Bridging Individual and Group Perspectives in Psychopathology

Computational Modeling Approaches using Ecological Momentary Assessment Data

Bridging Individual and Group Perspectives in Psychopathology

Computational Modeling Approaches using Ecological Momentary Assessment Data

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. Pamela Habibović
in accordance with the decision of the Board for Deans,
to be defended in public on Tuesday 15, April 2025 at 10.00 hours

by

MANDANI NTEKOULI

Supervisors:

Prof. Dr. Gerhard Weiss,	Maastricht University
Prof. Dr. Anne Roefs,	Maastricht University

Co-supervisors:

Dr. Gerasimos Spanakis,	Maastricht University
Dr. Lourens Waldorp,	University of Amsterdam

Assessment Committee:

Chair:	
Prof. Dr. Mark Winands	Maastricht University, chair
Members:	
Prof. Dr. Denny Borsboom	University of Amsterdam
Prof. Dr. Adriana Iamnitchi	Maastricht University
Dr. John Paparrizos	University of Ohio / Aristotle University of Thessaloniki
Dr. Konstantia Zarkogianni	Maastricht University

NSMD New Science of Mental Disorders

This study is part of the project 'New Science of Mental Disorders' (www.nsmd.eu), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.004.016).

SIKS Dissertation Series No. 2025-21 The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN 978-94-6510-566-6

Front and back covers designed by Yaru Zhang

To my parents.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1. STUDYING MENTAL DISORDERS

The field of psychopathology is dedicated to the scientific study of mental disorders, focusing on understanding their symptoms, causes, and treatments [1–3]. Mental disorders include a wide range of conditions that impact mood, cognition, and behavior, causing distress and mental dysfunction in people's daily lives [4]. Common mental disorders include depressive disorders, anxiety disorders, eating disorders, bipolar disorder, schizophrenia, and post-traumatic stress disorder (PTSD), among others [5].

Mental disorders are highly prevalent, affecting millions of individuals worldwide [6]. The roots of mental disorders have not been discovered, and it is questionable if that can ever be achieved using a reductionist approach that focuses on breaking down complex phenomena into simplest components (e.g., biological mechanisms) [3, 7]. Mental disorders are characterized by a wide range of signs and symptoms that can vary significantly between individuals and over time. For instance, their etiologies involve multiple and overlapping factors, such as biological, psychological, and social factors [8, 9]. This makes it challenging to diagnose, treat, and generally understand the underlying mechanisms.

Starting with the challenge of diagnosing, mental disorders can be difficult to identify, leading to cases where they remain untreated [10]. For instance, symptoms may not be severe or immediately recognized as part of a disorder, which can delay or prevent proper diagnosis and treatment [11, 12]. Additionally, some individuals may experience subclinical symptoms that do not fully meet the diagnostic criteria but still impact their well-being [12]. The complexity of diagnosis can be further supported by the fact that many mental disorders share overlapping symptoms, making it hard to accurately distinguish between them [13]. This overlap can lead to misdiagnoses or the assignment of multiple diagnoses, complicating treatment and care.

However, even if a disorder is identified and treated, many people do not benefit sufficiently from the current methods of treatment [14, 15].

There have been cases of patients not responding at all to any treatment, or initially improving for a short time before relapse [16]. Therefore, another significant challenge in the field is the effective treatment of mental disorders and the prevention of relapse.

Poor treatment outcomes point to a significant gap in scientific understanding and clinical practice, suggesting that current methods may not fully address the complexities of mental disorders. This gap may arise either from flawed conceptualizations of the underlying mechanisms of mental disorders or from ineffective methods of capturing these mechanisms.

1.2. CONCEPTUALIZING MENTAL DISORDERS

Historically, the traditional medical model (also known as the disease model or common cause model) conceptualizes mental disorders the same way as physical illnesses [3, 17, 18]. According to this model, both mental and physical disorders arise from common causes. That is, an underlying common cause - likely a biological mechanism - gives rise to a set of independent symptoms. In this framework, symptoms are viewed as indicators of underlying common causes. By focusing on identifying and treating this root cause, the medical model has been a dominant framework for diagnosing mental health disorders.

While this approach has proven effective for many physical illnesses, where a clear biological cause (such as a virus or genetic mutation) can often be identified, it falls short when applied to mental disorders. Mental health conditions are inherently more complex, usually lacking a singular, identifiable biological cause and involving multiple interactions between genetic, environmental, psychological, and social factors. As a result, traditional diagnostic models may oversimplify this complexity.

In the next subsection, relying on the medical model, the categorical model is examined that continues to dominate mental health diagnostics today, although it may not fully capture the complexity of mental disorders. Following, an alternative conceptualization is introduced, the network approach, which may offer a more accurate representation of the underlying mechanisms of mental disorders.

1.2.1. CATEGORICAL DIAGNOSING MODEL

This traditional medical model, which assumes that disorders arise from common underlying causes, has heavily influenced the development of diagnostic frameworks in mental health. As a result, it is common for professionals/practitioners in the field of mental health to rely heavily on categorical approaches for their diagnoses [19]. As noted by [20], attempts to organize and categorize the natural world can be traced back to the ancient Greeks, who relied on subjective judgments to discern similarities between organisms or objects. Similarly, in the context of mental disorders, unified categorical models for conceptualizing and diagnosing

mental health conditions are based on manuals, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM, currently DSM-5) [21, 22]. While the DSM offers a standard framework for organizing the phenomena of mental disorders and classifying them into categories (or classes), it incorporates several conceptual flaws that may contribute to insufficient treatment outcomes.

- According to DSM, the classification process is unified across individuals. It assumes that mental disorders manifest similarly in all individuals, meaning that all exhibiting the same symptoms fall into the same disorder/class [23]. However, such a scheme overlooks the significant variability in how symptoms present and progress. This one-size-fits-all model fails to account for the individual nature of mental health conditions, leading to treatment strategies that are not likely to be effective for everyone.

- DSM uses a categorical classification system. This system divides mental disorders into discrete categories based on the appearance and the frequency of symptom clusters. While this simplifies the diagnostic process, it does not fully reflect the complexity of mental disorders. Most individuals do not fit exactly into one category, as the same symptoms usually overlap in categorizing multiple disorders [24]. This hard classification can drive wrong diagnoses that further complicate treatment.

- Another issue is about the criteria to belong to a category. For diagnosis, DSM-5 relies on symptom checklists. By checking only the presence of some symptoms, other important aspects, such as the underlying mechanisms and context of these symptoms, are typically ignored. This reinforces the scheme of "common cause" that handles symptoms as the main causes of each disorder [7].

- Regarding DSM's checklists, the assumption for a diagnosis is that a fixed number of symptoms must be present. This criterion creates the possibility that two individuals with different symptom profiles may receive the same diagnosis [25]. For instance, one individual might display symptoms 1 through 5, while another might show symptoms 5 through 9, with both being diagnosed with the same disorder despite experiencing very different aspects of the condition. Consequently, this oversimplified diagnosis approach can impact treatment strategies, failing to address the unique needs of each individual.

Given the challenges of diagnosing mental disorders, the traditional medical model, which views symptoms as arising from a common underlying cause, may not fully capture the complexity of these conditions [7]. To address these limitations, recently, there has been a paradigm shift to the network approach to psychopathology.

1

1.2.2. NETWORK APPROACH TO PSYCHOPATHOLOGY

Unlike the traditional medical model, the network approach conceptualizes mental disorders not as the result of a single underlying cause, but as complex systems where psychopathological-related variables (or elements) interact with and influence each other directly [8, 24, 26]. This conceptualization shift is illustrated in Figure 1.1. The involved psychopathological-related variables extend beyond symptoms, including other types of psychopathological information, such as mood states, behavior, thoughts and context factors.

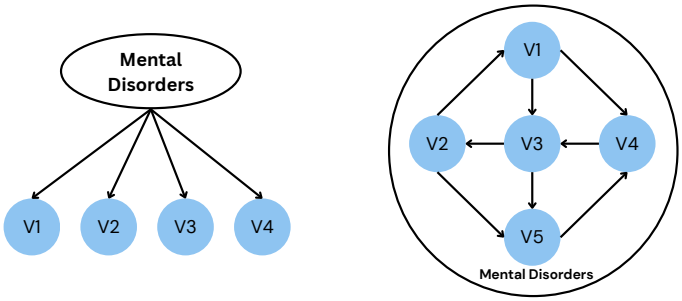


Figure 1.1: Conceptualization shift from the medical model (left), where mental disorders are the result of a single underlying cause, to the network approach to psychopathology (right), where mental disorders arise from the interacting symptoms (adapted from [27]).

According to the network approach to psychopathology, the psychopathology-related variables - instead of being indicators of a common cause - are assumed to cause one another [24, 28]. For example, by reinforcing or inhibiting one another, this paradigm shift highlights the importance of interplay between variables, where each one can influence and be influenced by others, creating a complex network of relationships. This way, mental disorders can be easily represented by networks (or graphs) that consist of variables (as nodes) and pairwise relations (as edges) between them [29]. Networks are a popular representation, also seen in other fields, such as social networks, where people are connected through relationships, and neural networks where neurons are connected through axons and dendrites [30, 31].

A central idea of conceptualizing mental disorders as networks is a more straightforward discovery of the pathways through which psychopathology-related variables influence each other [28]. This facilitates the identification of key variables that play a central role in the disorder, which is important for accurate diagnosis, but possibly also for treatment [32]. Instead of treating the disorder, the network's advantage is to facilitate treating the present crucial variables and relations or interactions. Specifically, intervening on these key variables and/or connections can potentially influence the structure of the entire network, offering new

approaches for treatment [33].

From a network perspective, this approach presents an additional benefit in hard or challenging cases, where psychopathology-related variables do not clearly correspond to a specific disorder, such as in comorbidities [24]. By analyzing the pathways of variables across different disorders, network models can reveal connections that might go undetected by traditional diagnostic methods.

COMORBIDITY

The network approach significantly enhances our understanding regarding comorbidities, which is the condition of an individual simultaneously exhibiting symptoms of two or more disorders [34]. Traditionally, comorbidities are handled independently, ignoring the complex interactions between the symptoms of different disorders. However, through a network, various psychopathology-related variables, belonging to several disorders, are taken into account. Therefore, the interactions among all the variables and consequently all the potentially involved disorders can be represented and analyzed [24]. Figure 1.2 illustrates an example of comorbidity between two disorders using the network approach. This insight is essential for recognizing the central variables that can help in developing more effective treatments, especially for hard comorbid conditions.

