05686	0.04927
dayAct_high & M10_7	0.2952	0.2592
dayAct_sedentary & mesor_14	0.2083	0.1926
RMSSD_daily & IV14	-0.2513	-0.2013
RMSSD_sleep & GoF_14	0.2477	0.2589
RMSSD_sleep & Sleep_imobile	-0.4429	-0.4895
sle_imobile_BefAftMS_1 & dayAct_sedentary	-0.1821	-0.1641
aft_fall & mesor_14	-0.03738	-0.05211
min_lag_aft_wake & dayAct_sedentary	-0.1172	-0.135
SD_rel_act_day & RMSSD_SD_act_day	-0.502	-0.4605
midn2morn & amplit_7	-0.1811	-0.1936

	Depression	
	train	test
L5_7 & mesor_14	0.4272	0.5611
IS6014 & IV14	-0.2561	-0.08683
L5_14 & M10_7	0.2526	0.2919
L5_time_14 & GoF_7	0.3438	0.4196
RA_14 & GoF_14	-0.1761	-0.1625
dayAct_high & M10_7	0.5195	0.2722
dayAct_sedentary & mesor_14	-0.06425	0.06841
RMSSD_daily & IV14	-0.02825	-0.2737
RMSSD_sleep & GoF_14	0.4618	0.5105
RMSSD_sleep & Sleep_imobile	-0.1441	0.03892
sle_imobile_BefAftMS_1 & dayAct_sedentary	-0.4268	-0.3558
aft_fall & mesor_14	0.202	0.07256
min_lag_aft_wake & dayAct_sedentary	-0.3361	-0.2826
SD_rel_act_day & RMSSD_SD_act_day	-0.2592	-0.4412
midn2morn & amplit_7	0.03559	-0.1909

Figure 5.9: Correlation coefficient: training vs. testing set for state-based split dataset.

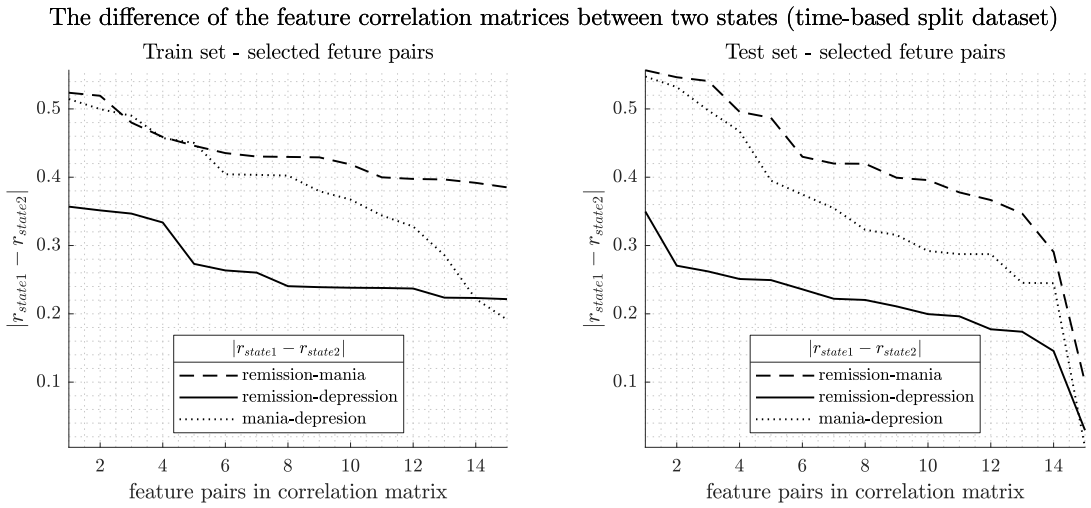


Figure 5.10: The difference of the feature correlation matrices between two states on time-based dataset

Correlation coefficient: training vs. testing set (time-based split dataset)

Remission

	train	test
IV14 & GoF_14	-0.2475	-0.2527
IV6014 & amplit_14	-0.6857	-0.685
RA_14 & GoF_14	0.05544	0.05294
dayAct_sedentary & mesor_14	0.2043	0.202
dayAct_sedentary & L5_14	0.2281	0.2379
Sleep_imobile & mesor_14	0.2253	0.2214
sle_imobile_BefAftMS_1 & GoF_14	0.1754	0.1635
aft_wake & IV6014	-0.3575	-0.3561
bef_fall_std & amplit_7	0.4496	0.4534
bef_fall_std & Sleep_imobile	0.2204	0.1869
lag_act_day & RMSSD_daily	0.2936	0.2862
miss_day_part & Miss_dur_daily	0.8815	0.8679
daily_act & RA	0.3749	0.3836
SampEnt & M10_14	-0.1764	-0.1697
SampEnt & dayAct_sedentary	0.37	0.3659

Mania

	train	test
IV14 & GoF_14	0.2717	0.2937
IV6014 & amplit_14	-0.162	-0.1442
RA_14 & GoF_14	-0.4035	-0.434
dayAct_sedentary & mesor_14	-0.2254	-0.2938
dayAct_sedentary & L5_14	-0.2022	-0.1921
Sleep_imobile & mesor_14	-0.2207	-0.3352
sle_imobile_BefAftMS_1 & GoF_14	-0.2537	-0.1833
aft_wake & IV6014	0.06122	0.04319
bef_fall_std & amplit_7	0.05289	0.03339
bef_fall_std & Sleep_imobile	-0.1714	-0.1037
lag_act_day & RMSSD_daily	-0.1863	-0.08033
miss_day_part & Miss_dur_daily	0.4462	0.9673
daily_act & RA	-0.02263	0.005813
SampEnt & M10_14	0.2088	0.226
SampEnt & dayAct_sedentary	-0.02985	-0.05381

Remission

	train	test
IS6014 & IV14	-0.4655	-0.455
M10_time_14 & phi_7	-0.7506	-0.6972
L5_14 & mesor_14	0.2028	0.2088
L5_14 & L5_time_7	0.03756	0.07909
L5_time_14 & mesor_14	0.03342	0.02897
L5_time_14 & L5_7	0.1404	0.1202
RA_14 & GoF_14	0.05544	0.05294
dayAct_sedentary & mesor_14	0.2043	0.202
dayAct_sedentary & M10_SD	-0.03186	-0.03506
RMSSD_sleep & Sleep_imobile	-0.4488	-0.4748
sle_imobile_BefAftMS_1 & dayAct_sedentary	-0.1722	-0.1846
sle_imobile_BefAftMS_2 & M10_SD	-0.02305	-0.01115
bef_fall & L5_14	0.07129	0.08413
midn2morn & L5_time_7	0.1329	0.138
miss_7 & Sleep_duration	-0.04544	0.009684

Depression

	train	test
IS6014 & IV14	-0.2265	-0.2329
M10_time_14 & phi_7	-0.4903	-0.5234
L5_14 & mesor_14	0.4759	0.4709
L5_14 & L5_time_7	0.3945	0.2754
L5_time_14 & mesor_14	0.297	0.2493
L5_time_14 & L5_7	0.4918	0.3907
RA_14 & GoF_14	-0.1816	-0.1466
dayAct_sedentary & mesor_14	-0.03385	-0.04903
dayAct_sedentary & M10_SD	-0.2555	-0.2124
RMSSD_sleep & Sleep_imobile	-0.1021	-0.125
sle_imobile_BefAftMS_1 & dayAct_sedentary	-0.41	-0.4205
sle_imobile_BefAftMS_2 & M10_SD	0.1985	0.1344
bef_fall & L5_14	0.3117	0.3335
midn2morn & L5_time_7	0.4667	0.3489
miss_7 & Sleep_duration	0.1778	0.03922

Figure 5.11: Correlation coefficient: training vs. testing set for time-based split dataset.

5.2.2 Leave-One-Patient-Out Cross-Validation

In this section we present the results of the LOOCV validation scenario. The results are divided into two sections based on the type of classified state (mania or depression). Details of the individual patient iterations are shown in the appendix (section A).

Mania

The results in table 5.7 obtained using LOOCV with the PCA feature selection approach for assessing mania–remission show that COV achieved the highest average accuracy of 61%. In contrast, median and SD showed comparable results of 56%, suggesting that these methods may be slightly more limited in their ability to distinguish mania from remission in all patients. On the other hand, in the case of feature selection using the max-diff method, a lower average accuracy of 50% was achieved by COV and median, while using the SD value accuracy was higher — 72%.

	Mania-remission (LOOCV)	
	PCA method	**max diff pairs method**
Method	*avg ACC*	*avg ACC*
COV	61.1%	50.0%
Median	55.6%	50.0%
SD	55.6%	72.2%

Table 5.7: Average accuracy (avg ACC) of classifying mania-remission using LOOCV method

Depression

The results of LOOCV using PCA distinguishing depression-remission states, shown in table 5.8, reveal that COV achieved the highest average accuracy — 61%. The remaining feature value methods achieved lower average accuracy, the median method performed with 56% accuracy, as in the case of remission and mania, and SD reached lower average accuracy of 42%. Again, the max-diff selection method achieved better results in the opposite cases than the PCA. A poor average accuracy of 50% was obtained for COV, median reached 61%, while for the SD method the value was higher 68%.

	Depression-remission (LOOCV)	
	PCA method	**max diff pairs method**
Method	*avg ACC*	*avg ACC*
COV	61.1%	50.0%
Median	55.6%	61.1%
SD	41.7%	66.7%

Table 5.8: Average accuracy (avg ACC) of classifying depression-remission using the LOOCV method

This result suggests that the SD with a combination of the PCA has the ability to discriminate both depression and mania from remission most effectively among the compared approaches. Although the average accuracy values are not very high, we can say that the PCA method of symptom selection is more suitable for COV, while the max-diff method is more suitable for other feature value methods.

5.2.3 7-day Window Across All Patients

In this section we present the results for the classification results of 7-day window across all patients validation scenario. The results are divided into three groups according to the type of dataset split and then the type of classified state (mania or depression).

5.2.3.1 Patient-Based Split Dataset

The figures below visualise the classification results on different days for different methods combined with both feature selection methods with accuracy evaluation on training and test datasets using ACC, SEN and SPEC. The following figure gives an overview of all methods and a table with an overview of evaluation metrics for all methods for the patient-based split dataset.

Mania

For the classification task for the mania-remission discrimination, the following results were obtained:

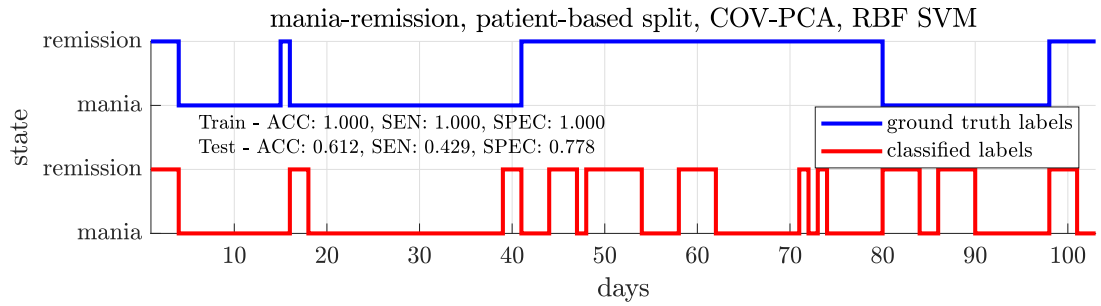


Figure 5.12: Classification of mania-remission using PCA selected COV structures on a patient-based split dataset.

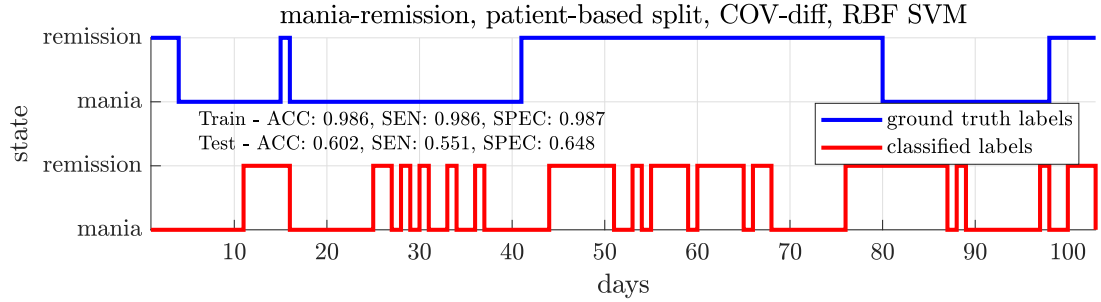


Figure 5.13: Classification of mania-remission using max-diff-based selected COV structures on a patient-based split dataset.

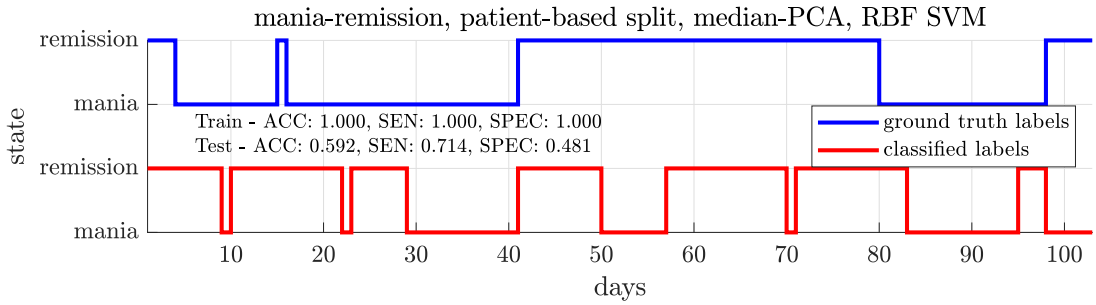


Figure 5.14: Classification of mania-remission using PCA selected median values on a patient-based split dataset.

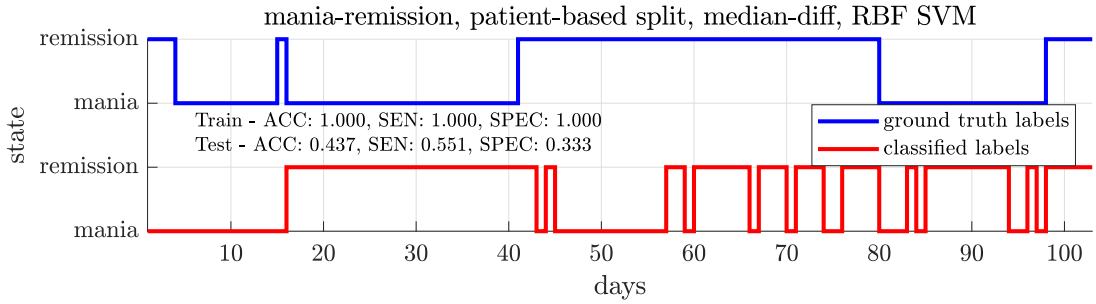


Figure 5.15: Classification of mania-remission using max-diff-based selected median values on a patient-based split dataset.

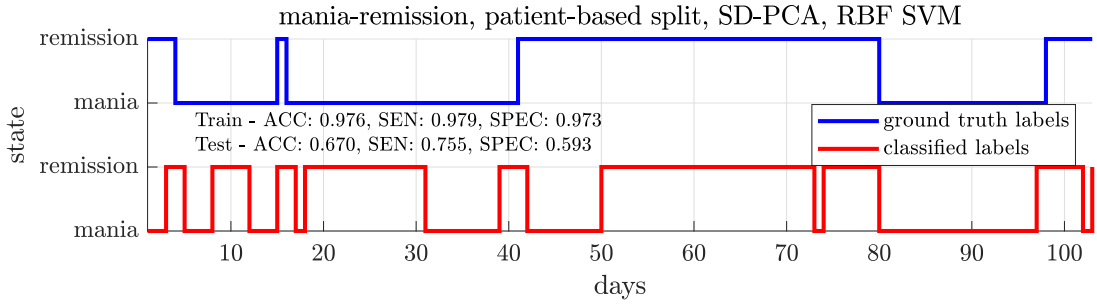


Figure 5.16: Classification of mania-remission using PCA selected SD values on a patient-based split dataset.