Consequently, this approach also supports a shift in understanding mental disorders, moving from a diagnosis-specific focus to a broader, trans-diagnostic perspective [35, 36]. This shift facilitates the integration of multiple clinical insights, enhancing our ability to address the complex nature of mental health conditions.

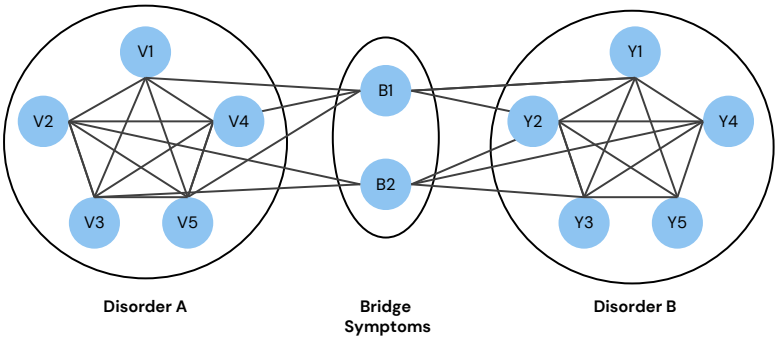


Figure 1.2: From a network perspective, comorbidity arises as a result of direct relations between the bridge symptoms (B1, B2), that overlap between disorders A and B (adapted from [24]).

This alternative conceptualization should be further investigated for its ability to improve the understanding and treatment of mental disorders.

1.3. MODELING THE NETWORK APPROACH TO PSYCHOPATHOLOGY

The primary objective of network models in the field of psychopathology is to describe the underlying complex relations among a diverse set of psychopathology-related variables [37]. Depending on the nature of the variables collected and the specific requirements of the experimental design, various network models can be applied. Such models are borrowed from the fields of multivariate statistics and network science to investigate the structure of relationships in multivariate data [29, 38]. Then, each model offers distinct methodologies for analyzing the data, which can lead to different interpretations and insights regarding the derived associations among the variables [37].

The most popular strategy of network modeling takes advantage of conditional associations to describe the network structure among a set of variables. A conditional association between two variables is established when these variables demonstrate probabilistic dependence, conditioned on all other examined variables within the dataset. Choosing the most appropriate model for estimating conditional association depends on the structure of the data, aiming to accurately capture the inter-dependencies across all variables.

1.3.1. STATIC CROSS-SECTIONAL MODELS

Cross-sectional network models play a vital role in mapping the relationships between psychopathology-related variables at a specific point in time (not necessarily the same time point for each individual) across individuals [2, 33]. More specifically, cross-sectional models analyze data collected at a single point in time by a large number of individuals. The associations between variables are driven by individual differences, providing a snapshot of their relationships. These models are useful for understanding the structure of variable relationships in a population at a specific time-point, which can help in identifying core variables or mechanisms that may be driving a disorder. Subsequently, potential central variables or states could be treated as targets for improved therapeutic interventions. Among many studies targeting specific mental disorders, some great examples of cross-sectional network studies can be found in [39, 40].

To construct these cross-sectional network models, the primary goal of network estimation is to uncover the relationships between variables, focusing on the strength and direction of the edges connecting them. These edges represent the associations between variables, and estimating them accurately is critical for understanding the structure of the network. Techniques such as partial correlation or regularized regression are commonly used to determine these connections, helping to define the edges between nodes [41, 42].

However, the use of cross-sectional models in psychopathology presents several significant challenges. First, it should be noted that the derived

network topologies describe differences between individuals, assuming that all individuals are homogeneous [37]. Such an assumption is not always true, leading to generalized effects that may not reflect the actual mechanisms that could characterize the diversity of individuals within the dataset. Second, analyzing cross-sectional networks at a single specific time-point yields static structures that do not capture the evolving nature of psychological phenomena [19]. This static representation shows a significant limitation, as it fails to reflect the continuous and dynamic changes in an individual's psychological state over time. Utilizing such static data is connected to some further concerns regarding network stability and replicability [2, 43, 44].

To better understand the evolving patterns at play within individuals, methodologies should go beyond cross-sectional analysis including temporal (or longitudinal) modeling approaches using time-series data. These methodologies allow capturing intra-individual measurements over time, providing a more detailed picture of mental health by adding the temporal dimension within individuals.

1.3.2. TEMPORAL MODELS

In response to the challenges posed by individual heterogeneity and the static nature of the widely used cross-sectional network models, there has been a growing interest in the development of temporal network models [2, 33, 45]. These models can handle time, as an additional dimension of the data, allowing for a dynamic analysis and understanding of mental health disorders. By incorporating temporal dynamics, these can capture the evolution of individual psychopathology-related variables as well as their inter-relations over time. Accounting for individual heterogeneity, these models focus on each individual separately, tailoring network structure to the unique dynamics of each one [46]. To apply such temporal network models, intensive time-series data is necessary. These studies involve monitoring the same individuals over an extended period, which facilitates observing changes and trends regarding various factors as they naturally occur.

1.4. CAPTURING MENTAL DISORDERS

Having discussed the conceptual challenges of understanding mental disorders, another critical issue lies in how these disorders are captured and measured. The methods used to collect data in the field of mental health are crucial, as they shape our understanding of mental disorders and inform clinical decisions. While retrospective methods have been widely used, they come with several limitations [47].

1.4.1. TRADITIONAL RETROSPECTIVE METHODS FOR CAPTURING MENTAL DISORDERS

Traditionally, studying the phenomena of mental disorders has been based on data collected through retrospective interviews during health assessment sessions [47, 48]. In this approach, individuals are asked by clinicians to recall and report their experiences and symptoms over a period of time, typically ranging from a few weeks to a few months. Based on this knowledge, clinicians make decisions regarding diagnosis and treatment [21]. However, this method of data measurement involves several limitations [47–49]:

- Individuals may not accurately recall past events and symptoms, or they may only partially recall them, ignoring useful information, leading to incomplete or biased data.
- Based on the retrospective nature of assessments, data is captured at discrete time points, missing the continuous and dynamic information regarding the progression of mental health symptoms.
- While clinical assessments often ask individuals to reflect on their everyday experiences, the data collected in these settings may not accurately capture the complexity or variability of those experiences as they occur in real-time. The clinical environment combined with reliance on retrospective recall can limit the real-world applicability and accuracy of the findings.

Such limitations of traditional retrospective assessments play an important role in poor treatment results and high relapse rates [14]. Consequently, there is an urgent need for more accurate, dynamic, and context-sensitive approaches to measuring mental disorders.

1.4.2. ECOLOGICAL MOMENTARY ASSESSMENT FOR CAPTURING MENTAL DISORDERS

To address the significant challenges inherent in the study and treatment of mental disorders, research in the field of psychopathology shifts to more innovative methodologies for collecting data, aiming to uncover the underlying nature of mental disorders. One of the most promising methodological advancements for capturing mental disorders is through Ecological Momentary Assessment (EMA) studies [50, 51]. EMA is also known as "Experience Sampling Methods (ESM)" [52], "Ambulatory Assessment (AA)" [53] or "Intensive-longitudinal Study Design" [54]. All refer to a research tool that allows us to collect repeated measurements on various psychopathology-related variables, such as individuals' symptoms, behaviors, and experiences in their natural environments. According to the EMA protocol for data collection, participants respond to digital questionnaires on their personal devices (such as smartphones), where they rate the perceived intensity of different psychopathology-related

questions (EMA items or variables) along with contextual information (such as their geo-location, activity and company). An example of such a questionnaire, repeated over time, is presented in Figure 1.3.

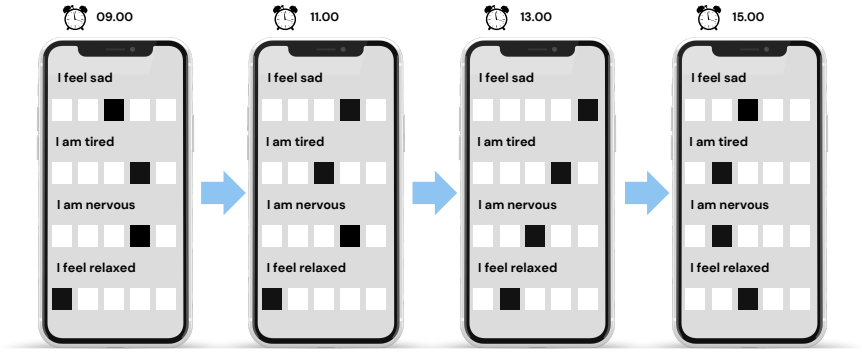


Figure 1.3: EMA data collection through digital questionnaires, captured multiple times a day to provide real-time insights.

Understanding the potential of EMA requires breaking down its three core components: ecological, momentary, and assessment.

- The term "ecological" refers to the context of collecting data on individuals in their natural environment. Unlike traditional methods that gather information in clinical settings, EMA captures data in real-world conditions where individuals actually live and interact daily. By this, the negative effects of retrospective measurements regarding recall bias [55] are prevented, ensuring that the data reflects real-time experiences. This makes the findings more relevant and applicable to real-life scenarios.

- "Momentary" emphasizes the importance of capturing data at specific moments in time as they are collected almost in real-time [56]. Particularly, prompts are sent to individuals' smartphones at regular, but randomized, intervals of 1-2h. Also, these measurements are typically taken multiple times each day (e.g., approximately 8 times) over several days, weeks, or even months. The collection of repeated measurements leads to a temporal (time-series) dataset of a frequency of every 1-2h, for each individual.

- "Assessment" refers to the structured and systematic process of measuring relevant psychological and behavioral variables over time. Through brief and minimally intrusive digital questionnaires delivered at regular intervals, EMA allows for a rich, fine-grained collection of psychopathology-related information.

EMA data is organized into 3 granularity levels, the variable-, temporal- and individual-level, as depicted in Figure 1.4. These are described as follows:

- Multi-dimensional (or multivariate) data: On each measurement moment, questions are asked about various psychopathology-related behaviors, experiences, and symptoms. These represent the variables of the EMA dataset, leading to a multivariate dataset.
- Time-series data: Collecting repeated measurements over time (frequency of every 1-2h) leads to a temporal (time-series) dataset for each individual.
- Across-individuals data: During an EMA study, data from multiple individuals are collected, all represented by the same set of variables.