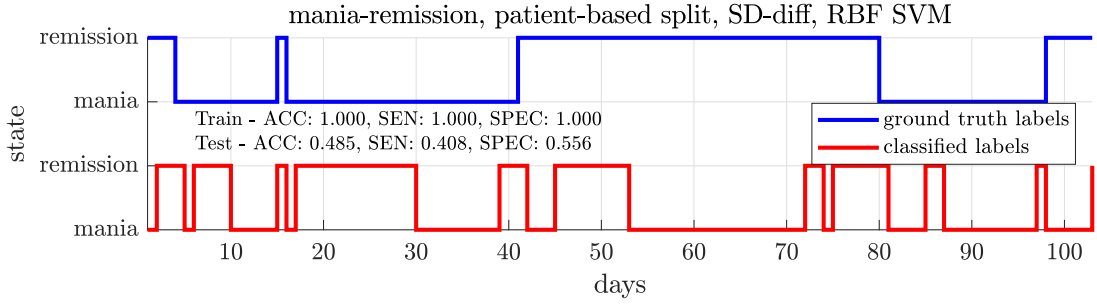


Figure 5.17: Classification of mania-remission using max-diff-based selected SD values on a patient-based split dataset.

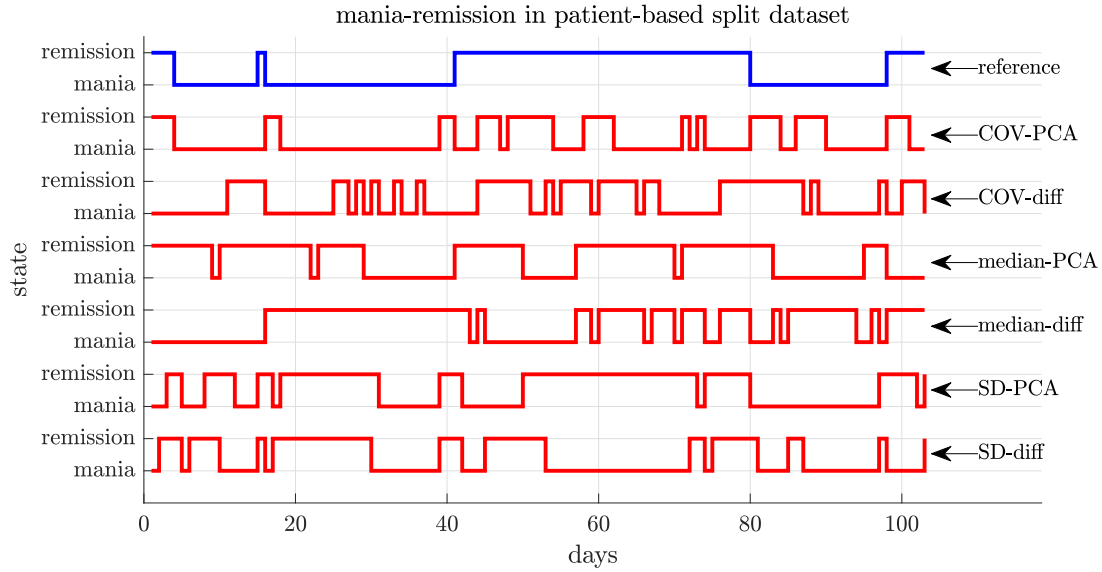


Figure 5.18: Classification overview of all feature selections on mania-remission using all methods on a patient-based split dataset.

	Mania-remission (patient-based split dataset)											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	61.22%	42.7%	72.8%	98.6%	98.6%	98.7%	60.2%	55.1%	64.8%
Median	100%	100%	100%	59.2%	71.4%	48.1%	100%	100%	100%	43.7%	55.1%	33.3%
SD	97.6%	97.9%	97.3%	67.0%	75.5%	59.3%	100%	100%	100%	48.5%	40.8%	55.6%

Table 5.9: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a patient-based split dataset.

Even though the results are relatively similar, the best results on the testing set according to the accuracy were achieved by the SD method using PCA dimension reduction with 67%. This was followed by the COV method which achieved 61% in the case of using PCA and 60 using the max-diff method. The median value method achieved noticeably better results using the PCA method, with 59%.

Depression

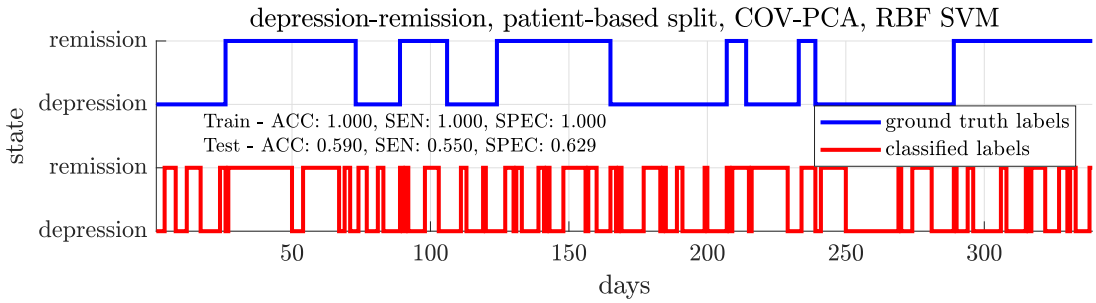


Figure 5.19: Classification of depression-remission using PCA selected COV structures on a patient-based split dataset.

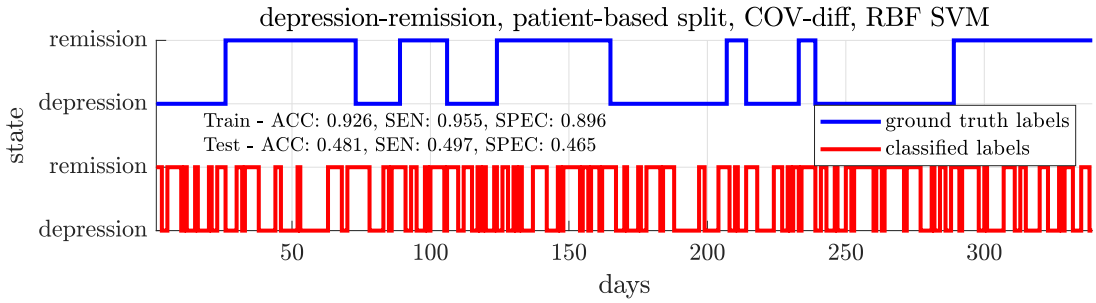


Figure 5.20: Classification of depression-remission using max-diff-based selected COV structures on a patient-based split dataset.

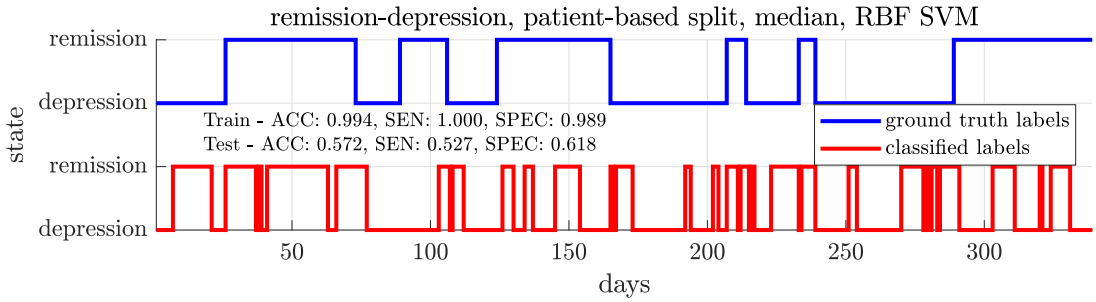


Figure 5.21: Classification of depression-remission using PCA selected median values on a patient-based split dataset.

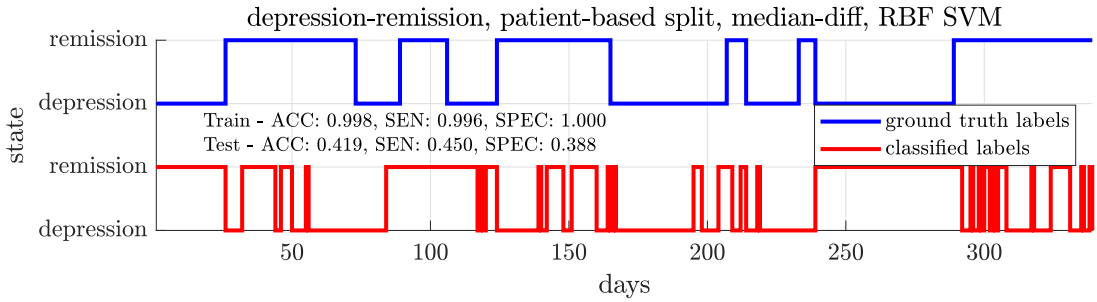


Figure 5.22: Classification of depression-remission using max-diff-based selected median values on a patient-based split dataset.

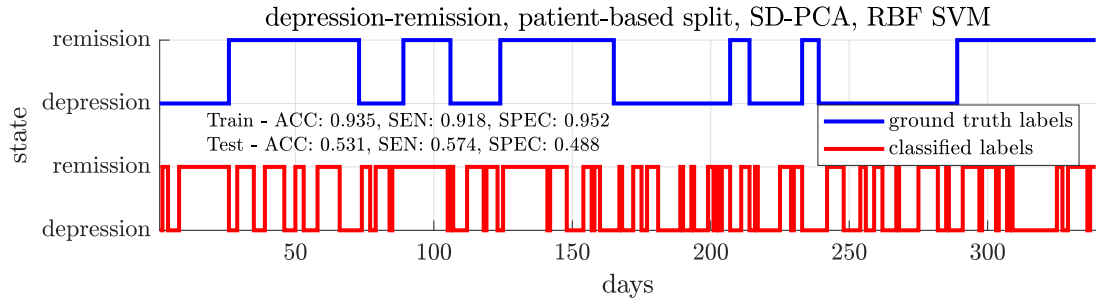


Figure 5.23: Classification of depression-remission using PCA selected SD values on a patient-based split dataset.

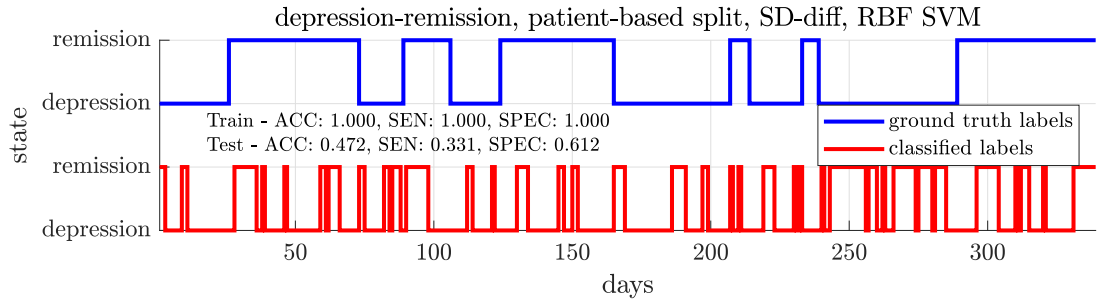


Figure 5.24: Classification of depression-remission using max-diff-based selected SD values on a patient-based split dataset.

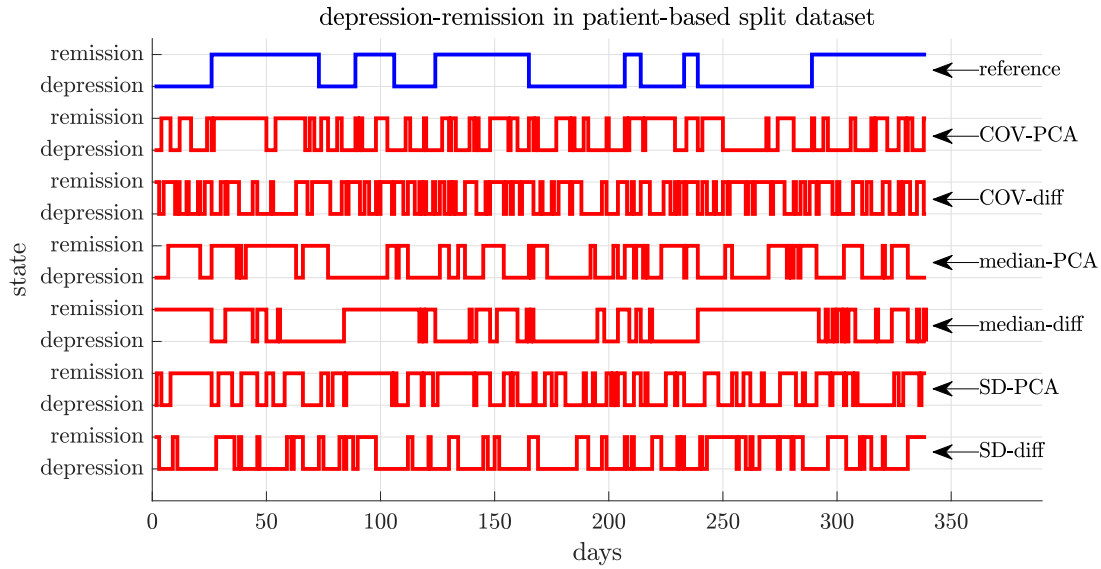


Figure 5.25: Classification overview of all feature selections on mania-remission using all methods on a patient-based split dataset.

	Depression-remission (patient-based split dataset)											
	PCA method						max diff pairs method					
	Train set			Test set			Train set			Test set		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	59.0%	55.0%	62.9%	92.6%	95.5%	89.6%	48.1%	49.7%	46.5%
Median	99.4%	100%	98.9%	57.2%	52.7%	61.8%	99.8%	99.6%	100%	41.9%	45.0%	38.8%
SD	93.5%	91.8%	95.2%	53.1%	57.4%	48.8%	100%	100%	100%	47.2%	33.1%	61.2%

Table 5.10: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a patient-based split dataset.

The distinction between depression and remission achieved significantly worse results compared to mania. This can be seen even by looking at the visualisations. All methods worked slightly better with PCA. The best accuracy was achieved by the COV 59% method followed by the median 57% and the worst was the SD method 53% on the testing set. All methods using the max-diff selection did not even reach the 50% accuracy.

5.2.3.2 State-Based Split Dataset

The following figures again show the classification results on each day using the different methods in combination with the two feature selection techniques — this time on state-based split dataset. Accuracy evaluation is performed on both training and test datasets using metrics such as accuracy (ACC), sensitivity (SEN) and specificity (SPEC). Again, a figure with a complex summary of all methods is included, supplemented by a table with overview of all results.

Mania

For the classification task for the mania-remission discrimination, the following results were obtained:

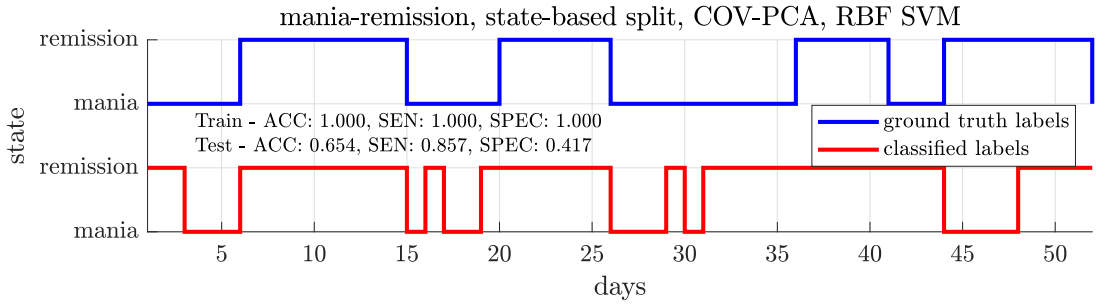


Figure 5.26: Classification of mania-remission using PCA selected COV structures on a state-based split dataset.