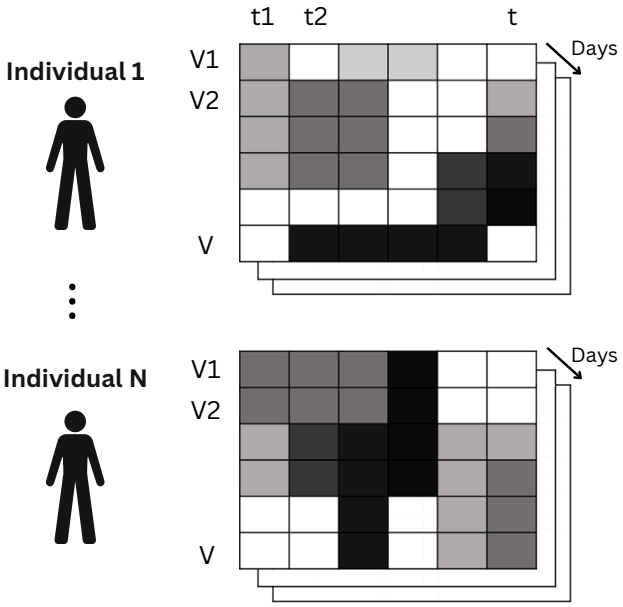


Figure 1.4: EMA data structure, organized in 3 granularity levels - variables (EMA items), time-points and individuals - capturing the multivariate, temporal and personalized nature of the data.

Therefore, EMA data with these characteristics is structured as a Multivariate Time-series (MTS). The multi-level structure of EMA provides a significant amount of information to better understand mental disorders.

1.5. TEMPORAL NETWORKS FOR MODELING EMA

The next steps of methodological research in the field of EMA are directed at developing statistical techniques capable of identifying potential network structures among psychopathology-related variables derived from empirically collected EMA data.

After collecting EMA time-series data, it can be easily utilized in temporal network models to capture the relations between psychopathology-related variables over time. This approach provides a more granular view of how psychopathology-related variables evolve and interact, offering great insights into understanding the day-to-day dynamics of mental disorders and developing more accurate interventions. Further details about the most popular temporal network models, such as the Vector Autoregressive (VAR) model, are given in Chapter 2.

One of the major strengths of network models is their interpretability. By representing the inter-relations between variables as edges in a network, they can be easily visualized. An example of such a network is shown in Figure 1.5. Consequently, understanding complex interactions becomes quite straightforward. Particularly, each edge in the network reflects the association's strength and direction of one node (variable) to another. For example, the directed edge $e12$ represents the strength of the association from node $V1$ to $V2$, which differs from the edge $e21$ in the opposite direction. Quantitatively, this association is represented by a numerical value, indicating that changes in one variable are assumed to directly and proportionally predict changes in another. Therefore, network models rely on linear assumptions about the data. Linearity facilitates mathematically approaching EMA data, making them computationally feasible and often easier to implement. Nevertheless, these linear network models may be insufficient to uncover the possible complicated interactions and describe the real complex nature of mental disorders [57]. Consequently, there has been an urgent need to develop more advanced statistical methods to model the underlying complex psychological mechanisms [57]. These new methods aim to provide deeper insights into the patterns and complex interplay of psychopathology-related variables of mental disorders.

1.6. MACHINE LEARNING FOR MODELING EMA

Despite the paradigm shift to temporal network models for psychopathology, to achieve better insight into mental disorders, the development of robust and accurate models remains essential [58–61]. In this context, robust models refer to reliable models that are resilient to variations in data, while accurate models are those capable of correctly predicting an outcome by capturing complex patterns and relationships within the data. Exploiting the rich EMA data structure and information can be valuable for building accurate personalized predictive models, but also for capturing the intricate interplay between psychopathology-related variables. These models aim to provide a comprehensive understanding of how variables influence one another over time, offering further insights beyond individual predictions. By focusing on different tasks, such as predicting a future event or the progression of individual symptomatology (course of mental disorder), models' accuracy and robustness can serve as an indicator of performance, ultimately assessing the represen-

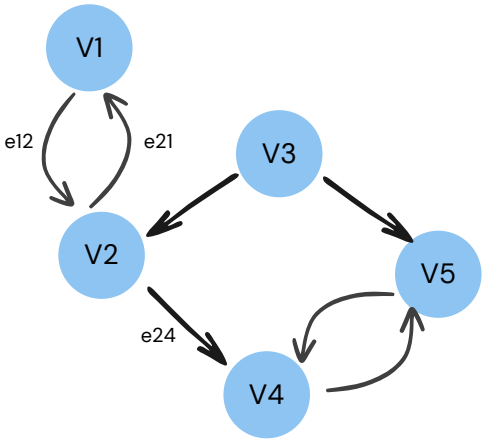


Figure 1.5: A directed network where EMA variables ($V1 - V5$) are represented by nodes and connections by edges. The edges represent direct relationships from one node to another, for example, $e12$ indicates an effect from $V1$ to $V2$, while $e21$ from $V2$ to $V1$.

tations of variables' interactions derived from these models [62–65].

A promising direction to discover complex and higher-order interactions between EMA variables is using non-linear machine learning (ML) models [58–61, 66]. ML models can enhance the ability to accurately predict the occurrence of different psychopathology-related variables by recognizing complicated patterns or relations between them in existing data. The necessity of using ML models becomes more evident when dealing with a large number of variables, as richer information can lead to a more complex system that is unlikely to be represented by simple linear patterns [66]. These advanced data-driven approaches are, also, not dependent on any formal assumptions regarding the structure of the data, unlike linear statistical models, which often require specific assumptions such as linearity, stationarity or normality (later discussed in Chapter 2.3.1). This flexibility allows data-driven methods to capture complex, non-linear relationships and interactions within the data that linear models might not reveal.

1.6.1. IDIOGRAPHIC AND NOMOTHETIC APPROACHES

The idiographic (also called personalized or individual) predictive approach is often the first method applied to a dataset, focusing on the unique data of each individual [36, 67, 68]. Given that there is large inter-individual heterogeneity in mental disorders, ideally separate predictive models are built for each individual. However, a significant challenge in building personalized models is the limited number of data points available per variable for each individual. Data collection in EMA studies is

often limited because these studies typically do not run for extended periods of time to avoid overburdening participants [69]. These small datasets potentially lead to training overfitted models without being capable of generalizing, or even to situations where models cannot be trained at all [45]. The latter scenario is particularly common in highly imbalanced datasets, where one of the outcomes may not be sufficiently represented in the whole dataset or when splitting it into training or test personalized datasets [70].

Moreover, inspired by the traditionally applied cross-sectional studies, research often extends the idiographic models to nomothetic (group-level) predictive methods. Specifically, information collected from other individuals in the same EMA study can prove beneficial for modeling [45]. Identifying common patterns and generalized trends within larger populations can provide broader insights to generally understand mental disorders. The most common way of integrating data of more than one individual in a model is to concatenate the data of all individuals together in a single dataset. The augmented dataset is then used to construct a group-level model.

Such models produce generalizable predictions that can be relevant to a wider range of individuals, beyond those included in the training sample. For example, a group-level model can be applied to new individuals who were not part of the training set and may even belong to different populations. An additional benefit is that it can be applied to individuals with limited data, where personalized modeling is not feasible due to an insufficient number of training points or imbalanced datasets. By capturing patterns that generalize across individuals, these models provide a flexible and alternative solution.

1.6.2. CLUSTER-BASED APPROACHES

Utilizing data from multiple individuals can significantly enhance the generalizability of a model, allowing for broader applications across the population (referring to individuals of the same data collection). However, this approach has its own challenges, especially when there is a large heterogeneity among the individuals whose data is included. Large variability, while valuable for capturing a broad range of experiences, can weaken the model's accuracy, making it less effective for a specific group or individual.

In a way to further refine the nomothetic approaches, advanced cluster-based approaches could split the population into more homogeneous subgroups. These clusters can be formed based on identifying individuals with various similar characteristics, such as demographic factors, symptom profiles, etc. [71]. Exploring various characteristics and clustering methods is essential for enhancing the models' utility and efficacy. This exploration enables the development of models that not only maintain generalizability across the population but also offer more precise insights and solutions compared to traditional nomothetic

approaches that use all data collectively.

By effectively grouping similar individuals, more sophisticated cluster-based models can bridge the gap between capturing the diversity of entire populations and recognizing the unique needs of individual subgroups. This approach enables a deeper, more practical application of psychological research, aligning broad data collection with focused, effective solutions.

1.7. RESEARCH STATEMENT AND QUESTIONS

This dissertation focuses on applying advanced data analysis techniques for several tasks. More specifically, there are two main directions of analysis: (1) application of advanced non-linear methods and (2) exploiting the nomothetic predictive approach. Five research questions are examined:

- **Research Question 1 (RQ1):** Are non-linear individual models capable of outperforming the linear network models?

  Traditional temporal network models have been widely used due to their simplicity and interpretability. These models typically assume linear relationships among variables, making them straightforward to apply and analyze. However, mental disorders are considered complex systems of dynamically interacting variables. Thus, EMA data in psychopathology is expected to exhibit complex, non-linear relationships that network models may not capture effectively. Non-linear machine learning models, such as tree- or boosting-based algorithms [72], offer promising alternatives [61, 66]. In particular, these models have the potential to better capture the data patterns as well as interactions between psychopathology-related variables. However, according to the literature, their application to EMA data is limited. Given this gap, this dissertation aims to explore various non-linear methods that still preserve the aspect of explainability, a crucial factor for understanding the model's decision-making process, especially when the model is not as transparent or interpretable as linear models. This investigation will focus on assessing whether these advanced non-linear models can outperform traditional temporal network models in predicting future psychopathology-related variables as an outcome.

- **Research Question 2 (RQ2):** Could nomothetic modeling approaches, by integrating more data, exceed the predictive performance of individual models?

  According to the nomothetic predictive approach, information from multiple individuals can be utilized in the modeling process, by integrating data from a larger number of individuals. This raises the question of how to effectively incorporate such information into modeling to enhance individual performance. Initial nomothetic

approaches utilize all available data (using-all-data or aggregated models), which can provide a baseline model with a potentially improved predictive accuracy due to increased data volume. Nevertheless, to further optimize the integration of additional information, more sophisticated methodologies need to be employed. Specifically, clustering techniques that group individuals based on similar characteristics can be proven more effective. By identifying and grouping similar individuals, models can be tailored to these specific subgroups, thereby training them on more homogeneously relevant data. This cluster-based approach enhances the predictive performance of the models, but also retains a level of personalization.