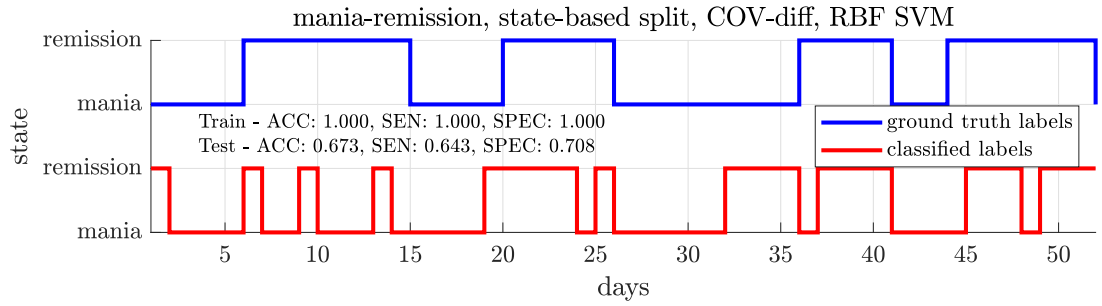


Figure 5.27: Classification of mania-remission using max-diff-based selected COV structures on a state-based split dataset.

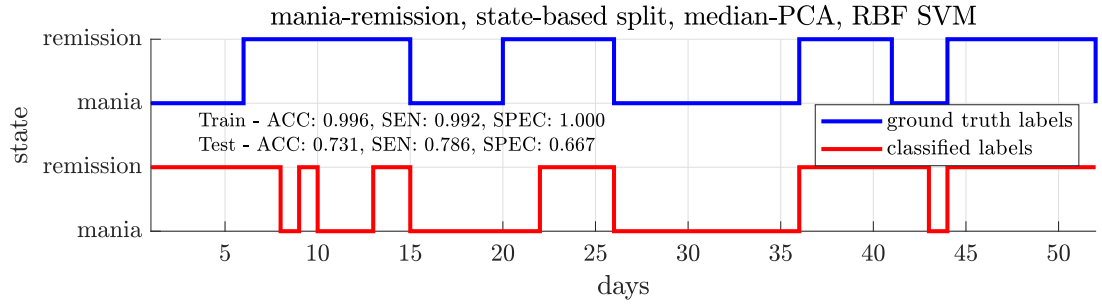


Figure 5.28: Classification of mania-remission using PCA selected median values on a state-based split dataset.

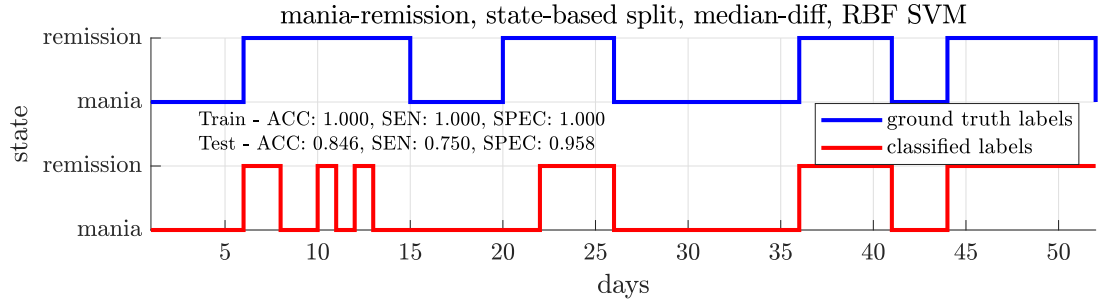


Figure 5.29: Classification of mania-remission using max-diff-based selected median values on a state-based split dataset.

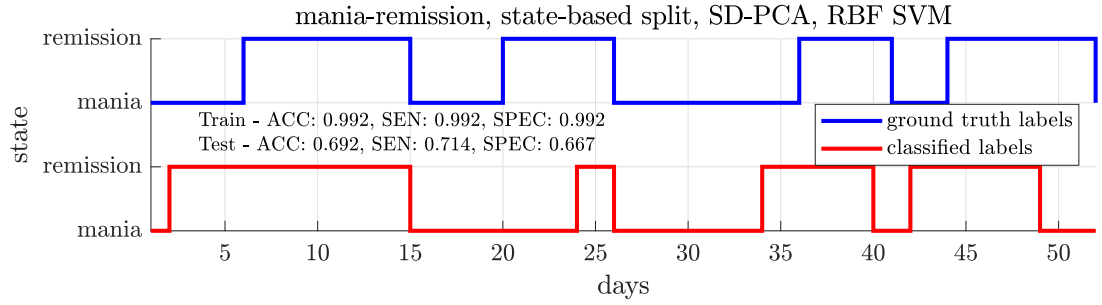


Figure 5.30: Classification of mania-remission using PCA selected SD values on a state-based split dataset.

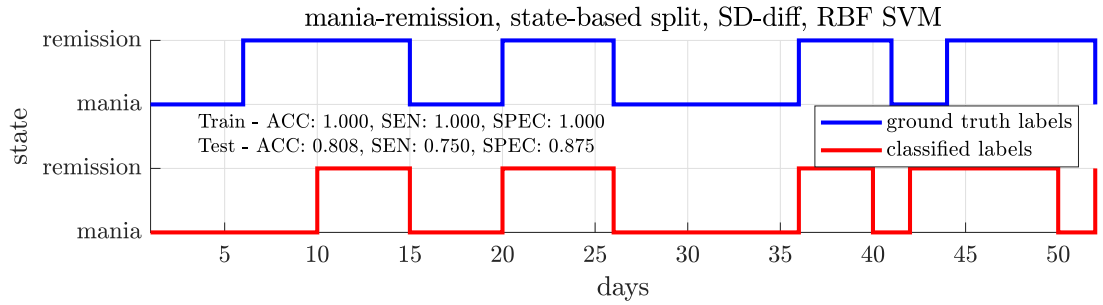


Figure 5.31: Classification of mania-remission using max-diff-based selected SD values on a state-based split dataset.

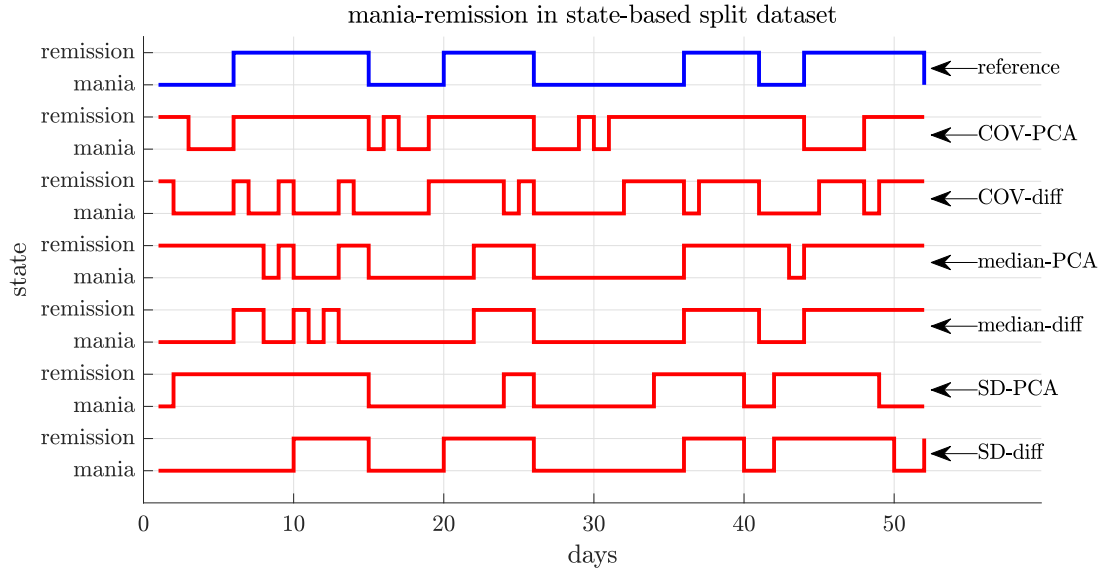


Figure 5.32: Classification overview of all feature selections on mania-remission using all methods on a state-based split dataset.

	Mania-remission (state-based split dataset)											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	65.4%	85.7%	41.7%	100%	100%	100%	67.3%	64.3%	70.8%
Median	99.6%	99.2%	100%	73.1%	78.6%	66.7%	100%	100%	100%	84.6.0%	75.0%	95.8%
SD	99.2%	99.2%	99.2%	69.2%	71.4%	66.7%	100%	100%	100%	80.8%	75.0%	87.5%

Table 5.11: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a state-based split dataset.

The best accuracy was achieved by the median value method with max-diff selection of attributes (85%). This was followed by the SD method (81%) and the COV method performed the worst (67%). In general, higher accuracy was achieved by max-diff-based reducing the dimension of the feature space for all methods.

Depression

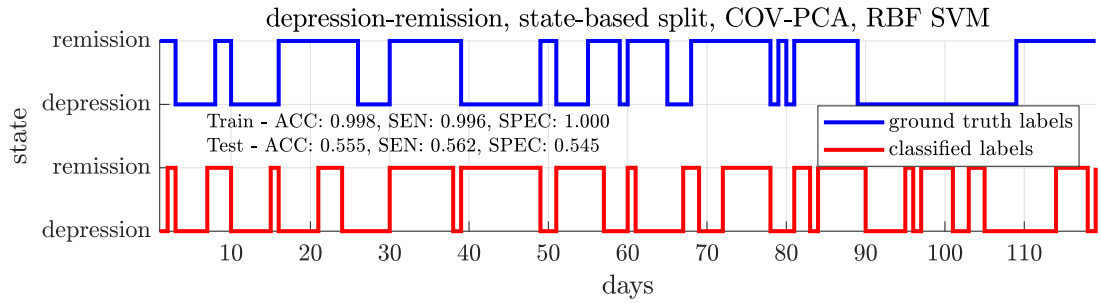


Figure 5.33: Classification of depression-remission using PCA selected COV structures on a state-based split dataset.

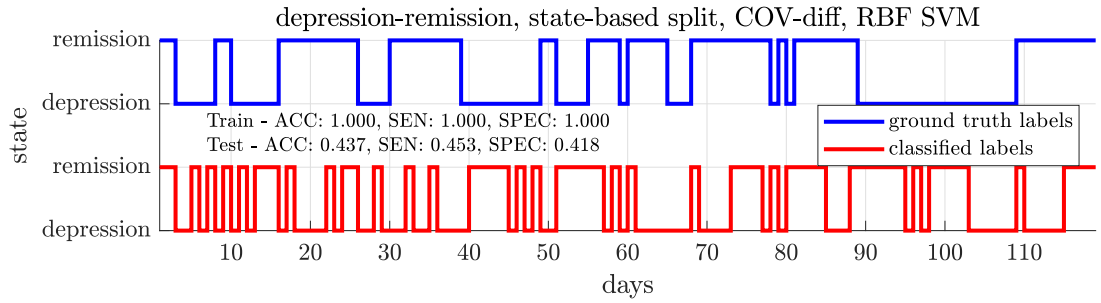


Figure 5.34: Classification of depression-remission using max-diff-based selected COV structures on a state-based split dataset.

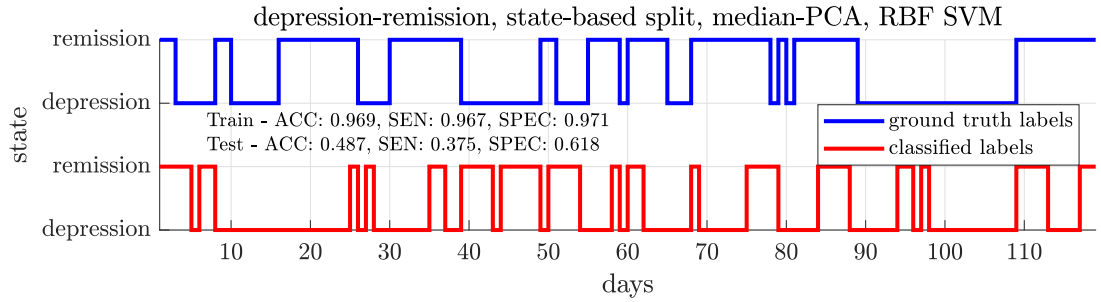


Figure 5.35: Classification of depression-remission using PCA selected median values on a state-based split dataset.

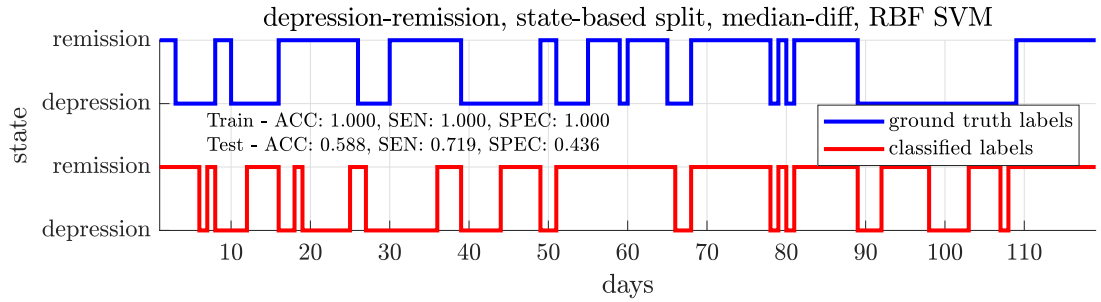


Figure 5.36: Classification of depression-remission using max-diff-based selected median values on a state-based split dataset.

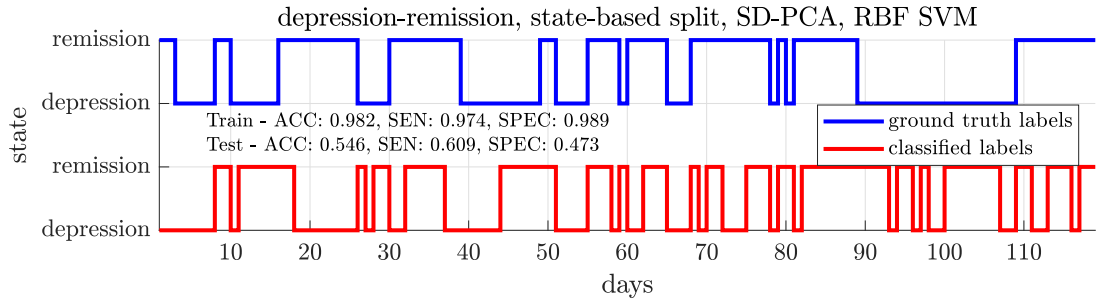


Figure 5.37: Classification of depression-remission using PCA selected SD values on a state-based split dataset.

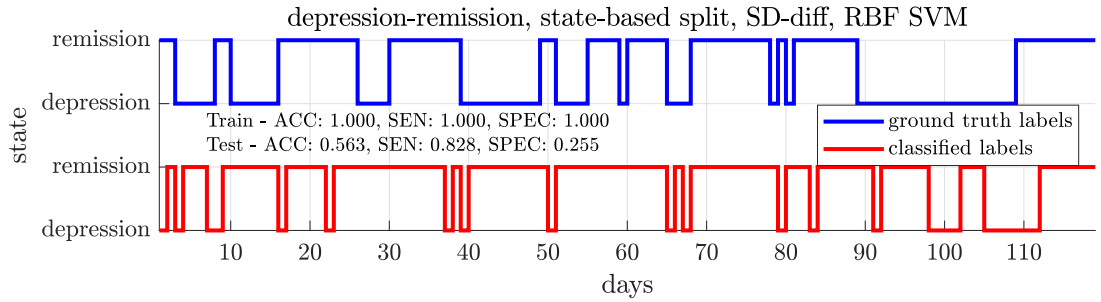


Figure 5.38: Classification of depression-remission using max-diff-based selected SD values on a state-based split dataset.