- **Research Question 3 (RQ3):** How could nomothetic modeling approaches effectively integrate group-based information while maintaining the focus on individual data?

  Although nomothetic and cluster-based models offer advanced opportunities compared to only using individual data, in most cases, the utilized knowledge is too broad and potentially does not reflect every individual separately. This brings up the question of how to balance a focus on the individual with useful group-based information. Individual focus can derive from several aspects of knowledge learned during the training process. Specifically, this involves refining algorithms that can dynamically adjust to the unique information of an individual while utilizing the predictive power provided by a broader group. Borrowing approaches from the field of advanced machine learning, such as transfer learning [73] and knowledge distillation [74], different methodologies are investigated and employed to enhance the predictive performance of individual models. Such methodologies have the potential to balance the benefits of both individual and nomothetic models.

- **Research Question 4 (RQ4):** What individual characteristics extracted from time-series can be used to effectively group individuals into homogeneous clusters?

  While cluster-based modeling can provide a way of integrating individual and nomothetic models, it is essential to utilize effective and homogeneous clustering-uncovered groups of individuals. Particularly, a crucial step in this process is determining what characteristics (types of information) from each individual can be used to accurately cluster multivariate time-series (MTS) data. The most straightforward clustering approaches use raw time-series data and explore the most suitable time-based similarity measures for comparing time-series. Beyond raw data, given the complexity and high dimensionality of EMA data, clustering could be based on several types of representational information. For example, model-derived information, such as model's coefficients, could be also utilized, reflecting another promising clustering category, that is model-based

clustering [75]. The investigation should identify such key individual characteristics that can influence the clustering process to achieve meaningful insights about individual grouping and similar EMA patterns.

- **Research Question 5 (RQ5):** How can we evaluate the time-series clustering results derived from different unsupervised clustering algorithms?

  Various clustering approaches can be employed in a group-based predictive approach. However, clustering is an unsupervised problem and the true underlying groups are not commonly known. Because of the unsupervised nature of the problem and the large number of possible clustering parameters, it is quite difficult to evaluate the produced results. Each method, according to its objective function and parameters, aims to separate data in the most appropriate way, every time leading to a different group separation. Thus, all clustering-related choices demand ways to examine and validate possible EMA clustering approaches for different scenarios. Beyond methodologies relying only on internal evaluation measures, explanations are further provided to examine the effectiveness of the derived groups of individuals.

1.8. THESIS OVERVIEW

This dissertation is organized into 8 chapters. Starting from the current chapter of the Introduction, the foundation of the current research field is provided, where motivation and context are discussed. The rest of the chapters focus on describing the data analysis pathway, ranging from idiographic and nomothetic to group-based predictive approaches. More specifically, the rest of the chapters are structured as follows:

- **Chapter 2** provides the methodological background on the typical modeling approaches applied to EMA data. Starting from the widely used linear models, such as Vector Autoregressive (VAR) models, this chapter leads to more advanced interpretable models and their adaptation for EMA data. Moreover, it introduces the EMA datasets analyzed in this work, along with the specific model output and prediction tasks that are examined.

- **Chapter 3** goes a bit further from the traditional linear models. This chapter explores the trade-off between linear and non-linear models by trying to integrate the strengths of both worlds (i.e. accuracy and interpretability). First, according to **RQ1**, it focuses on using non-linear interpretable ML models in the context of individual classification problems. ML models can enhance the ability to accurately predict the occurrence of different psychopathology-related

variables by recognizing complicated patterns in the data. Second, apart from individual approaches, this chapter partially contributes to **RQ2** and **RQ3** by investigating two different nomothetic approaches to integrate data of more than one individual, one using all data directly during training and one based on more sophisticated approaches, such as knowledge distillation. To evaluate both questions, the performance of various ensembles of trees is compared to linear models using imbalanced synthetic and real-world datasets.

- **Chapter 4** utilizes ML to identify similar patterns in EMA data across different individuals through clustering. This approach aims to refine the previously examined nomothetic methods by incorporating clustering results into group-based strategies to address **RQ2** and enhance personalized performance. Particularly, it focuses on clustering EMA data of individuals based on the raw MTS data, thereby also contributing to **RQ4**. Since clustering is an unsupervised problem, it is challenging to assess whether the resulting grouping is successful. Therefore, various clustering methods are assessed using simulated data designed to resemble EMA patterns as well as a real-world dataset collected as part of our project, NSMD [36]. Additionally, several internal evaluation measures, such as the Silhouette coefficient, are examined, contributing to **RQ5**.

- **Chapter 5** continues the clustering exploration using a different EMA data representation, such as model-based information. In an attempt to additionally address **RQ4**, two different model-based clustering approaches are examined. The first clustering method is based on model-extracted parameters of individual models, whereas the second is optimized on the model-based forecasting performance. Similar to Chapter 4, it also contributes to addressing the challenges of clustering evaluation of **RQ5**. Both methods are analyzed using intrinsic clustering evaluation measures (e.g. Silhouette coefficients) as well as the performance of a downstream forecasting scheme, where each forecasting group model is devoted to describing all individuals belonging to one cluster.

- **Chapter 6** extends the work on clustering evaluation, essentially targeting **RQ5**. Apart from the previously investigated structure and quality of the clustering-derived results, another important aspect of evaluation is clustering explainability. In particular, this chapter proposes an attention-based interpretable framework to identify the important time-points and variables that play primary roles in distinguishing between clusters. A key part of this study is to examine ways to analyze, summarize, and interpret the attention weights as well as evaluate the patterns underlying the important segments of the data that differentiate across clusters.

- **Chapter 7** further explores more advanced modeling strategies

with the goal of not only taking advantage of group-based approaches but also prioritizing individual-level information. This contributes to addressing part of **RQ3**. More specifically, transfer learning approaches are applied to improve predictions for a specific individual (target domain) by incorporating data from similar individuals (source domain). This chapter focuses on boosting-based methodologies, which are adapted to EMA data and methodologically enhanced regarding their modeling process. To evaluate the effectiveness of all the proposed enhancements, such as the optimal selection of similar source domains and their weighting strategies, their impact on performance is mainly investigated.

- **Chapter 8** summarizes the proposed modeling-specific enhancements and discusses to what extent these significantly contribute to a better understanding of mental disorders. In particular, it considers how each research question is addressed and the impact on the field. Moving forward, it provides several future directions for analyzing EMA MTS data.

A summary of all main chapters (3-7) is presented in Table 1.1. This includes a more structured format of associating the main building blocks, that is chapters and research questions along with their interconnections.

Table 1.1: Summary of the main chapters (3-7) in the dissertation, providing their interconnections and the links to the research questions (RQs)

Chapter	Connection to other Chapters and RQs
3	• Extends linear models (presented in Chapter 2) by exploring non-linear personalized models (RQ1). • Investigates the modeling concept of nomothetic approaches (introduced in Chapter 1) by applying more advanced methodologies including more individuals (RQ2). • Explores a more sophisticated way to balance idiographic and nomothetic approaches, using a 2-step knowledge distillation method (RQ3).
4	• Refines the nomothetic approaches (investigated in Chapter 3) by using cluster-based methods (RQ2). • Explores clustering based on time-series EMA data (RQ4). • Assesses clustering through internal cluster evaluation (RQ5).
5	• Extends clustering exploration (presented in Chapter 4) based on model-based information (RQ4). • Evaluates model-based clustering through performance (RQ5).
6	• Extends clustering evaluation (presented in Chapter 4 and Chapter 5) by exploring clustering explanations (RQ5).
7	• Balances the advanced group-based modeling strategies (proposed in Chapters 4 and 5) with individual-level data (RQ3). • Explores the concept of boosting-based transfer learning.

2

METHODOLOGICAL BACKGROUND: FROM THE LINEAR NETWORK APPROACH TO NON-LINEAR MODELING

2.1. INTRODUCTION

In the last decade, the collection of time-intensive, repeated, intra-individual measurements in psychology has grown, sparked by recent technological and methodological developments [33, 45, 51, 76]. This method is called Ecological Momentary Assessment (EMA). This chapter starts by describing the EMA data and their special time-series characteristics. Subsequently, a literature overview outlines the challenges of typical modeling approaches applied to EMA data. This discussion then leads to the introduction of more advanced and interpretable models [62].

2.2. EMA DATA

EMA data is organized into three levels of granularity: individuals, variables, and time-points. More specifically, in EMA, participants are prompted several times a day to answer questions on their smartphones, during a certain time period, which typically ranges between 2-4 weeks. Depending on the topic of the study, questions may probe for participants' mood states, craving for food, social circumstances, etc.

In practice, an EMA dataset consists of N independent and identically distributed (i.i.d.) individuals, each represented by a multivariate time-series (MTS). The dataset can be defined as $X = x_1, ..., x_N$. Then, each MTS x_i ($i = 1, ..., N$) of the i-th individual consists of a sequence of V ($v = 1, ..., V$) univariate time-series (UTS), where each one has a same-

length size T_i ($t = 1, ..., T_i$). A UTS represents each of the EMA items. An EMA item captures the responses or ratings (on a Likert or a visual analogue scale (VAS), see [77]) over time of each individual on a question. In general, the items asked in an EMA questionnaire, include individuals' daily-life experiences and emotions along with context information. To account for inter-individual differences in responses, regarding the perceived rating scales, data are typically normalized or scaled per person, leading to data that is often treated as continuous. Hence, a complete EMA dataset X is an N–dimensional MTS of V variables with a varying length of time-points T_i, as algebraically represented in Equation 2.1.

$$X = x_{1,1..V,1..T_1}, x_{2,1..V,1..T_2}, .., x_{N,1..V,1..T_N} \tag{2.1}$$

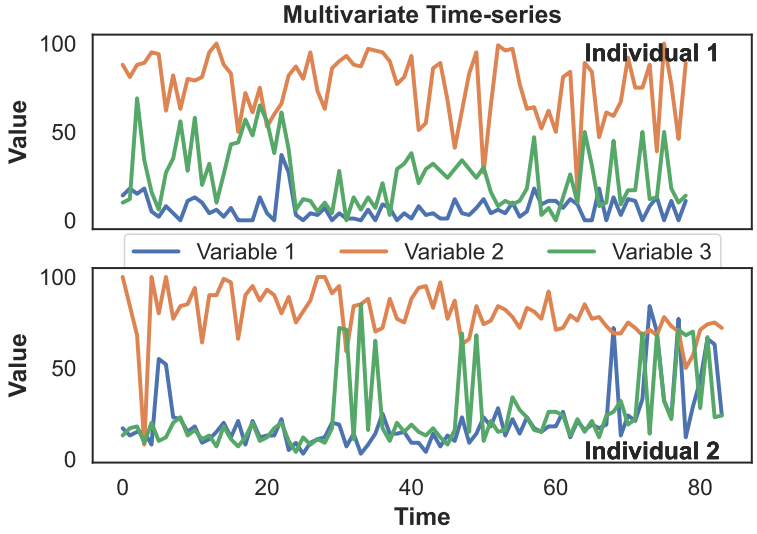


Figure 2.1: An example of EMA MTS data of 2 individuals, each measured across three variables over time.