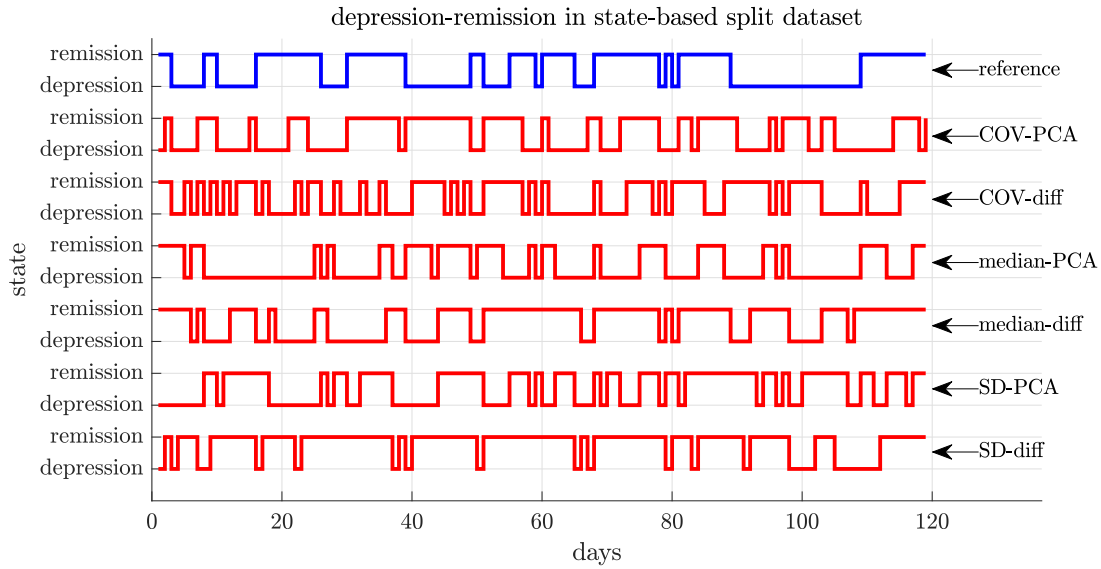


Figure 5.39: Classification overview of all feature selections on depression-remission using all methods on a state-based split dataset.

	Depression-remission (state-based split dataset)											
	PCA method						max diff pairs method					
	Train set			Test set			Train set			Test set		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	99.8%	99.6%	100%	55.5%	56.2%	54.5%	100%	100%	100%	43.7%	45.3%	41.8%
Median	96.9%	96.7%	97.1%	48.7%	37.5%	61.8%	100%	100%	100%	58.8%	71.9%	43.6%
SD	98.2%	97.4%	98.9%	54.6%	60.9%	47.3%	100%	100%	100%	56.3%	82.8%	25.5%

Table 5.12: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a state-based split dataset.

As in the previous dataset, the classification of depression in this case also achieved a significantly lower accuracy than mania. Best classification results were obtained using the median value method (59%) followed by SD method (57%) with max-diff selection of actigraphic features. Whereas the COV method performed better with the use of COV method (55%).

5.2.3.3 Time-Based Split Dataset

As above, the following figures show again the classification results of each day using different methods in combination with both feature selection techniques, this time on the time-based split dataset. The accuracy evaluation is once again performed on both training and test datasets using metrics: accuracy (ACC), sensitivity (SEN) and specificity (SPEC). Again, a figure with a comprehensive summary of all methods is shown, accompanied by a table presenting the evaluation metrics.

Mania

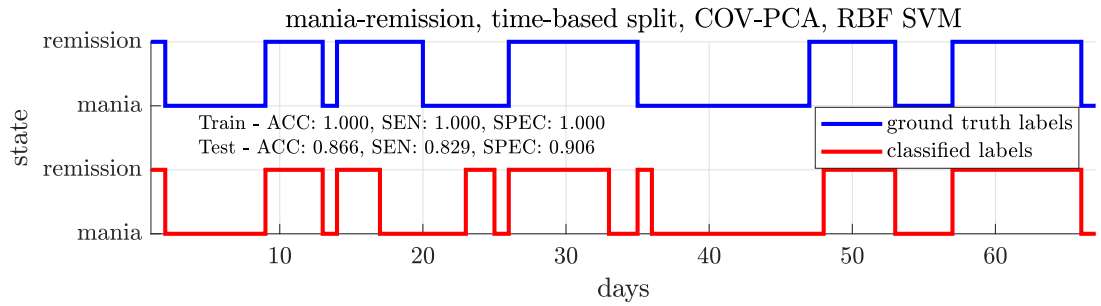


Figure 5.40: Classification of mania-remission using PCA selected COV structures on a time-based split dataset.

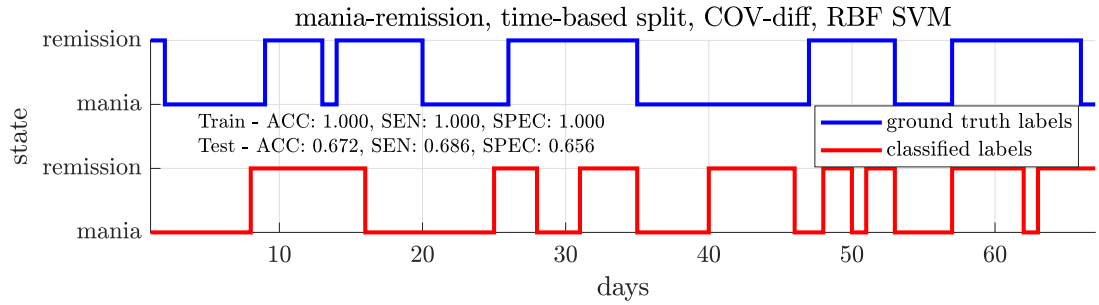


Figure 5.41: Classification of mania-remission using max-diff-based selected COV structures on a time-based split dataset.

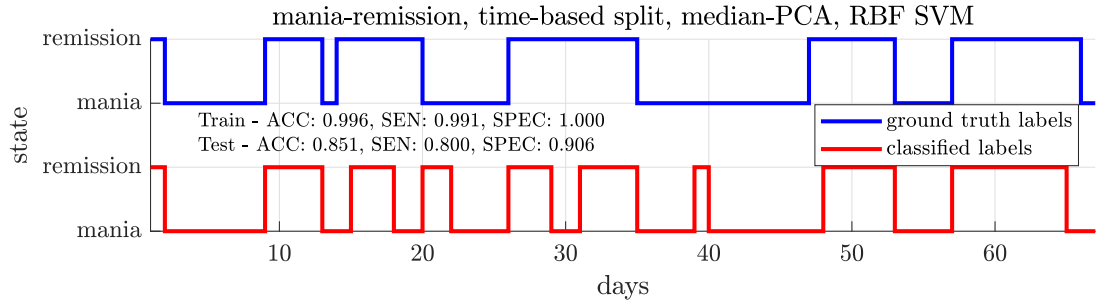


Figure 5.42: Classification of mania-remission using PCA selected median values on a time-based split dataset.

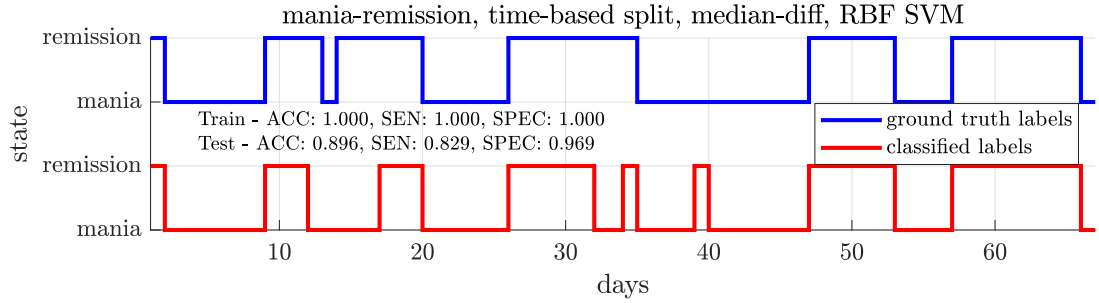


Figure 5.43: Classification of mania-remission using max-diff-based selected median values on a time-based split dataset.

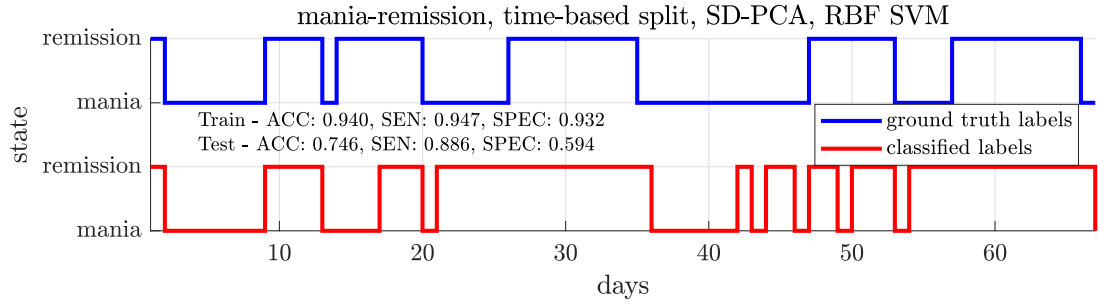


Figure 5.44: Classification of mania-remission using PCA selected SD values on a time-based split dataset.

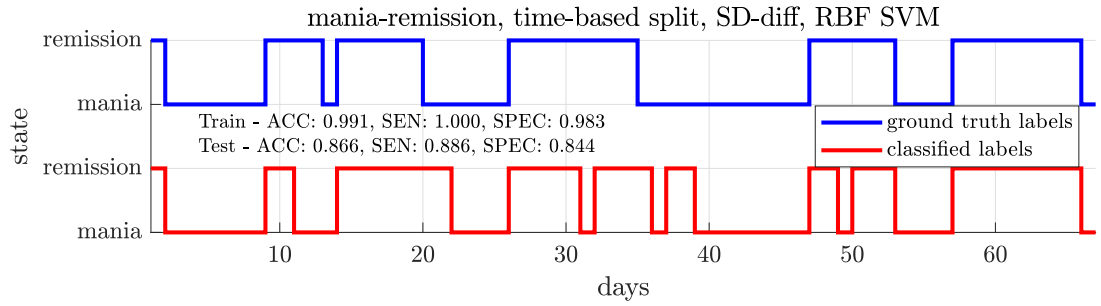


Figure 5.45: Classification of mania-remission using max-diff-based selected SD values on a time-based split dataset.

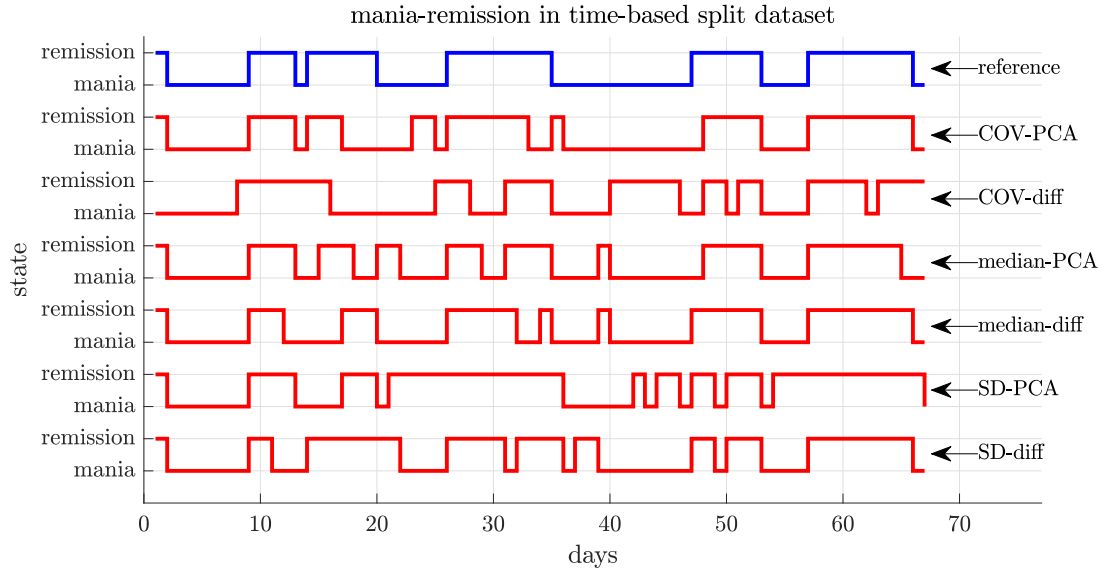


Figure 5.46: Classification overview of all feature selections on mania-remission using all methods on a time-based split dataset.

	Mania-remission (time-based split dataset)											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	86.6%	82.9%	90.6%	100%	100%	100%	67.2%	68.6%	65.6%
Median	99.6%	99.1%	100%	85.1%	80.0%	90.6%	100%	100%	100%	89.6%	82.9%	96.9%
SD	94.0%	94.7%	93.2%	74.2%	88.6%	59.4%	100%	100%	100%	86.6%	88.6%	84.4%

Table 5.13: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a time-based split dataset.

The best classification results for mania state on this dataset were achieved by the median value method with the use of max-diff feature selection with an accuracy of 90%. The COV method with PCA achieved the same accuracy as the SD value method with max-diff dimensionality reduction — 87%. Such high values compared to previous datasets are due to the high dependence

of the training and testing set.

Depression

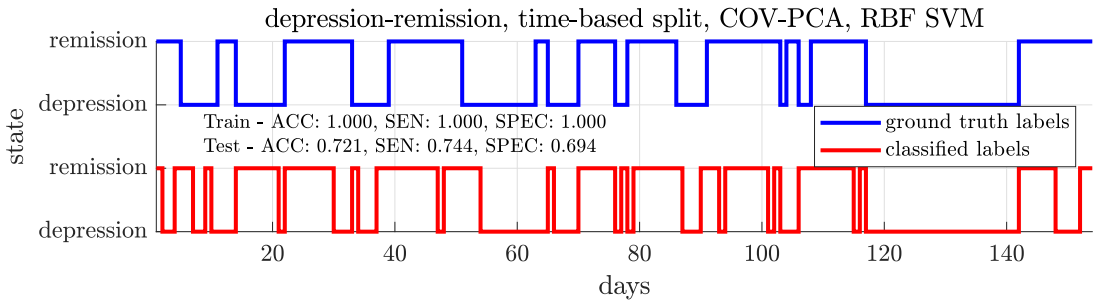


Figure 5.47: Classification of depression-remission using PCA selected COV structures on a time-based split dataset.

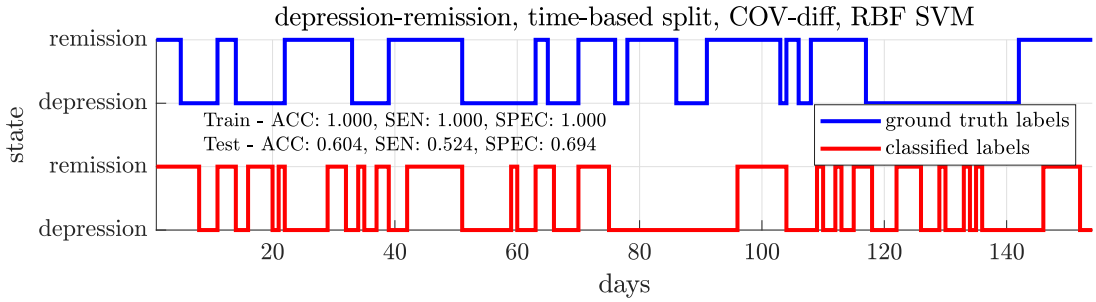


Figure 5.48: Classification of depression-remission using max-diff-based selected COV structures on a time-based split dataset.

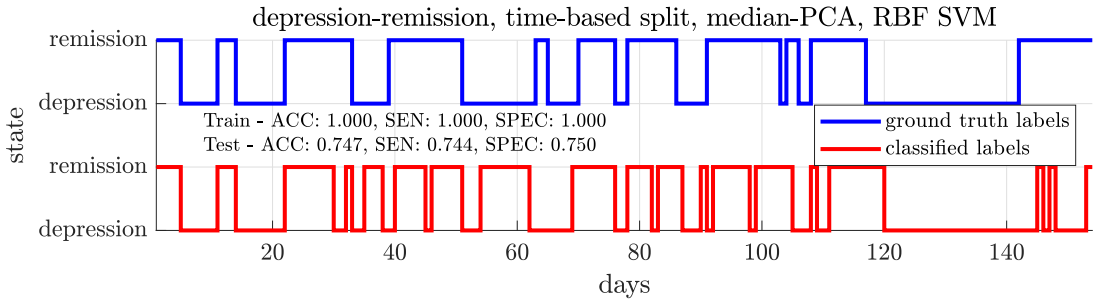


Figure 5.49: Classification of depression-remission using PCA selected median values on a time-based split dataset.

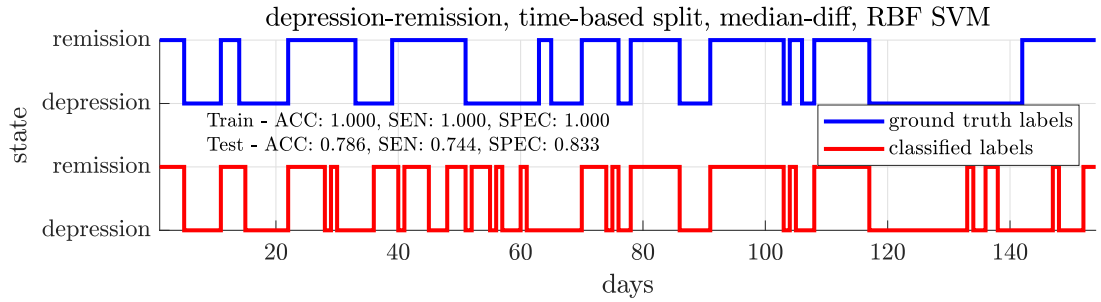


Figure 5.50: Classification of depression-remission using max-diff-based selected median values on a time-based split dataset.

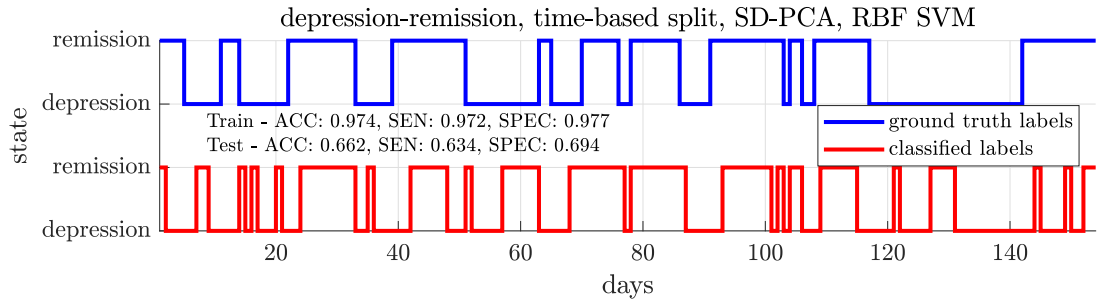


Figure 5.51: Classification of depression-remission using PCA selected SD values on a time-based split dataset.

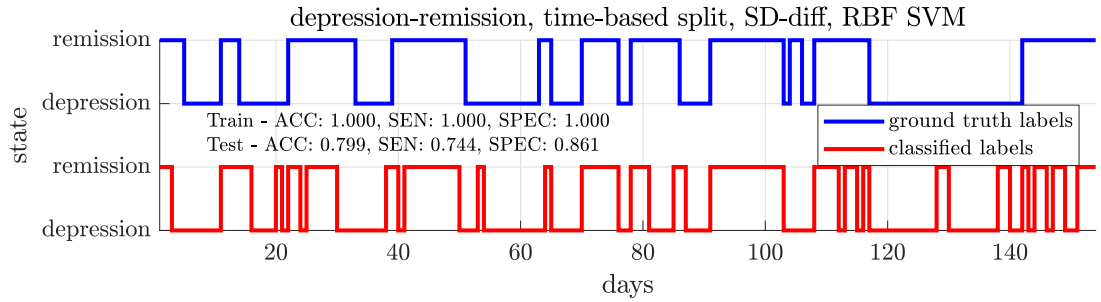


Figure 5.52: Classification of depression-remission using max-diff-based selected SD values on a time-based split dataset.

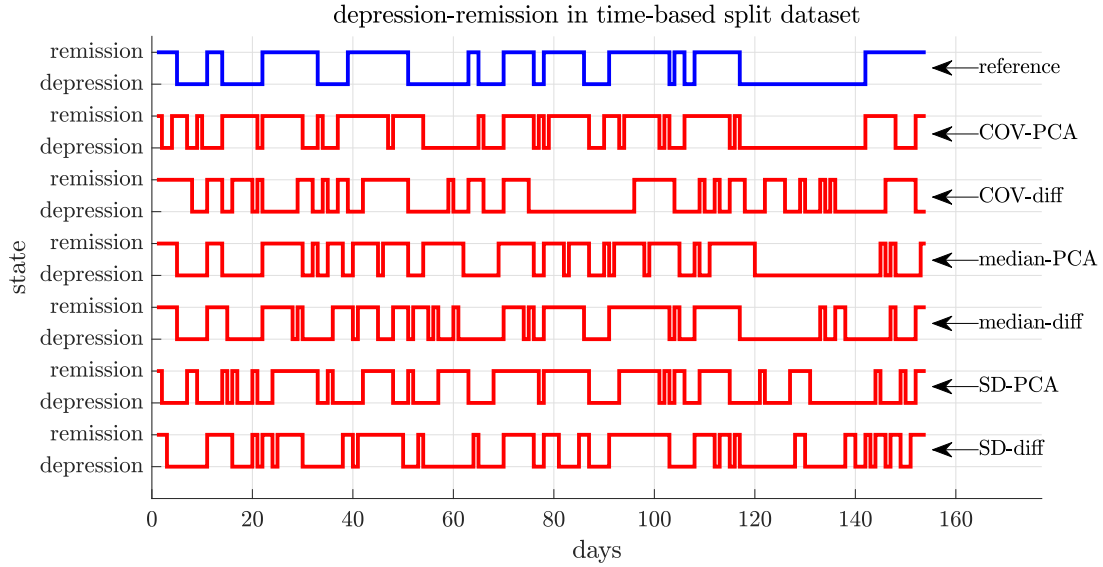


Figure 5.53: Classification overview of all feature selections on mania-remission using all methods on a time-based split dataset.

	Depression-remission (time-based split dataset)											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	72.1%	74.4%	69.4%	100%	100%	100%	60.4%	52.4%	69.4%
Median	100%	100%	100%	74.7%	74.4%	75.0%	100%	100%	100%	78.6%	74.4%	83.3%
SD	97.4%	97.2%	97.7%	66.2%	63.4%	69.4%	100%	100%	100%	79.9%	74.4%	86.1%

Table 5.14: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections on a time-based split dataset.

Even in the case of the classification of the depression state, the values are considerably higher due to the dependence of the training and testing set compared to previous datasets. Despite this, the accuracies are again lower compared to the mania classification. The highest accuracy was achieved by the SD method with max-diff feature selection (80%). The median value method with max-diff feature selection reached 79% accuracy and the COV method using PCA 72% yielded classification accuracy.

5.2.3.4 Results Summary

To summarise all the results obtained by the validation method using a 7-day window on the data across all patients, let us compare the results from the individual methods and datasets. In general, the mania-remission classification had better accuracy than depression-remission on all datasets. On the patient-based split dataset, the mania classification achieved an accuracy of 67%, whereas the depression classification achieved only 59%. For classification on the state-based

dataset, the accuracy of mania-remission was 85% and depression-remission again only 59%. Notably, the accuracy for classification of clinical states on the time-based split dataset whose training and testing sets contained strongly dependent samples was the highest of all, 90% for mania and 80% for depression. Optimal methods on the patient-based split dataset were the SD method with PCA for classification of mania-remission and COV with PCA for depression states. For the other two state-based split and time-based split datasets, the median value method with max-diff selection of attributes was the most successful. The increasing accuracy with data dependence between the training and testing sets suggests that the low classification accuracies are due to large inter-patient variability and also differences in actigraphic values between clinical states.

5.2.4 7-day Window for Single Patient

In this section we present the results of the 7-day window classification for single patient. The results are divided into two groups according to the type of data set distribution and further according to the type of condition classified (mania or depression). Due to the large number of patients and the very similar results between individuals, we present only few patients as examples (id 465 and 528 for mania and id 808 and 864 for depression).

5.2.4.1 State-Based Split Dataset

The following figures show the classification results on different days for the different methods combined with the two feature selection methods with accuracy evaluation on the training and test datasets using ACC, SEN and SPEC. Again, a figure with a comprehensive summary of all methods is shown, accompanied by a table presenting the evaluation metrics for all methods for the split state-based dataset for a single patient.

Mania

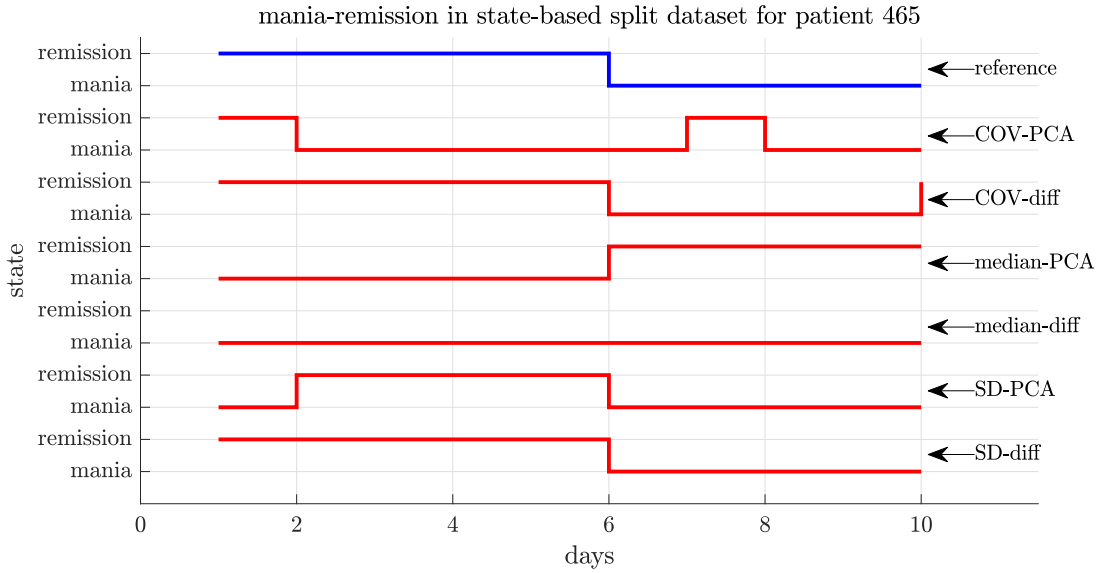


Figure 5.54: Classification overview of all feature selections on mania-remission using all methods for patient 465 on a state-based split dataset.

	Mania-remission (state-based split dataset) patient 465											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	50.0%	20.0%	80.0%	100%	100%	100%	90.0%	100%	80.0%
Median	100%	100%	100%	0.0%	0.0%	0.0%	100%	100%	100%	50.0%	0.0%	100%
SD	100%	100%	100%	90.0%	80.0%	100%	100%	100%	100%	100%	100%	100%

Table 5.15: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 465 on a state-based split dataset.

In the case of patient id 465, the max-diff feature selection achieved the best results for all methods. The highest mania classification accuracy of 100% was obtained by the SD value method.

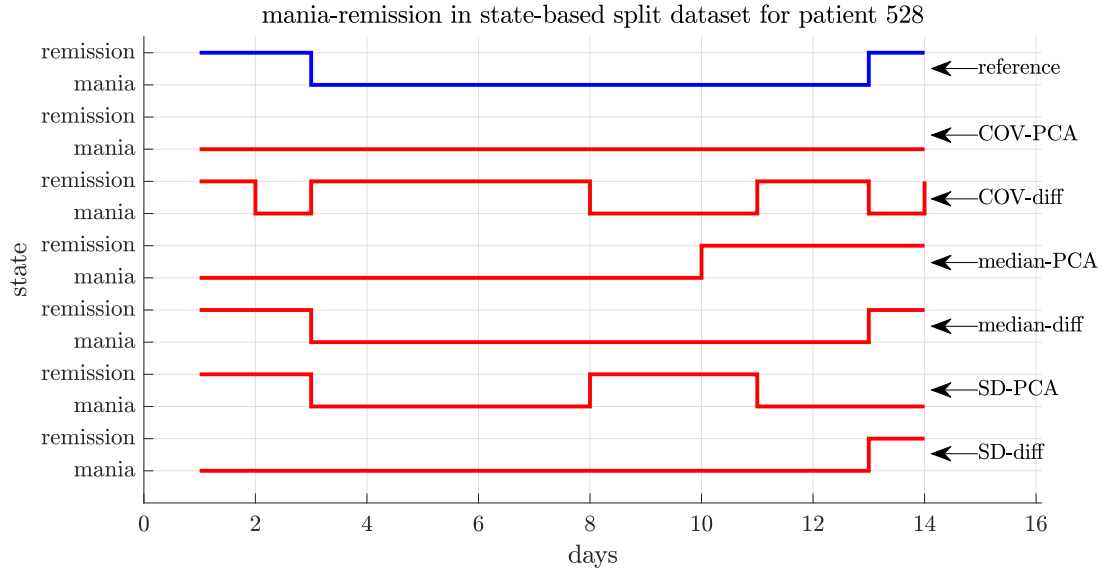


Figure 5.55: Classification overview of all feature selections on mania-remission using all methods for patient 528 on a state-based split dataset.

	Mania-remission (state-based split dataset) patient 528											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	71.4%	0.0%	100%	100%	100%	100%	35.7%	50.0%	30.0%
Median	100%	100%	100%	64.3%	50.0%	70.0%	100%	100%	100%	100%	100%	100%
SD	100%	100%	100%	64.3%	50.0%	70.0%	100%	100%	100%	85.7%	50.0%	100%

Table 5.16: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 528 on a state-based split dataset.

For patient id 528, the best results were obtained using the max-diff selection of feature value median and SD. While the COV method performed better using PCA. The highest mania classification accuracy of 100% was obtained by the max-diff-based selected median value method.