2.2.1. EMA CHARACTERISTICS

EMA data have several characteristics that need to be considered in analyses. The following section introduces these characteristics along with potential solutions to address each challenge effectively.

MISSING MEASUREMENTS

First, some measurements can be missing for several reasons, mostly because of a technical problem or because a participant was not able to respond to an EMA prompt. This leads to datasets with incomplete time-series, meaning that some MTS have less than the maximum length

of T time points. Differences in the number of missing values among individuals make the MTS to be of variant length (where potentially $T_1 \neq T_2 \neq T_N$), as observed in Figure 2.1). Missing points also affect the time intervals between two consecutive measurements. When missing points exist, data are characterized as irregularly spaced MTS. An example of an incomplete individual time-series of a variable (or feature) is given in Figure 2.2, where gaps are apparent throughout time due to missing values.

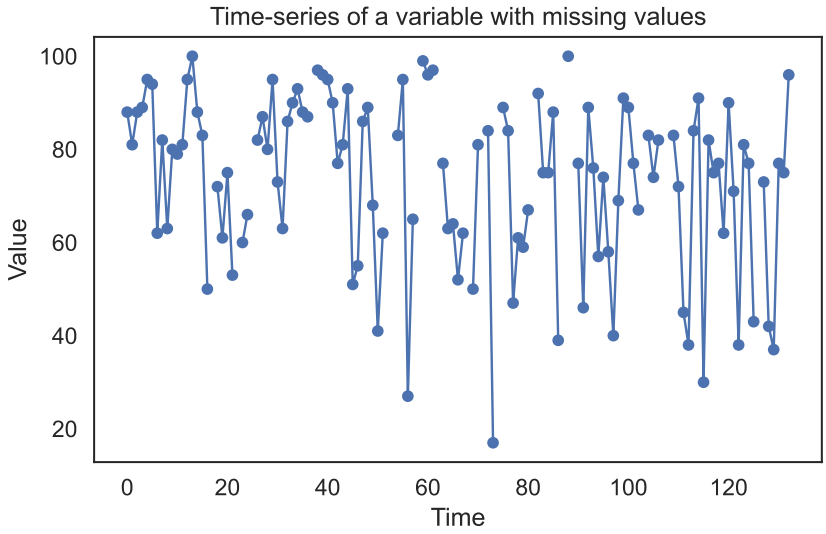


Figure 2.2: An example of a time-series variable with random missing values.

There are various strategies to handle such incomplete data [78]. First, all missing values can be omitted from the original dataset [79]. However, this approach seems valid only when data is missing completely at random (MCAR). Another common approach is to apply an imputation method to the data, which assumes that the data is either missing at random (MAR) or, in some cases, missing not at random (MNAR), depending on the chosen imputation technique [80]. During preprocessing, methods based on smoothing or interpolation, but also on machine learning (ML) algorithms are widely applied. These methods are needed to fill the gaps in data based on already existing patterns.

Beyond omission and imputation strategies, there are still ways to process data with missing values without relying on possibly biased techniques. A widely proposed approach is to apply a kernel to the raw data [81]. Kernel methods have dominated ML because of their effectiveness in dealing with a variety of learning problems [79]. To tackle these problems, a kernel works by mapping data into a higher-dimensional feature

space, called a reproducing kernel Hilbert space (RKHS) [82]. An RKHS is a nonlinear transformation that enables the smoothing or reshaping of data, effectively re-describing it such that (linear) separation becomes easier. The success of kernel methods relies on the fact that nonlinear data structures, like high dimensional MTS, can be transformed based on the type of kernel to a space where they are finally linearly separable.

VARIABILITY IN MEASUREMENTS SCALE

Apart from length invariances, resulting from missing values, EMA time-series data can also exhibit different characteristics in terms of measurement scale [83]. Regarding scaling, although EMA responses are usually recorded on a Likert (with 5 or 7 categories) or VAS scale, the range of given responses may differ per participant. For example, some individuals may tend to be biased towards the middle values, avoiding all the extreme scores, whereas others may do the opposite, resulting in a higher skewness in the data of some items, like negative emotions.

TIMING SHIFTS IN MEASUREMENTS

Additionally, different individuals' time-series can exhibit variations in timing, known as shift invariances [84]. A time-series represents the evolution of an individual's emotion, behavior or other variable. Thus, among different individuals, similar patterns of behavior can be seen, but shifted in time. For example, two individuals should be considered similar when they both show a similar pattern, e.g. a stable and then an increasing trend, even if the timing of the transition differs, with one individual transitioning quicker than the other. In a multivariate setting, similarities between all variables are considered, taking into account the relationships between multiple variables across time. When it is necessary to compare the shifted patterns across individuals, such as in clustering, an appropriate alignment method should be applied. For instance, alignment issues can be taken into account by an appropriate distance measure such as Dynamic Time Warping (DTW) [85]. Such measures will be further discussed in Section 4.3.1.

2.3. OVERVIEW OF TEMPORAL NETWORK MODELS

The network approach to psychopathology has gained momentum over the past decade [26, 86]. According to this approach, mental disorders can be represented by a network of interconnected psychopathology-related variables. Within this framework, all variables are represented as nodes, while the connections (edges) between them indicate their interrelations (or interactions) and mutual influences. Figure 2.3 shows an example of a directed network among the EMA variables.

One of the key aspects of the network approach lies in its straightforward visualization, representing the complex interplay between relevant variables. Its visualization enhances the simplicity and interpretability of

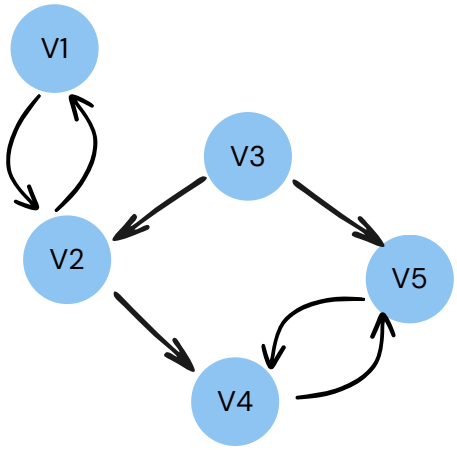


Figure 2.3: A directed network where EMA variables ($V1 - V5$) are represented by nodes and connections by edges. The directed edges (arrows) indicate the directional relationships between variables. For example, an arrow from $V1$ to $V2$ represents a direct effect or influence of $V1$ on $V2$.

the model. However, the effectiveness of the network approach relies heavily on accurately discovering the interconnections between these variables [87]. Thus, robust statistical methods are necessary to identify reliable connections.

Time-series statistical modeling is crucial for analyzing EMA data, which captures temporal sequences of psychopathology-related variables, such as symptoms, behaviors and experiences. The most popular class of time-series statistical modeling is linear Autoregressive (AR) models [88]. AR models predict the value of a variable at a specific time-point taking into account its values at previous time-points. The number of previous steps the model takes into account refers to the number of lags in an AR model. For instance, when considering only one step back, it is referred to as a 1-lag AR model.

However, in the case of EMA, where multiple variables are involved, AR models should be expanded to handle multivariate time-series data. The most commonly applied model to EMA data is the Vector Autoregressive (VAR) model [57]. Similar to AR, VAR models predict the variance of a variable at a specific time-point considering both its past values (self-loops or autoregressive effects) as well as the past values of the other variables (cross-lagged effects). Consequently, VAR models provide a detailed understanding of how psychopathology-related (or EMA) variables evolve and influence each other over time.

The 1-lag VAR model is described in Equation 2.2. In particular, in VAR, each variable v of individual i at time-point t, $x_{i,v,t}$, is modeled as a linear combination of all the other variables j at time point $t-1$ added to a constant term ϵ, representing both the intercept and error. The pa-

rameters (or coefficients) $w_{j,v}$ of the model play a key role because they quantify the linear dependency between the variables j and the output v. Therefore, the VAR-extracted coefficient reflects a linear relationship, as illustrated in Figure 2.4a, and effectively corresponds to an edge in the network, as depicted in Figure 2.4b. Thus, linearity simplifies the estimation of the model and the interpretation of the underlying phenomena.

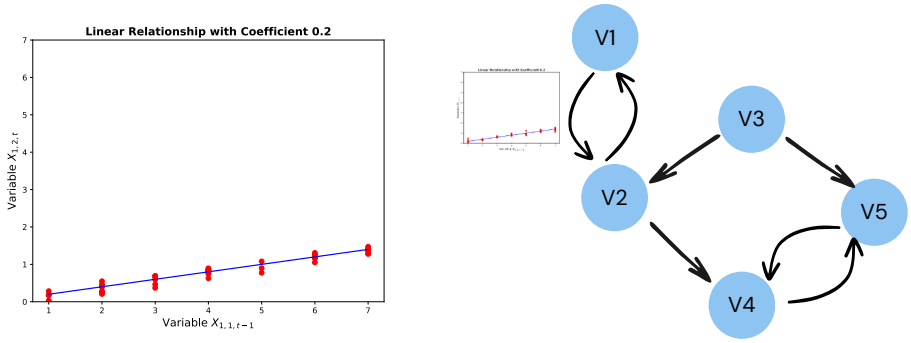
$$
\begin{aligned}
x_{i,1,t} &= \sum_{j}^{V} w_{j,1} \cdot x_{i,j,t-1} + \epsilon_1 \\
x_{i,2,t} &= \sum_{j}^{V} w_{j,2} \cdot x_{i,j,t-1} + \epsilon_2 \\
&.. \\
x_{i,V,t} &= \sum_{j}^{V} w_{j,V} \cdot x_{i,j,t-1} + \epsilon_V
\end{aligned}
\tag{2.2}
$$

2.3.1. CHALLENGES IN APPLYING VAR

Despite the widespread application of the VAR model to the theory of psychopathology, several assumptions inherent to the model are often violated in the context of collected EMA data. These violations present significant challenges that need to be taken into account, as outlined below and elaborated in [89, 90]:

- A crucial assumption of the VAR models is data stationarity, meaning that the statistical properties (e.g., mean values) of each variable are not expected to change over time. This assumption is necessary so that the model-derived coefficients are consistent throughout the entire time series. Nevertheless, given the complexity of mental disorders, the dynamics of EMA data are rarely constant. Consequently, when VAR is applied to EMA data, the identified relationships are likely ancestral relations (historical associations that persist over time) rather than direct causal interactions. This means that the model primarily captures long-term associations rather than immediate, direct interactions. Additionally, the presence of confounding factors (unmeasured variables) can further obscure these interactions, making it difficult to disentangle the underlying dynamics of the disorder [91].

- Another key assumption of VAR models is that the relationship between the variables is linear, meaning that changes in one variable result in linearly proportional changes in the output variable. Despite the given interpretation, in the case of real-world data, it is more realistic that EMA data exhibits non-linear relationships.



(a) (b)

Figure 2.4: (a) Linear relationship of input $x_{i,1,t-1}$ (or $V1$) to output $x_{i,2,t}$ (or $V2$), representing $w_{1,2} = 0.2$. (b) The linear relationship is reflected in the directed connection between $V1$ and $V2$.

- Although one of the advantages of the VAR-uncovered interactions is its interpretability reflecting the dynamic influence among variables, the VAR model does not derive true causality. Instead, VAR parameters show partial correlations which could only provide indications or hypotheses about the causal pathways among variables. Often, Granger causality is applied within VAR frameworks to infer directional dependencies [92]. However, Granger causality only suggests temporal precedence rather than actual causation, as these relationships are derived from observational data without experimental manipulation [93]. According to simulation studies [94, 95], VAR coefficients do not always reliably represent the underlying causal interactions.

- The estimation of the VAR parameters (coefficients) can be challenging, especially with high-dimensional data. When many variables are involved, the need for more data points increases to achieve sufficient statistical power. In such cases, techniques such as regularization, feature selection or dimensionality reduction should be employed [96].

- EMA data is expected to be equidistant, meaning the time interval between two consecutive measurements is almost equal. However, due to missingness in the data, this assumption is commonly violated.

To overcome these assumptions, a number of VAR extension models, such as mixed VAR [97], multi-level VAR [98], time-varying VAR [99], and VAR with Bayesian Dynamic Modeling [100] have been considered promising analysis methods. However, it seems unrealistic for linear models to uncover all possible complex interactions between variables

that are relevant to mental disorders. Exploring the application of more advanced ML models might be beneficial for the modeling of EMA.

2.4. ADVANCED NON-LINEAR INTERPRETABLE MODELS

Interpretable machine learning is an emerging research area focused on developing algorithms that can provide clear explanations for their predictions [101]. While accuracy is a necessary prerequisite of any ML algorithm, interpretability is another property that is important for a successful predictive model. However, most ML models are considered complicated black boxes, producing predictions without the whole decision-making process being transparent. Especially in case of critical and high-risk applications, it is important to understand how these decisions are made.

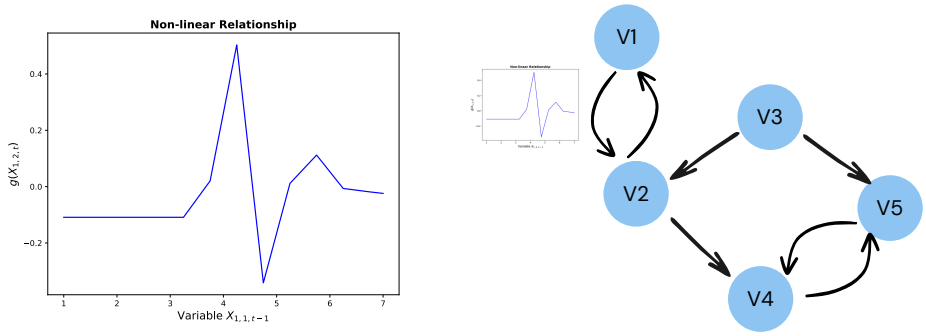
A first step towards interpretability is to know how much each variable (or feature) contributes to the output prediction. A clear example of a model providing such information is the linear regression model. In linear models, the prediction's outcome is modeled as a weighted sum of the existing features, with each weight indicating the feature's contribution.

Building on this, a natural extension of linear models is the more flexible Generalized Additive Models (GAMs) [102–104]. The main concept of GAMs remains the same as of the linear ones, expecting for the outcome to be an additive model of feature effects, but relaxing the restriction of the linear relationship. It allows the use of arbitrary functions to represent the features' effects. Mathematically, the relationship in a GAM for an individual $i = 1$ is presented in Equation 2.3, where f_j are the feature functions of a variable j and g is the link function (e.g., identity or logistic).

$$g(x_{1,v,t}) = \sum_j f_j(x_{1,j,t-1}) \tag{2.3}$$

The f functions can be based on regression spline models and tree-based models such as single trees or ensembles of bagged trees, boosted trees or combinations of boosted-bagged trees [104]. This allows more flexible, non-linear feature functions to be incorporated. An example of such a feature function is shown in Figure 2.5a. Similar to the network models, such feature functions can reflect interrelations between nodes, as illustrated in Figure 2.5b.

However, there is still a significant gap between the flexible GAMs and full-complexity models, such as ensembles of trees, regarding accuracy [104]. The main reason for this limitation is that GAMs take into account only univariate terms without considering any interaction (or interrelationship) between features. To deal with this drawback, a more advanced method was developed, called Generalized Additive Models plus Interactions (GA^2Ms), which additionally incorporates pairwise interactions between features [105]. Similar to GAMs, this model describes any variable $x_{1,v,t}$ according to Equation 2.4 in the following form:



(a) (b)

Figure 2.5: (a) Non-linear relationship of input $x_{1,1,t-1}$ (or $V1$) to output $x_{1,2,t}$ (or $V2$). (b) The non-linear relation is again reflected in the directed connection between $V1$ and $V2$.

$$g(x_{1,v,t}) = \sum_j f_j(x_{1,j,t-1}) + \sum_{l \neq j} f_{lj}(x_{1,l,t-1}, x_{1,j,t-1}) \tag{2.4}$$

where f_{lj} is the function for feature interactions for all combinations of variables (l and j). This model can still be interpretable, using heat maps for representing the pairwise features' interactions, as well as accurate, reaching the performance of the state-of-the-art ML models. An example of a pairwise features interaction is presented in Figure 2.6.

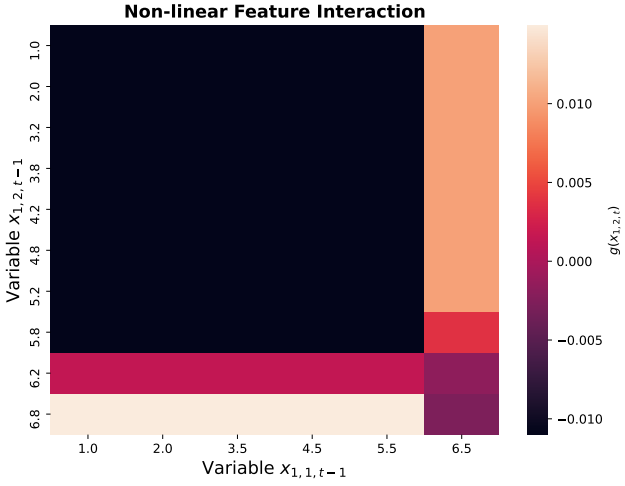


Figure 2.6: Pairwise feature interaction between $x_{1,1,t-1}$ and $x_{1,2,t-1}$, colored by the effect on the output $g(x_{1,2,t})$.

This presents a heatmap of 2 dimensions, with each axis representing the values of a feature. The heatmap shows the scores $g(x_{1,2,t})$ associated with a combination of feature values for features $x_{1,1,t-1}$ and $x_{1,2,t-1}$. Higher scores indicate a greater probability of predicting the positive class (in the case of a binary classification setting) when those specific combinations of feature values occur.

In this work, a fast implementation of the GA^2Ms algorithm is used, called Explainable Boosting Models (EBMs), which is part of Microsoft's open-source Python package, called InterpretML [106, 107]. The EBMs' learning process makes use of the gradient boosting algorithm with shallow tree ensembles, as described in detail in Figure 2.7. At each boosting round, a tree is built on a single feature and its residuals are used for training the tree of the following feature. This is repeated for all different features. After several boosting rounds, each feature's trees of all rounds can be combined, leading to tree ensembles as the final features' representation. Typically, the number of boosting rounds is initially set and controlled by a predefined tolerance threshold. The additive property of trees for each feature is illustrated in Figure 2.8. On top of this, functions for pairwise features' interactions can be additionally incorporated. The FAST method is used to detect and rank features' interactions, subsequently keeping the most significant ones [105]. This prevents an extensive checking of all possible combinations [105]. Similarly, the same training process is performed for the specified pairs.

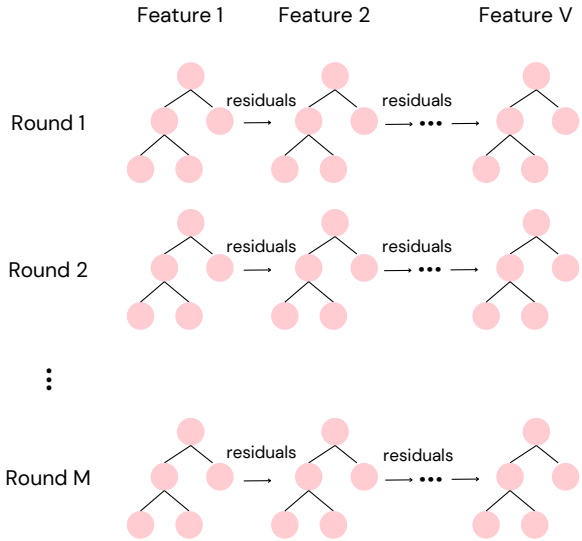


Figure 2.7: The learning process of EBMs: Feature functions are iteratively updated at each boosting round to minimize the prediction error (residuals).