Depression

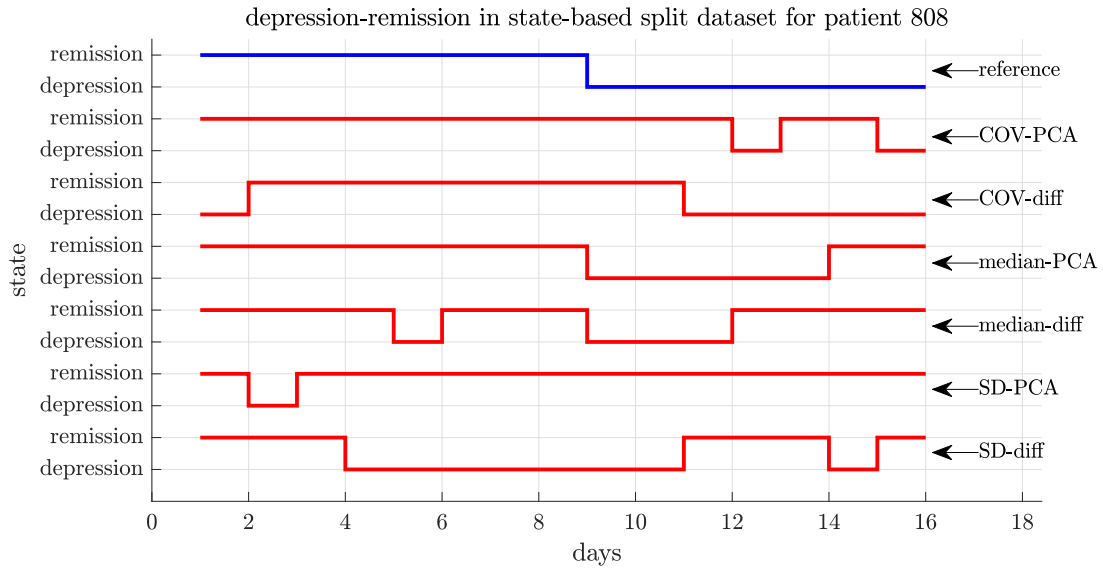


Figure 5.56: Classification overview of all feature selections on depression-remission using all methods for patient 808 on a state-based split dataset.

	Deprssion-remission (state-based split dataset) patient 808											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	68.8%	100%	37.5%	100%	100%	100%	81.2%	87.5%	75.0%
Median	100%	100%	100%	81.2%	100%	62.5%	100%	100%	100%	62.5%	87.5%	37.5%
SD	100%	100%	100%	43.8%	87.5%	0.0%	100%	100%	100%	37.5%	37.5%	37.5%

Table 5.17: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 808 on a state-based split dataset.

In the case of the classification of depression in patient id 808, the best results were achieved only in the case of COV method. While median and SD methods performed better with PCA. The highest overall classification accuracy of 81% was then obtained by both median value methods using PCA and the max-diff dimensionality reduction of COV method.

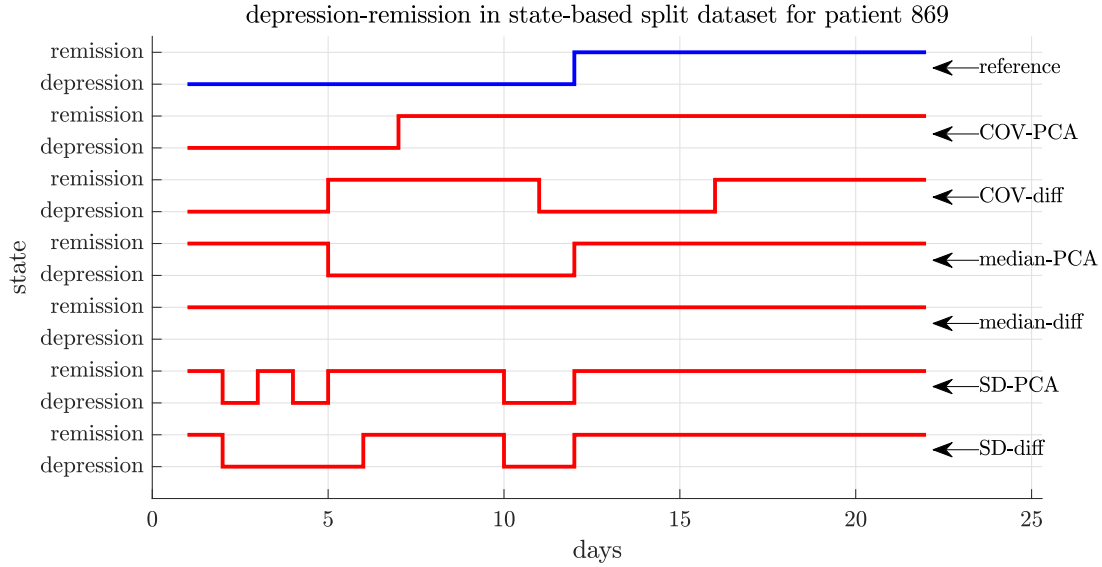


Figure 5.57: Classification overview of all feature selections on depression-remission using all methods for patient 869 on a state-based split dataset.

	Deprssion-remission (state-based split dataset) patient 869											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	77.3%	100%	54.5%	98.5%	100%	97.0%	54.5%	63.6%	45.5%
Median	100%	100%	100%	81.8%	100%	63.6%	100%	100%	100%	50.0%	100%	0.0%
SD	95.5%	94.1%	97.0%	68.2%	100%	36.4%	100%	100%	100%	77.3%	100%	54.5%

Table 5.18: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 869 on a state-based split dataset.

In the case of the classification of depression in patient id 869, PCA feature selection achieved the best results only in the case of the median value of the method. While the COV and SD value methods performed better using the max-diff approach. The median value method using PCA yielded the highest classification accuracy of 82%.

5.2.4.2 Time-Based Split Dataset

In the following figures, we show the classification results on different days for different methods combined with two feature selection methods with accuracy evaluation on training and test datasets using ACC, SEN and SPEC. Once again, a figure with a comprehensive summary of all methods is presented, accompanied by a table that shows the evaluation metrics of all methods for the time-based dataset for a single patient.

Mania

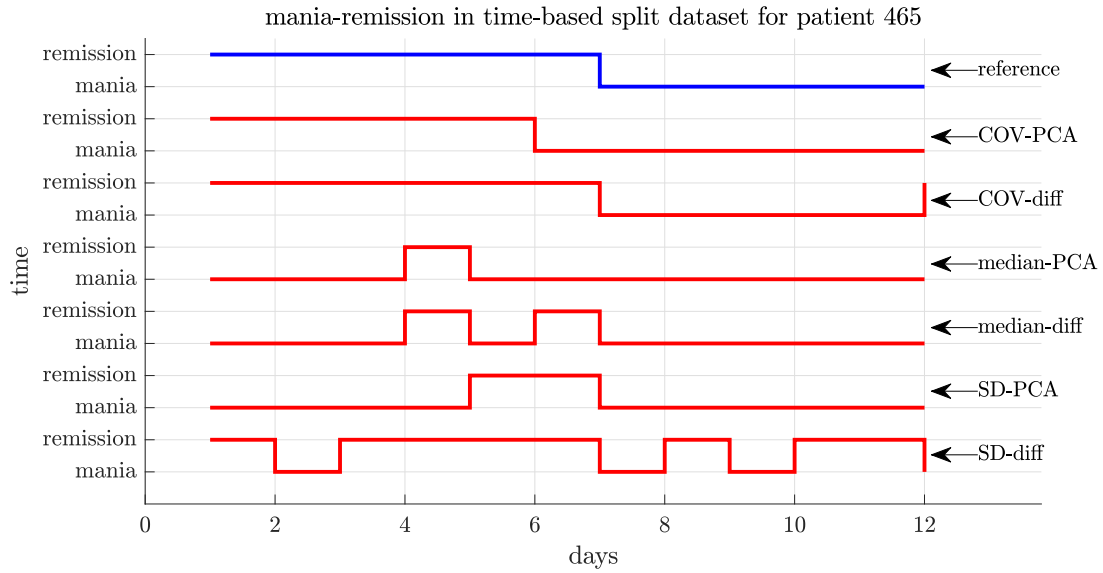


Figure 5.58: Classification overview of all feature selections on mania-remission using all methods for patient 465 on a time-based split dataset.

	Mania-remission (time-based split dataset) patient 465											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	91.7%	83.3%	100%	100%	100%	100%	91.7%	100%	83.3%
Median	100%	100%	100%	58.3%	16.7%	100%	100%	100%	100%	66.7%	33.3%	100%
SD	100%	100%	100%	66.7%	33.3%	100%	100%	100%	100%	66.7%	83.3%	50.0%

Table 5.19: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 465 on a time-based split dataset.

In the case of the classification of mania state in id 465 patient, PCA and max-diff feature selection achieved comparable results. Only for the median value method the use of the dimensionality reduction approach appeared to be more appropriate from an accuracy point of view. The highest classification accuracy of 92% was then achieved by the COV method in general.

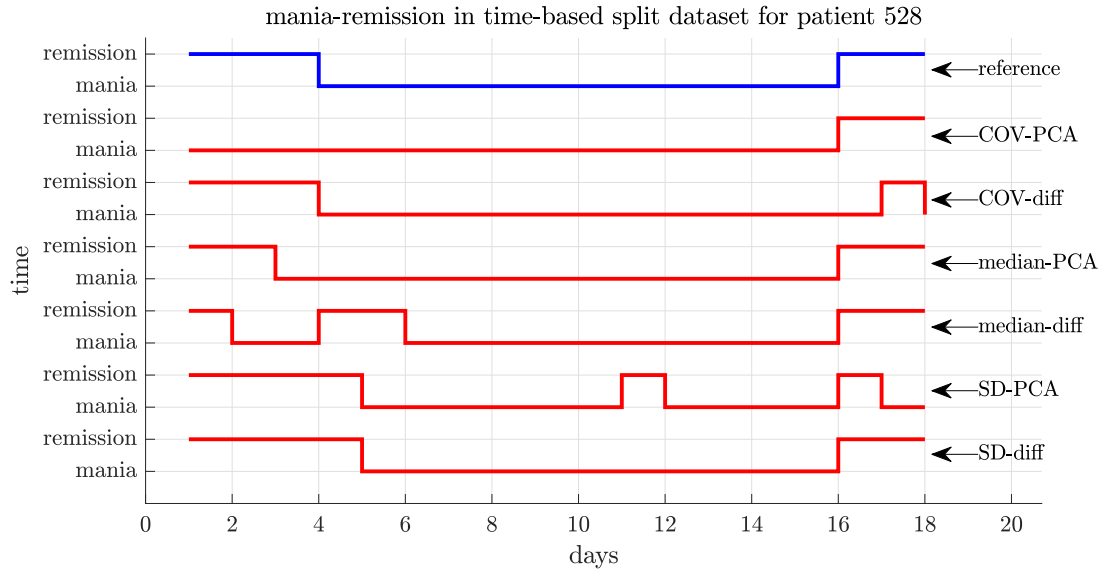


Figure 5.59: Classification overview of all feature selections on mania-remission using all methods for patient 528 on a time-based split dataset.

	Mania-remission (time-based split dataset) patient 528											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	83.3%	50.0%	100%	100%	100%	100%	88.9%	66.7%	100%
Median	100%	100%	100%	94.4%	83.3%	100%	100%	100%	100%	77.8%	66.7%	83.3%
SD	96.2%	100%	93.1%	77.8%	66.7%	83.3%	100%	100%	100%	94.4%	100%	91.7%

Table 5.20: Performance metrics of classification mania-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 528 on a time-based split dataset.

In the case of classification of mania state in patient id 528, the median value of the PCA method of feature selection achieved better results than the max-diff approach. On the other hand, for the COV method and the SD value method, the use of the max-diff dimensionality reduction appeared to be more appropriate in terms of accuracy. The highest classification accuracy 95% for this patient was obtained by both median value with PCA and SD value with max-diff selection.

Depression

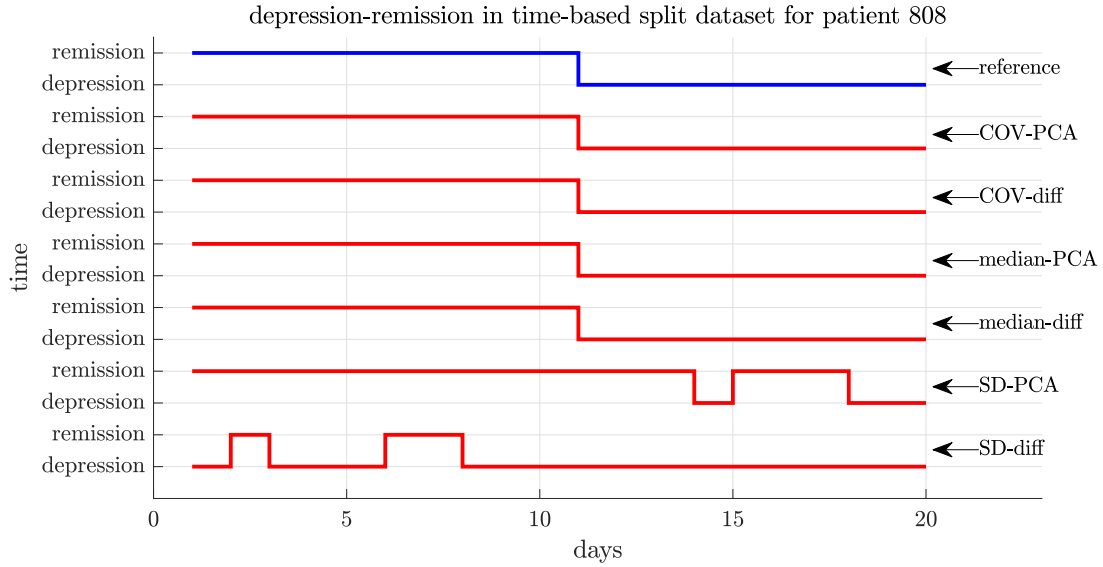


Figure 5.60: Classification overview of all feature selections on depression-remission using all methods for patient 808 on a time-based split dataset.

	Depression-remission (time-based split dataset) patient 808											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Median	100%	100%	100%	100%	100%	100%	100%	100%	100%	92.3%	84.6%	100%
SD	100%	100%	100%	70.0%	100%	40.0%	100%	100%	100%	65.0%	30.0%	100%

Table 5.21: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 808 on a time-based split dataset.

In the case of depression classification in patient with id 808, the COV method achieved the same results with both PCA and max-diff approach. For the median and SD value methods, the use of PCA for dimensionality reduction seemed to be more appropriate from the accuracy point of view. The highest possible classification accuracy of 100% for this particular patient was achieved by both COV methods with both dimensionality reduction approaches and median value with PCA.

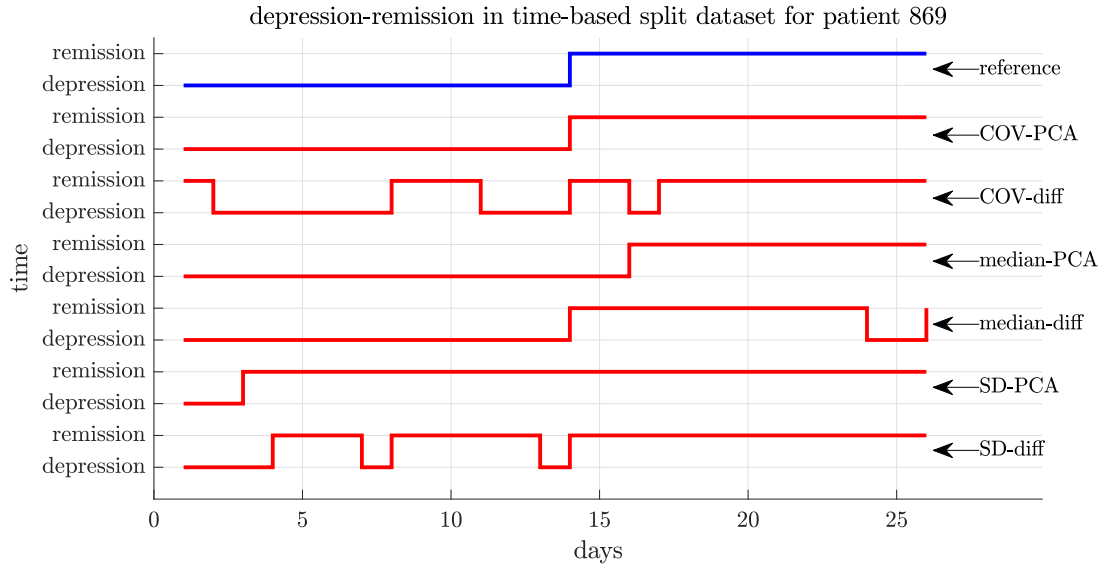


Figure 5.61: Classification overview of all feature selections on depression-remission using all methods for patient 869 on a time-based split dataset.