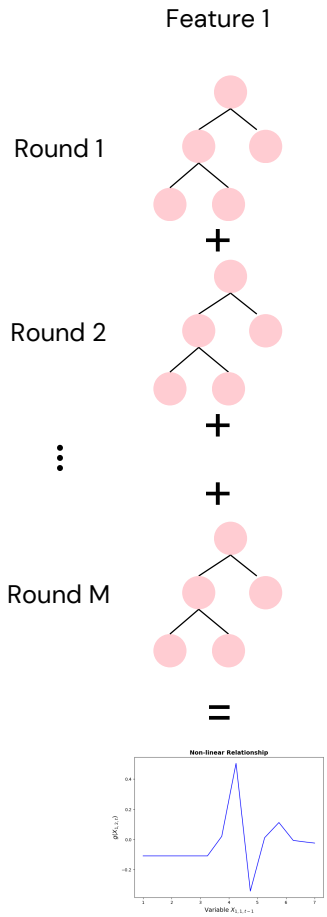


Figure 2.8: Interpretability of Feature 1 by summing the learned trees of all M rounds.

2.5. OUTPUT TASKS

Up to this point, the focus was to approximate the network approach using more advanced non-linear models, showing that more detailed and accurate information can be derived regarding the underlying processes. However, deciding on the type of model depends on the output task that needs to be addressed.

In the network approach, the applied models have been mainly discussed in a multivariate regression setting. Nevertheless, advanced models could be built targeting different output tasks. For instance, the modeling approach varies depending on the type of output variable being targeted. Therefore, in this section, an overview of all the examined output tasks is presented, each providing insights into different aspects of mental disorders. The tasks are also summarized in Figure 2.9.

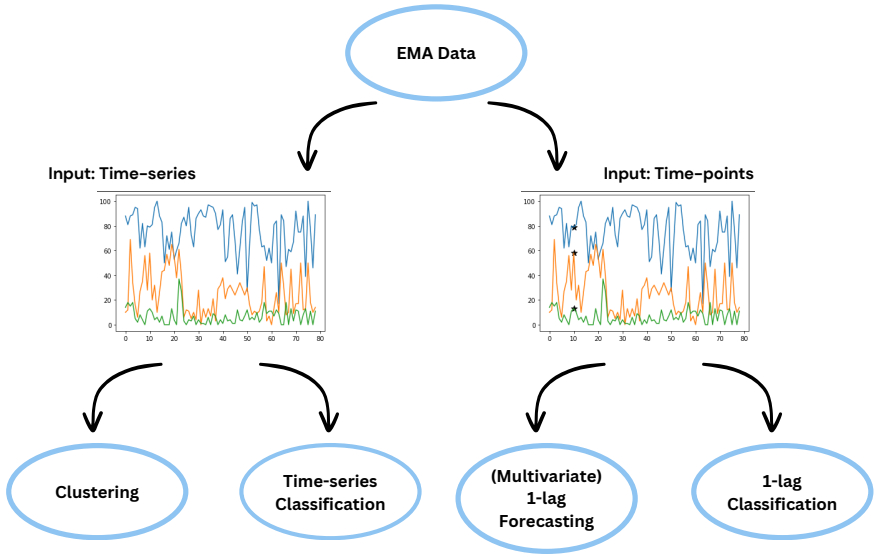


Figure 2.9: Overview of the examined output tasks in relation to the input (time-series or time-points).

2.5.1. 1-LAG BINARY CLASSIFICATION FOR EVENTS PREDICTION

Classification models are developed to predict a particular categorical outcome, such as an event or variable. In this context, the focus is on on binary classification models predicting the occurrence or not of an event, such as elevated negative emotions, within a given time frame (e.g., 1-lag or next time-point). To create binary outputs from EMA variables, cut-offs are applied to transform continuous or ordinal measurements into binary indicators of event occurrence. These models use the previous states of EMA variables to identify patterns and risk factors associated with one step ahead of the future events (variables) of interest. When many variables are involved, this is referred to as a 1-lag Multivariate Binary Classification.

2.5.2. 1-LAG MULTIVARIATE FORECASTING

Multivariate regression models are developed to forecast the future values of multiple (or all) EMA variables, such as mood states (e.g., positive and negative affect). In the case of 1-lag forecasting, the goal is to predict the values of the next time-point; however, these models can also be extended to forecast multiple future time points if needed. This approach provides an overall forecast of an individual's mental health status, as represented by all measured variables. In case of a focused task, this scenario can be simplified from multivariate to one EMA variable, referred to as a 1-lag forecasting task.

2.5.3. TIME-SERIES CLUSTERING

Given the high heterogeneity across individuals collected in an EMA study, further analysis is necessary to uncover subgroups with homogeneous EMA patterns or profiles. In particular, EMA data can be used to uncover patterns and similarities in their data, leading to meaningful groups of similar individuals. To address the task of individual grouping, clustering algorithms are applied to EMA data. Clustering is an unsupervised method, meaning that there are no available labels (clusters) to optimize the learning process. Instead, it only relies on similarities of various characteristics extracted from individual time-series. Effective clustering plays a significant role in understanding how similar patterns form distinct EMA profiles, which can be useful for targeted interventions and personalized treatment plans.

2.5.4. TIME-SERIES CLASSIFICATION FOR CLUSTERING EXPLANATIONS

Given that there is no definitive answer about the true clustering results, evaluating all possible groupings is infeasible due to the vast computational complexity involved. Instead, smart algorithms have been developed to identify high-quality clustering solutions. Besides some evaluation measures examining the clustering quality, another approach can be also given by using a prediction model to evaluate and explain some "optimal" (meeting quality criteria) clustering results (cluster labels). Specifically, classification models are used to classify individuals into their respective cluster labels. In this case, classification models are used in a different setting than the first one (1-lag Binary Classification), where instead of inputting the EMA values of one time-point, the input is the whole MTS. Therefore, appropriate classification models should be selected that are capable of handling MTS data as input. Depending on the number of clusters involved, this can be split into binary (2-cluster) or multi-class (multi-cluster) MTS classification. Such models are used to provide explanations about the common characteristics of each cluster, that is the common features of individuals within each group.

2.6. DATASETS

In this section, the real-world EMA datasets that are explored in this thesis are presented. Because the availability of open-source datasets is limited, additional synthetic data are generated and used for evaluation, but these are described in the corresponding chapters regarding the targeted task.

All real-world datasets' parameters are briefly reported in Table 2.1 and more extensively described as follows. Additionally, Figure 2.10 presents an illustrative example of a single individual from each of the three examined datasets, showing the patterns of three distinct variables over time. It is important to note that the individuals depicted are all different,

with each participating in only one of the studies or datasets. Additionally, while the variables are labeled similarly, they can differ between datasets.

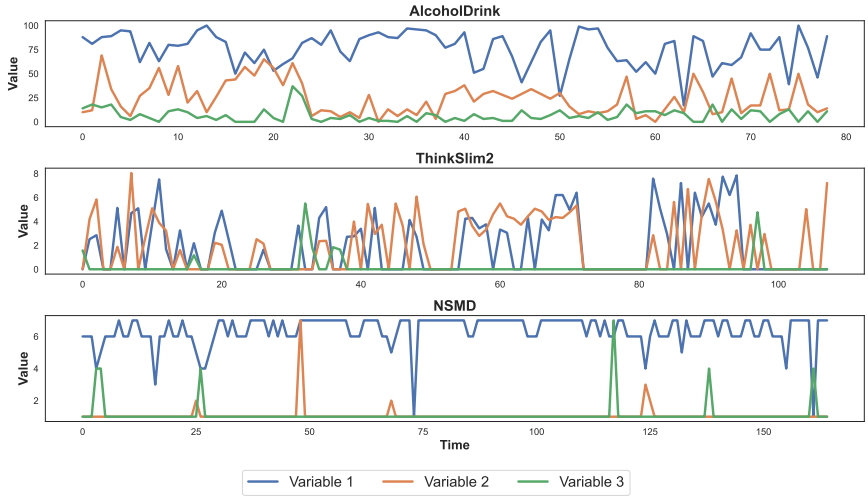


Figure 2.10: A time-series example of the 3 examined datasets. Each subplot corresponds to an individual (showing 3 variables) of one of the datasets AlcoholDrink, ThinkSlim2 and NSMD. The 3 variables differ between datasets.

2.6.1. ALCOHOLDRINK DATASET

The AlcoholDrink dataset is a real-world and open-source dataset, obtained by a study described in [108]. It was a 2-week collection of data from 33 individuals through mobile notifications. The captured variables included positive and negative emotions, drinking cravings and expectancies, perceived alcohol consumption, impulsivity, as well as social context. All these variables were measured on a visual analogue scale (VAS) from 0 to 100.

During data preparation, each participant's EMA data was analyzed separately. Each individual dataset was assessed for the frequency of daily observations as well as the frequency and distribution of the outcome events. First, individuals having very few observations per day or in total were removed, with the threshold being set at 80% compliance. The number of individuals retained was 26. For a more detailed exploration of the data's characteristics, additional figures can be found in the supplementary material of this chapter, specifically in Figures 2.13 and 2.14.

Table 2.1: Characteristics of the examined real-world datasets regarding the number of individuals, features and time-points (mean and standard deviation across all individuals) after the initial preprocessing steps.

Dataset	#Individuals	#Features	#Time-points
AlcoholDrink	26	15	92.0(48.1)
ThinkSlim2	65	37	119.3(54.5)
NSMD	187	12	167.1(27.4)

2.6.2. THINKSLIM2: HEALTHY/UNHEALTHY (HU) EATING DATASET

The ThinkSlim2 dataset is a real-world dataset, obtained by a study described in more detail in [109]. The dataset consisted of information collected from 135 overweight individuals throughout the day for eight weeks via a mobile application.

Each participant's EMA data was prepared for analysis separately. After checking data compliance, 76 individuals were retained for further analysis. Additionally, only a subset of the captured variables was selected again. This was necessary because the majority were categorical variables and some categories were infrequently represented in the dataset, reducing their utility for robust analysis. Therefore, their distribution was evaluated. Infrequent variables were then removed to finally retain only the 13 informative ones.