	Depression-remission (time-based split dataset) patient 869											
	PCA method						**max diff pairs method**					
	Train set			**Test set**			**Train set**			**Test set**		
Method	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*	*ACC*	*SEN*	*SPEC*
COV	100%	100%	100%	100%	100%	100%	100%	100%	100%	80.8%	92.3%	69.2%
Median	95.2%	90.6%	100%	92.3%	84.6%	100%	100%	100%	100%	100%	100%	100%
SD	95.2%	96.9%	93.5%	57.7%	100%	15.4%	100%	100%	100%	69.2%	100%	38.5%

Table 5.22: Performance metrics of classification depression-remission: accuracy (ACC), sensitivity (SEN) and specificity (SPEC) for all methods using all feature selections for patient id 86 on a time-based split dataset.

In the case of depression classification in patient id 869, the COV method achieved better results with the use of PCA. On the other hand, for the median and SD methods, the use of PCA for dimensionality reduction seemed to be more appropriate in terms of accuracy. The highest possible classification accuracy of 100% for this particular patient was achieved by both COV with PCA and median value with max-diff selection.

5.2.4.3 Results Summary

For the summary of the performance of the validation method using a 7-day window on individual patient data, let us compare the results of the used methods and datasets. In general, as in the previous analogous validation method over all patients, the mania-remission classification had higher accuracy than the depression-remission on all datasets used. For the classification of selected individual patients on the state-based dataset, the accuracy of mania-remission and depression-remission was 100% and slightly lower at 81% and 82%, respectively. Because of the strongly dependent samples in the time-based split dataset, the accuracy of classification of clinical states of individuals was higher only for depression 100%, whereas for mania, this accuracy slightly decreased to 92%, respectively 95%. The best performing method differed depending on the classified patient and the selected dataset, thus it is not possible to single out a specific one. The expected higher accuracy with data dependency between training and testing datasets was only observed for the mania-remission classification, not for depression. This may be due to the small number of samples for the same person and therefore very similar dependencies.

Chapter 6

Discussion

The initial aim of our research focused on established correlations between actigraphy-based activity measures (specifically activity levels and sleep patterns) — and the clinical status of individuals with BD. Our chosen novel approach was to investigate whether the covariance structure among various actigraphy features underwent changes across different clinical states. To achieve this our initial idea was to apply time series clustering using TICC algorithm. However, it is important to note certain limitations of the TICC method, such as it is unsupervised approach and dependence on a substantial number of data points. Given the daily granularity of our dataset and the brevity of clinical states (e.g. median duration of mania episodes in our dataset was 21 days and depression 25 days), these limitations required us to change our analytical framework.

Due to the above limitations and considering the availability of clinical state labels, the logical step in our methodology was to move to a supervised scenario. This transition facilitated a more fine-grained and structured exploration of the complex relationship between the covariance structure of actigraphy parameters and clinical conditions. We added methods based on the median and SD value of actigraphy attributions in parallel to the COV methods. In addition, we introduced dimension reduction approaches based on PCA and max-diff method based on the maximal difference of the feature pair. We applied these methods on a clinician-labelled days with actigraphy data from 92 BD patients. Unfortunately, only 18 patients had epoch of clinical depression and remission at the same time, and (patients fulfilled the same criteria for mania.

Each of these three methods was used with one of two methods of reducing the dimensionality of the symptom space - using the well-known PCA and using our proposed max-diff-based method based on the maximal difference of feature pairs between clinical states. To evaluate these methods, we used three validation scenarios: the leave-one-patient out cross-validation method — distinguishing the entire clinical states of mania-remission, respectively depression-remission, and two methods classifying time-series in a 7-day window across all patients and for individuals. For this evaluation, three methods of creating training and testing sets were used: patinet, state and time-based split, with the aim of creating independent, partially and maximally dependent datasets to illustrate the upper and lower bounds for performance of the

investigated methods.

The results of the supervised method show that distinguishing mania-remission is easier than depression-remission. Another finding is that the variability between patients is quite high. Using leave-one-patient-out cross-validation, an average accuracy of 72% for the classification of mania-remission and 67% for the classification of depression-remission was achieved in both cases using the SD value method with max-diff-based dimensionality reduction. Classification on patient-based split data using the 7-day window across all patients achieved 67% accuracy in the case of mania using the SD method and 59% accuracy in the case of depression using the SD value method, in both cases with PCA features selection. On the state-based split dataset, significantly better results of 85% were achieved in the case of mania, and the same accuracy of 59% was achieved in the case of depression as in the previous dataset. Each time using the method of median values with max-diff-based feature selection. For the time-based split dataset, the most dependent one, the best results were achieved, as expected. For the classification of mania 90% accuracy was obtained and 80% for depression, using the median value method with max-diff-based feature selection. For the last validation scenario for a 7-day window for each patient, 100% accuracy was achieved in the case of mania and 81%, 82%, respectively, in the presented patients on the state-based split data. For time-based split data, depression classification improved to 100%, while mania decreased to 92% and 95% respectively. This is due to the very small number of sample per patient and thus the small difference between the data splitting methods.

Considering our awareness of overfitting in supervised methods, it is crucial to comment on the delicate balance between model complexity and generalisation. Despite our efforts to introduce constraints and prevent overfitting on the classification set by tuning hyperparameters and exploring different kernel functions, the results led to a significant deterioration in classification accuracies even on the test set. Exploring these challenges contributes to a more comprehensive understanding of the limitations and underscores the complex nature of the dataset and the task underscores the complex nature of the dataset and the task.

6.1 Limitations

Results of this study need to be interpreted considering the following limitations: One of the main limitations arises from the composition of the sample and validation sets. The data used in this study includes a relatively limited number of samples, both in terms of patients and days. The rarity of these samples is further exacerbated by the overall size of the original data set. It is important to note that the scarcity of the data is attributed to the low frequency of manic and depressive states in BD, which, despite their lower occurrence, should not undermine their severity. Another limitation is the low sampling frequency of actigraphic features and their annotations since only one sample per day was obtained.

6.2 Future Work

For future work, the following main steps are to:

- Use supervised principal component analysis (sPCA) method to incorporate clinical state label information into PCA to increase the usefulness of the extracted features for this classification task.
- Include all removed clinical states (onsets and offsets of mania and depression) and convert the classification task into a multi-class one.
- Include healthy control subjects in this study.

Chapter 7

Conclusion

The aim of this work was to investigate the relationships between clinical status measured by clinical self-report scales and actigraphic variables such as changes in sleep and physical activity. Previous studies have shown significant but heterogeneous associations between actigraphic parameters and clinical status. Therefore, we decided to focus on changes in the structures of relationships between multiple actigraphic parameters during changes in clinical status. The original idea was to use a very promising unsupervised approach to multivariate time series clustering based on Toeplitz inverse covariance matrices [1]. However, by deeper investigation and evaluation with suitable criteria on simulated data led us to the conclusion, it did not appear to be a suitable method for our real data. Therefore, we implemented three supervised methods using the SVM classifier beyond the scope of the assignment. The first method based on the structures of covariance matrices (COV), and the other two based on the median and SD values of actigraphic features.

The applied supervised methods showed the classification of clinical states based on multivariate actigraphy is feasible, however, the classification accuracy was relatively low, especially in strict validation scenarios, where the models were applied to completely unseen patients. This implicates that some level of patient-based individualisation or other approaches using per-patient models should be investigated.

The approach based on the covariance structure showed higher performance when coupled with the PCA, rather than with the automatic maximally differing feature pairs - the latter being better performing when coupled with the feature SD or median values. Also, contrary to our expectations, the covariance structure-based method did not outperform the variance (SD) or value (median)-based multivariate methods.

To conclude, while the variability of activity markers has previously been shown to be connected with clinical worsenings in BD, finding appropriate multivariate methods poses a challenge, especially with respect to the difficulty of obtaining large-enough dataset on clinical population.

Bibliography

[1] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data", *Association for Computing Machinery*, pp. 215–223, 2017, ISSN: 10450823. DOI: 10.1145/3097983.3098060. arXiv: 1706.03161. [Online]. Available: https://doi.org/10.1145/3097983.3098060.

[2] GBD, *Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019)*, 2021. DOI: https://doi.org/10.6069/1D4Y-YQ37. [Online]. Available: https://vizhub.healthdata.org/gbd-results?params=gbd-api-2019-permalink/d370406f3a8ee21fbd20e0c8cb0cfaf4.

[3] APA, *Diagnostic and statistical manual of mental disorders : DSM-5, American Psychiatric Association. Arlington, VA Washington, D.C.: American Psychiatric Association.* 2013.

[4] K. Látalová, *BIPOLÁRNÍ AFEKTIVNÍ PORUCHA*, 1st ed. Grada Publishing, a.s., 2010, p. 256, ISBN: 978-80-247-7390-2.

[5] G. Murray and A. Harvey, "Circadian rhythms and sleep in bipolar disorder", *Bipolar Disorders*, vol. 12, no. 5, pp. 459–472, 2010, ISSN: 13985647. DOI: 10.1111/j.1399-5618.2010.00843.x.

[6] L. B. Alloy, T. H. Ng, M. K. Titone, and E. M. Boland, "Circadian Rhythm Dysregulation in Bipolar Spectrum Disorders", *Current Psychiatry Reports*, vol. 19, no. 4, 2017, ISSN: 15351645. DOI: 10.1007/s11920-017-0772-z.

[7] A. Millar, C. A. Espie, and J. Scott, "The sleep of remitted bipolar outpatients: A controlled naturalistic study using actigraphy", *Journal of Affective Disorders*, vol. 80, no. 2-3, pp. 145–153, 2004, ISSN: 01650327. DOI: 10.1016/S0165-0327(03)00055-7.

[8] P. A. Geoffroy, C. Boudebesse, F. Bellivier, *et al.*, "Sleep in remitted bipolar disorder: A naturalistic case-control study using actigraphy", *Journal of Affective Disorders*, vol. 158, pp. 1–7, 2014, ISSN: 01650327. DOI: 10.1016/j.jad.2014.01.012. [Online]. Available: http://dx.doi.org/10.1016/j.jad.2014.01.012.

[9] F. Bellivier, P. A. Geoffroy, B. Etain, and J. Scott, "Sleep- and circadian rhythm-associated pathways as therapeutic targets in bipolar disorder", *Expert Opinion on Therapeutic Targets*, vol. 19, no. 6, pp. 747–763, 2015, ISSN: 17447631. DOI: 10.1517/14728222.2015.1018822.

[10] A. K. Gold and L. G. Sylvia, "The role of sleep in bipolar disorder", *Nature and Science of Sleep*, vol. 8, pp. 207–214, 2016, ISSN: 11791608. DOI: 10.2147/NSS.S85754.

[11] G. Scola, H. K. Kim, L. T. Young, and A. C. Andreazza, "A fresh look at complex i in microarray data: Clues to understanding disease-specific mitochondrial alterations in bipolar disorder", *Biological Psychiatry*, vol. 73, no. 2, e4, 2013, ISSN: 18732402. DOI: 10.1016/j.biopsych.2012.06.028. [Online]. Available: http://dx.doi.org/10.1016/j.biopsych.2012.06.028.

[12] A. C. Andreazza, A. Duong, L. T. Young, and L. T, "Bipolar disorder as a mitochondrial disease", *Biological Psychiatry*, vol. 83, no. 9, pp. 720–721, 2018, ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2017.09.018. [Online]. Available: https://doi.org/10.1016/j.biopsych.2017.09.018.

[13] WHO, "Global challenge for movement on mental health kicks off as lack of investment in mental health leaves millions without access to services", Joint release by the World Health Organization, United for Global Mental Health and the World Federation for Mental Health, Tech. Rep., 2020. [Online]. Available: https://www.who.int/news/item/07-10-2020-global-challenge-for-movement-on-mental-health-kicks-off-as-lack-of-investment-in-mental-health-leaves-millions-without-access-to-services.

[14] J. N. Miller and D. W. Black, "Bipolar Disorder and Suicide : a Review Bipolar Disorder and Suicide : a Review", *Springer*, vol. 22, no. 6, pp. 0–10, 2020. DOI: 10.1007/s11920-020-1130-0.

[15] A. Schaffer, E. T. Isometsä, L. Tondo, *et al.*, "Epidemiology, neurobiology and pharmacological interventions related to suicide deaths and suicide attempts in bipolar disorder: Part I of a report of the International Society for Bipolar Disorders Task Force on Suicide in Bipolar Disorder", *Australian and New Zealand Journal of Psychiatry*, vol. 49, no. 9, pp. 785–802, 2015, ISSN: 14401614. DOI: 10.1177/0004867415594427.

[16] F. K. Goodwin and K. R. Jamison, *Manic-depressive illness*. Oxford University Press, 1990.

[17] R. S. McIntyre, M. Berk, E. Brietzke, *et al.*, "Bipolar disorders", *The Lancet*, vol. 396, no. 10265, pp. 1841–1856, 2020, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(20)31544-0. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(20)31544-0.

[18] T. P. Hui, A Kandola, L Shen, G Lewis, D. P. J. Osborn, and J. R. Geddes, "A systematic review and meta-analysis of clinical predictors of lithium response in bipolar disorder", *Acta Psychiatr Scand*, vol. 140, pp. 94–115, 2019. DOI: 10.1111/acps.13062.

[19] IQWiG, *Cognitive behavioral therapy*, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279297/.

[20] K. Özdel, A. Kart, and M. H. Türkçapar, "Cognitive Behavioral Therapy in Treatment of Bipolar Disorder", *Noro Psikiyatr Ars*, vol. 58, no. Supplement 1, pp. 66–76, 2021. DOI: 10.29399/npa.27419..

[21] S. A. Montgomery and M. Asberg, "A new depression scale designed to be sensitive to change", *British Journal of Psychiatry*, vol. 134, no. 4, pp. 382–389, 1979, ISSN: 00071250. DOI: 10.1192/bjp.134.4.382.

[22] M. Hamilton, "A rating scale for depression", *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, pp. 56–62, 1960. DOI: https://doi.org/10.1136/jnnp.23.1.56.

[23] A. J. Rush, C. M. Gullion, M. R. Basco, R. B. Jarrett, and M. H. Trivedi, "The Inventory of Depressive Symptomatology (IDS): psychometric properties", *Psychological medicine*, vol. 26, no. 3, pp. 477–486, 1996. DOI: https://doi.org/10.1017/s0033291700035558.

[24] M. Berk, G. S. Malhi, C. Cahill, *et al.*, "The Bipolar Depression Rating Scale (BDRS): its development , validation and utility", *Bipolar Disorders*, vol. 9, no. 6, pp. 571–579, 2007. DOI: 10.1111/j.1399-5618.2007.00536.x.

[25] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, "A Rating Scale for Mania : Reliability , Validity and Sensitivity", *The British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978. DOI: 10.1192/bjp.133.5.429.

[26] P. Bech, O. J. Rafaelsen, P. Kramp, and T. G. Bolwig, "The mania rating scale: Scale construction and inter-observer agreement", *Neuropharmacology*, vol. 17, no. 6, pp. 430–431, 1978. DOI: 10.1016/0028-3908(78)90022-9.

[27] E. G. Altman, D. R. Hedeker, P. G. Janicak, J. L. Peterson, and J. M. Davis, "The clinician-administered rating scale for mania (CARS-M): Development, reliability, and validity", *Biological Psychiatry*, vol. 36, no. 2, pp. 124–134, 1994. DOI: 10.1016/0006-3223(94)91193-2.

[28] S. Krüger, L. Quilty, M. Bagby, T. Lippold, F. Bermpohl, and P. Bräunig, "The Observer-Rated Scale for Mania (ORSM): development , psychometric properties and utility", *Journal of Affective Disorders*, vol. 122, no. 1-2, pp. 179–183, 2010, ISSN: 0165-0327. DOI: 10.1016/j.jad.2009.07.022. [Online]. Available: http://dx.doi.org/10.1016/j.jad.2009.07.022.

[29] K. D. Denicoff, G. S. Leverich, W. A. Nolen, *et al.*, "Validation of the prospective NIMH-Life-Chart Method (NIMH-LCMTM-p) for longitudinal assessment of bipolar illness", *Psychological Medicine*, vol. 30, no. 6, pp. 1391–1397, 2000. DOI: 10.1017/s0033291799002810.

[30] G. S. Sachs, C Guille, and S. L. McMurrich, "A clinical monitoring form for mood disorders", *Bipolar Disorders*, vol. 4, no. 5, pp. 323–327, 2002. DOI: 10.1034/j.1399-5618.2002.01195.x.

[31] E. B. Dennehy, T. Suppes, M. L. Crismon, M. Toprac, T. J. Carmody, and A. J. Rush, "Development of the Brief Bipolar Disorder Symptom Scale for patients with bipolar disorder", *Psychiatry Research*, vol. 127, no. 1-2, pp. 137–145, 2004. DOI: 10.1016/j.psychres.2004.02.009.

[32] J. M. Gonzalez, C. L. Bowden, M. M. Katz, *et al.*, "Development of the Bipolar Inventory of Symptoms Scale : concurrent validity , discriminant validity and retest reliability", *International Journal of Methods in Psychiatric Research*, vol. 17, no. 4, pp. 198–209, 2008. DOI: 10.1002/mpr.262.

[33] C. Acebo, A. Sadeh, R. Seifer, *et al.*, "Estimating sleep patterns with activity monitoring in children and adolescents: How many nights are necessary for reliable measures?", *Sleep*, vol. 22, no. 1, pp. 95–103, 1999, ISSN: 01618105. DOI: https://doi.org/10.1093/sleep/22.1.95Tt.

[34] W. W. Tryon, *Activity Measurement in Psychology and Medicine*. Plenum Press, 1991. DOI: https://doi.org/10.1007/978-1-4757-9003-0.

[35] A. Sadeh and C. Acebo, "The role of actigraphy in sleep medicine", *Sleep Medicine Reviews*, vol. 6, no. 2, pp. 113–124, 2002, ISSN: 10870792. DOI: 10.1053/smrv.2001.0182.

[36] M. T. Smith, C. S. McCrae, J. Cheung, *et al.*, "Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Clinical Practice Guideline", *Journal of Clinical Sleep Medicine*, vol. 14, no. 7, pp. 1231–1237, 2018, ISSN: 15509397. DOI: 10.5664/jcsm.7702.

[37] M. Ocisková and J. Praško, *Stigmatizace a sebestigmatizace u psychických poruch*. 2015, ISBN: 978-80-247-5199-3.

[38] A. Sadeh, K. M. Sharkey, and M. A. Carskadon, "Activity-based sleep-wake identification: An empirical test of methodological issues", *Sleep*, vol. 17, no. 3, pp. 201–207, 1994, ISSN: 01618105. DOI: 10.1093/sleep/17.3.201.

[39] C Violani, P Testa, and M Casagrande, "Actigraphic motor asymmetries during sleep", *Sleep*, vol. 21, no. 5, pp. 472–6, 1998.

[40] S. Thomée, "Mobile phone use and mental health. A review of the research that takes a psychological perspective on exposure", *International Journal of Environmental Research and Public Health*, vol. 15, no. 12, 2018, ISSN: 16604601. DOI: 10.3390/ijerph15122692.

[41] K. Krane-Gartiser, T. E. G. Henriksen, G. Morken, A. Vaaler, and O. B. Fasmer, "Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder", *PLoS ONE*, vol. 9, no. 2, 2014, ISSN: 19326203. DOI: 10.1371/journal.pone.0089574.

[42] J. Schneider, "Long-Term Actigraphy in Bipolar Disorder : Processing , Analysis , and Applications", Ph.D. dissertation, CZECH TECHNICAL UNIVERSITY IN PRAGUE, 2021.

[43] J. Schneider, E. Bakštein, M. Kolenič, *et al.*, "Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls", *CNS Spectrums*, vol. 27, no. 1, pp. 82–92, 2022, ISSN: 10928529. DOI: 10.1017/S1092852920001777.

[44] Y. Tazawa, M. Wada, Y. Mitsukura, *et al.*, "Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis", *Journal of Affective Disorders*, vol. 253, no. April, pp. 257–269, 2019, ISSN: 15732517. DOI: 10.1016/j.jad.2019.04.087. [Online]. Available: https://doi.org/10.1016/j.jad.2019.04.087.

[45] C. A. Janney, A. Fagiolini, H. A. Swartz, J. M. Jakicic, R. G. Holleman, and C. R. Richardson, "Are adults with bipolar disorder active? Objectively measured physical activity and sedentary behavior using accelerometry", *Journal of Affective Disorders*, vol. 152-154, no. 1, pp. 498–504, 2014, ISSN: 01650327. DOI: 10.1016/j.jad.2013.09.009. [Online]. Available: http://dx.doi.org/10.1016/j.jad.2013.09.009.

[46] B. S. Gonçalves, T. Adamowicz, F. M. Louzada, C. R. Moreno, and J. F. Araujo, "A fresh look at the use of nonparametric analysis in actimetry", *Sleep Medicine Reviews*, vol. 20, pp. 84–91, 2015, ISSN: 15322955. DOI: 10.1016/j.smrv.2014.06.002. [Online]. Available: http://dx.doi.org/10.1016/j.smrv.2014.06.002.

[47] C. N. Kaufmann, A. Gershon, C. A. Depp, S. Miller, J. M. Zeitzer, and T. A. Ketter, "Daytime midpoint as a digital biomarker for chronotype in bipolar disorder", *Journal of Affective Disorders*, vol. 241, pp. 586–591, 2018, ISSN: 15732517. DOI: 10.1016/j.jad.2018.08.032. [Online]. Available: https://doi.org/10.1016/j.jad.2018.08.032.

[48] J. Scott, "Clinical parameters of circadian rhythms in affective disorders", *European Neuropsychopharmacology*, vol. 21, no. SUPPL.4, S671–S675, 2011, ISSN: 0924977X. DOI: 10.1016/j.euroneuro.2011.07.006. [Online]. Available: http://dx.doi.org/10.1016/j.euroneuro.2011.07.006.

[49] P. S. Ritter, C. Marx, N. Lewtschenko, *et al.*, "The characteristics of sleep in patients with manifest bipolar disorder, subjects at high risk of developing the disease and healthy controls", *Journal of Neural Transmission*, vol. 119, no. 10, pp. 1173–1184, 2012, ISSN: 03009564. DOI: 10.1007/s00702-012-0883-y.

[50] K. A. Kaplan, L. S. Talbot, J. Gruber, and A. G. Harvey, "Evaluating sleep in bipolar disorder: Comparison between actigraphy, polysomnography, and sleep diary", *Bipolar Disorders*, vol. 14, no. 8, pp. 870–879, 2012, ISSN: 13985647. DOI: 10.1111/bdi.12021.

[51] J. St-Amand, M. D. Provencher, L. Bélanger, and C. M. Morin, "Sleep disturbances in bipolar disorder during remission", *Journal of Affective Disorders*, vol. 146, no. 1, pp. 112–119, 2013, ISSN: 01650327. DOI: 10.1016/j.jad.2012.05.057.

[52] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings", *Knowledge and Information Systems*, vol. 19, no. 3, pp. 361–394, 2009, ISSN: 02193116. DOI: 10.1007/s10115-008-0150-6.

[53] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A Review of Subsequence Time Series Clustering", *The Scientific World Journal*, vol. 2014, pp. 1–19, 2014. DOI: 10.1155/2014/312521.

[54] J. Alneberg, B. S. Bjarnason, I. De Bruijn, *et al.*, "Binning metagenomic contigs by coverage and composition", *Nature Methods*, vol. 11, no. 11, pp. 1144–1146, 2014. DOI: 10.1038/nmeth.3103.

[55] D. Herath, S. L. Tang, K. Tandon, D. Ackland, and S. K. Halgamuge, "CoMet: A workflow using contig coverage and composition for binning a metagenomic sample with high precision", *BMC Bioinformatics*, vol. 18, no. Suppl 16, 2017, ISSN: 14712105. DOI: 10.1186/s12859-017-1967-3.

[56] Y. Wang, K. Wang, Y. Y. Lu, and F. Sun, "Improving contig binning of metagenomic data using d2S oligonucleotide frequency dissimilarity", *BMC Bioinformatics*, vol. 18, no. 425, pp. 1–14, 2017, ISSN: 14712105. DOI: 10.1186/s12859-017-1835-1.

[57] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, pp. 273–297, 1995.

[58] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016, ISSN: 1364503X. DOI: 10.1098/rsta.2015.0202.

[59] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, ISBN: 978-0387-31073-2. DOI: 10.4324/9780203733332.

[60] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data: Data Mining, Inference, and Prediction*, Second. Springer, 2009, vol. 26. [Online]. Available: http://www-stat.stanford.edu/$\sim$tibs/ElemStatLearn/.

[61] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation", *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pp. 57–64, 2005. [Online]. Available: https://proceedings.mlr.press/r5/chapelle05b.html.

[62] MATLAB, *(version 9.12.0.2009381 R2022a) Update 4*, Natick, Massachusetts, 2022.

[63] J. Anýž, E. Bakštein, A. Dally, *et al.*, "Validity of the aktibipo self-rating questionnaire for the digital self-assessment of mood and relapse detection in patients with bipolar disorder: Instrument validation study", *JMIR Mental Health*, vol. 8, no. 8, pp. 1–17, 2021, ISSN: 23687959. DOI: 10.2196/26348.

Supplementary

A Detailed Results of Leave-One-Patient-Out Cross-Validation Method

						COV			median			SD		
id	**Days total**	**Days state**		**Reference**		**Prediction**		**Result**	**Prediction**		**Result**	**Prediction**		**Result**
		rem	**man**	**rem**	**man**	**rem**	**man**		**rem**	**man**		**rem**	**man**	
350	40	20	20	0	1	0	1	Both	0	1	Both	0	1	Both
353	43	21	22	0	1	0	0	One	0	1	Both	1	0	None
464	40	20	20	0	1	1	0	None	0	0	One	1	1	One
465	108	54	54	0	1	0	0	One	0	0	One	1	1	One
468	36	19	17	0	1	0	0	One	1	1	One	1	1	One
528	106	53	53	0	1	0	1	Both	0	0	One	0	0	One
577	33	17	16	0	1	1	1	One	0	0	One	0	1	Both
684	64	31	33	0	1	0	0	One	1	0	None	0	0	One
874	48	24	24	0	1	0	1	Both	0	0	One	1	1	One

Table 1: Classification results of mania-remission by LOOCV using COV, median, and SD method with PCA approach

						COV			median			SD		
id	**Days total**	**Days state**		**Reference**		**Prediction**		**Result**	**Prediction**		**Result**	**Prediction**		**Result**
		rem	**man**	**rem**	**man**	**rem**	**man**		**rem**	**man**		**rem**	**man**	
350	40	20	20	0	1	0	0	One	0	0	One	0	1	Both
353	43	21	22	0	1	0	0	One	1	1	One	1	1	One
464	40	20	20	0	1	0	0	One	0	0	One	0	1	Both
465	108	54	54	0	1	0	0	One	0	0	One	0	0	One
468	36	19	17	0	1	0	0	One	0	1	Both	0	0	One
528	106	53	53	0	1	0	0	One	0	0	One	1	1	One
577	33	17	16	0	1	0	0	One	0	0	One	0	1	Both
684	64	31	33	0	1	0	0	One	1	0	None	0	0	One
874	48	24	24	0	1	0	0	One	0	0	One	0	1	Both

Table 2: Classification results of mania-remission by LOOCV using COV, median, and SD method with max-diff approach

						COV			median			SD		
id	**Days total**	**Days state**		**Reference**		**Prediction**		**Result**	**Prediction**		**Result**	**Prediction**		**Result**
		rem	**dep**	**rem**	**dep**	**rem**	**dep**		**rem**	**dep**		**rem**	**dep**	
144	112	56	56	0	1	0	0	One	1	1	One	1	1	One
331	32	16	16	0	1	0	0	One	1	1	One	1	1	One
333	84	41	43	0	1	0	1	Both	1	1	One	1	1	One
339	42	21	21	0	1	1	0	None	1	1	One	0	1	Both
350	36	18	18	0	1	1	0	None	1	1	One	0	0	One
465	50	24	26	0	1	0	0	One	1	1	One	1	1	One
468	36	19	17	0	1	0	0	One	0	1	Both	0	1	Both
573	106	53	53	0	1	1	0	None	1	0	None	1	1	One
574	68	33	35	0	1	0	0	One	0	0	One	1	0	None
575	52	25	27	0	1	1	1	One	0	1	Both	1	0	None
577	40	20	20	0	1	0	0	One	1	1	One	1	1	One
681	42	21	21	0	1	0	0	One	1	1	One	1	1	One
684	64	31	33	0	1	0	1	Both	0	1	Both	1	0	None
804	42	22	20	0	1	0	1	Both	0	0	One	1	0	None
806	47	24	23	0	1	0	1	Both	1	1	One	1	1	One
808	95	48	47	0	1	0	1	Both	1	1	One	1	1	One
811	50	25	25	0	1	0	1	Both	1	1	One	1	0	None
869	113	56	57	0	1	0	1	Both	1	1	One	1	1	One

Table 3: Classification results of depression-remission by LOOCV using COV, median, and SD method with PCA approach

						COV			median			SD		
id	**Days total**	**Days state**		**Reference**		**Prediction**		**Result**	**Prediction**		**Result**	**Prediction**		**Result**
		rem	**dep**	**rem**	**dep**	**rem**	**dep**		**rem**	**dep**		**rem**	**dep**	
144	112	56	56	0	1	0	0	One	1	1	One	0	0	One
331	32	16	16	0	1	0	0	One	1	1	One	0	1	Both
333	84	41	43	0	1	0	0	One	1	1	One	0	1	Both
339	42	21	21	0	1	0	0	One	0	0	One	1	1	One
350	36	18	18	0	1	0	0	One	1	1	One	0	1	Both
465	50	24	26	0	1	0	0	One	1	1	One	0	0	One
468	36	19	17	0	1	0	0	One	1	1	One	0	0	One
573	106	53	53	0	1	0	0	One	0	0	One	0	0	One
574	68	33	35	0	1	0	0	One	0	0	One	0	0	One
575	52	25	27	0	1	0	0	One	0	1	Both	0	1	Both
577	40	20	20	0	1	0	0	One	1	1	One	0	0	One
681	42	21	21	0	1	0	0	One	1	1	One	0	0	One
684	64	31	33	0	1	0	0	One	0	1	Both	0	0	One
804	42	22	20	0	1	0	0	One	0	0	One	0	1	Both
806	47	24	23	0	1	0	0	One	1	1	One	1	1	One
808	95	48	47	0	1	0	0	One	0	1	Both	0	1	Both
811	50	25	25	0	1	0	0	One	0	1	Both	0	0	One
869	113	56	57	0	1	0	0	One	1	1	One	1	1	One

Table 4: Classification results of depression-remission by LOOCV using COV, median, and SD method with max-diff approach