The final variables included various positive and negative emotions, location, activity, social context, and type of consumed food. The emotion-related variables were measured on a scale from 0 to 10. All the other variables (e.g., activity and location) were categorical, including a long set of predefined choices for each one. For example, the categorical variable "activity" was summarized into 11 separate categories. Following, these needed to be transformed into numerical variables suitable for machine learning algorithms. A common approach for this transformation is one-hot encoding, converting each category of a categorical variable into a new binary variable. In this encoding scheme, if an instance (or sample) belongs to a particular category, the corresponding variable takes the value 1, while all other columns are 0. This strategy leads to a significant increase in the total number of variables used in the analysis, from 3 categorical to 21 binary variables. Given the complexity introduced by the large number of binary variables, emotion variables were also split in 2 (high and low) or 3 categories (high, medium, low). This categorization helps simplify the overall data structure. Additional figures illustrating the dataset's characteristics are available in the supplementary material for this chapter, specifically in Figures 2.15 and 2.16.

2.6.3. NSMD DATASET

The NSMD dataset is a real-world EMA dataset collected as part of the New Science of Mental Disorders (NSMD) project. It is a study in which data on mental health was collected from students in Dutch universities [36, 110]. A set of 288 individuals were monitored eight times a day for 28 days, leading to a total of 224 time-points per individual. However, due to missing data, not all individuals had sufficient data for analysis. After setting the compliance to 50%, that is a minimum of 112 time points, 187 individuals were included in the analysis. At each time-point a set of 65 psychopathology-related variables was assessed. Most variables were rated on a 7-point Likert scale. According to domain experts, to reduce the number of examined features, some variables were not included in the analyses, either due to limited within-person variance or due to relatively less relevance, while some connected EMA items were merged and averaged together. For instance, some variables, such as positive affect (PA) and negative affect, were averaged across all their relevant variables (e.g., happy, calm, etc. and sad, angry, etc., respectively). An overview of the examined variables is given in Table 2.2.

Table 2.2: Overview of the EMA variables of the NSMD dataset.

Variable	Description	Raw or Averaged EMA
PA	Positive Affect	Averaged
NA	Negative Affect	Averaged
Som_neg	Somatic Negative Affect	Averaged
Self_esteem	Self Esteem	Averaged
Enj_act	Enjoyment of activities	Averaged
Enj_social	Enjoyment of social encounters	Averaged
Crave_Food	Craving food	Raw
Crave_Other	Craving other	Raw
In control	In control	Raw
Concentrated	Concentrated	Raw
Worried	Worried	Raw
Impulsivity	Impulsivity	Raw

Before analysis, further exploration of the data's characteristics is necessary to get a better understanding of each variable. First, examining the distributions of individual variables provides useful insights regarding their spread and counts. In Figure 2.11, histograms are used to assess the distributions of each variable separated regarding the whole dataset. Similarly, the distribution of the variables could be investigated at an individual level.

Additionally, estimating different statistical properties, such as mean values, standard deviation and variance of each variable could provide

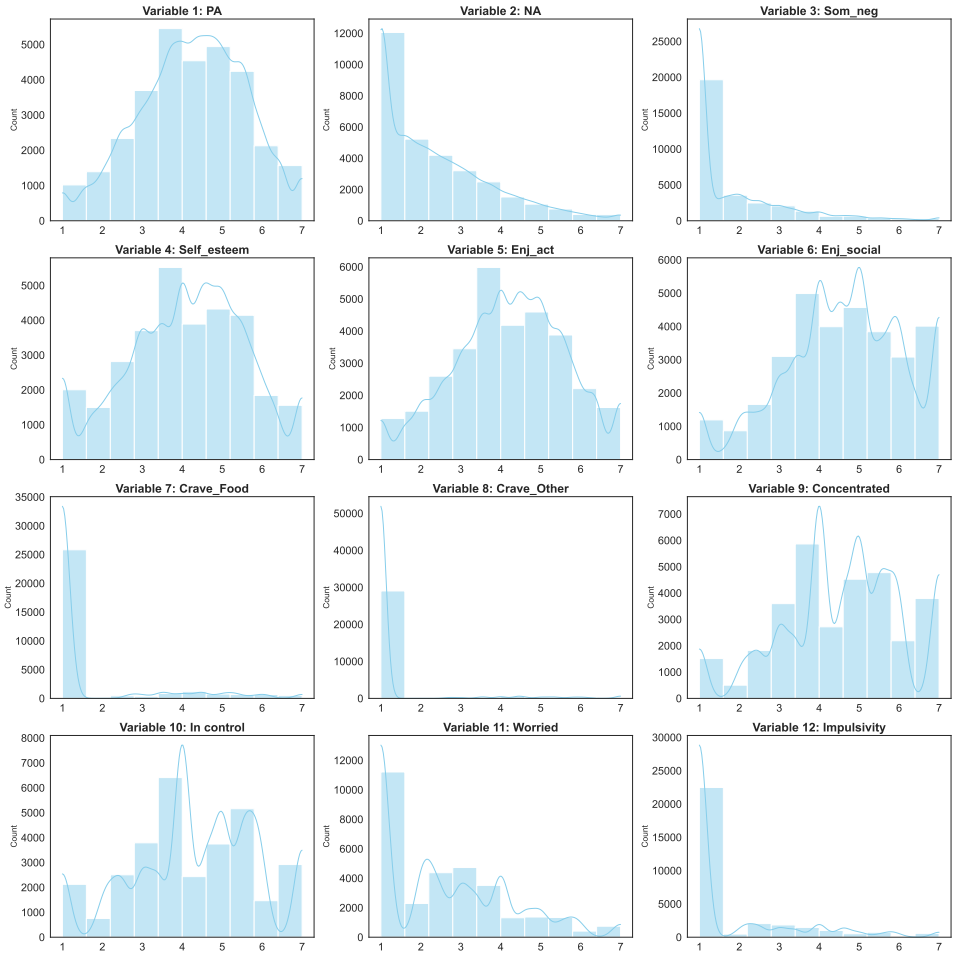


Figure 2.11: NSMD Dataset: Histograms of the frequency and spread of each variable.

more insights about the examined data. As shown in Figure 2.12, visual representations through boxplots can summarize these within-individual statistical properties across all individuals separately for each variable. In particular, variance is an important property indicating the level of variability within the dataset. Such exploration facilitates identifying uninformative variables and outliers. As a next step, further data exploration could follow, including variable correlation and pairwise associations among them.

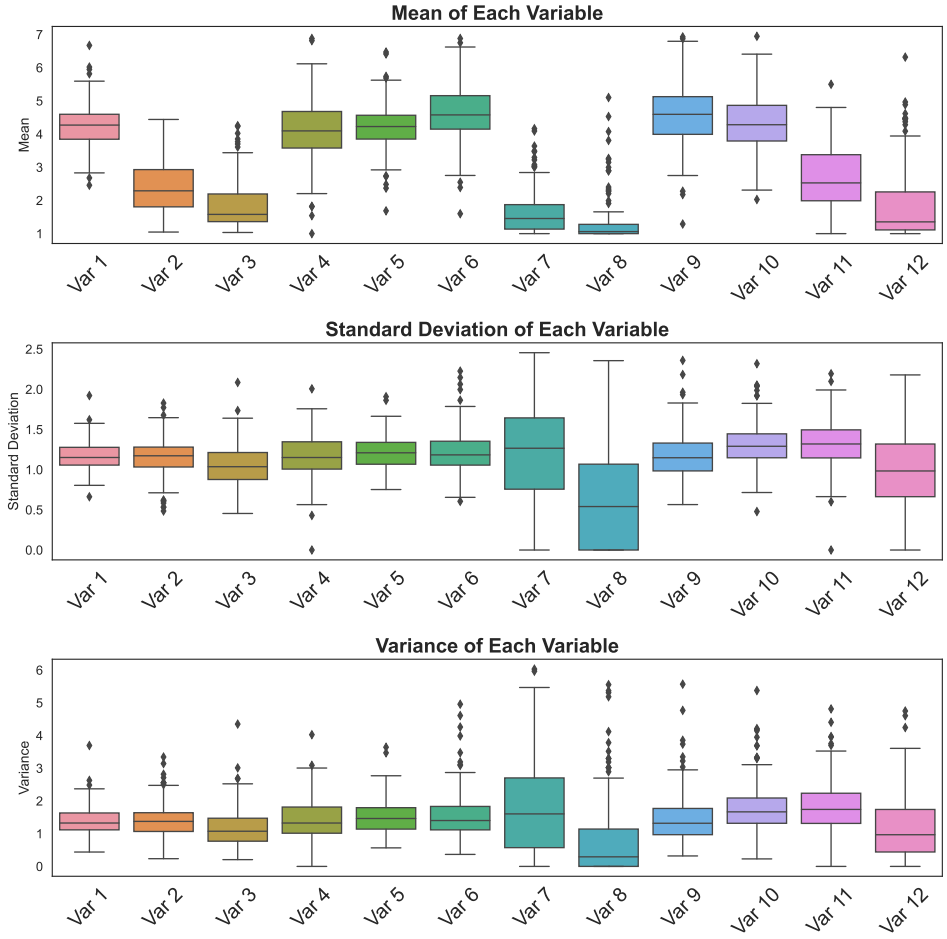


Figure 2.12: NSMD Dataset: Distributions regarding 3 statistical properties (mean, standard deviation and variance) of each variable.

2.7. CONCLUSIONS

In this chapter, the background information of this dissertation is provided. Beyond describing the complex structure of the multivariate time-series EMA data, we followed with more methodological aspects of the research. Starting from the linear VAR models, we explore the challenges arising and the opportunities emerging when applying more advanced non-linear models. The focus is on the interpretable non-linear EBM model that is applied in the following chapters (Chapter 3, 5 and 7) of this dissertation. However, it is important to note that different models can be applied depending on the task that we need to address. In detail, an overview of all the investigated output tasks along with the 3 utilized

real-world EMA datasets is presented. To summarize, the connection of all examined datasets and output tasks of the main chapters (3-7) is depicted in Table 2.3.

Table 2.3: Summary of the datasets and outputs tasks examined in each of the main chapters (3-7) in the dissertation.

Chapter	Dataset	Output Task
3	AlcoholDrink ThinkSlim2	Binary Classification
4	NSMD	Clustering
5	NSMD	Clustering
6	NSMD	MTS Classification
7	NSMD	Binary Classification

2.8. SUPPLEMENTARY MATERIAL

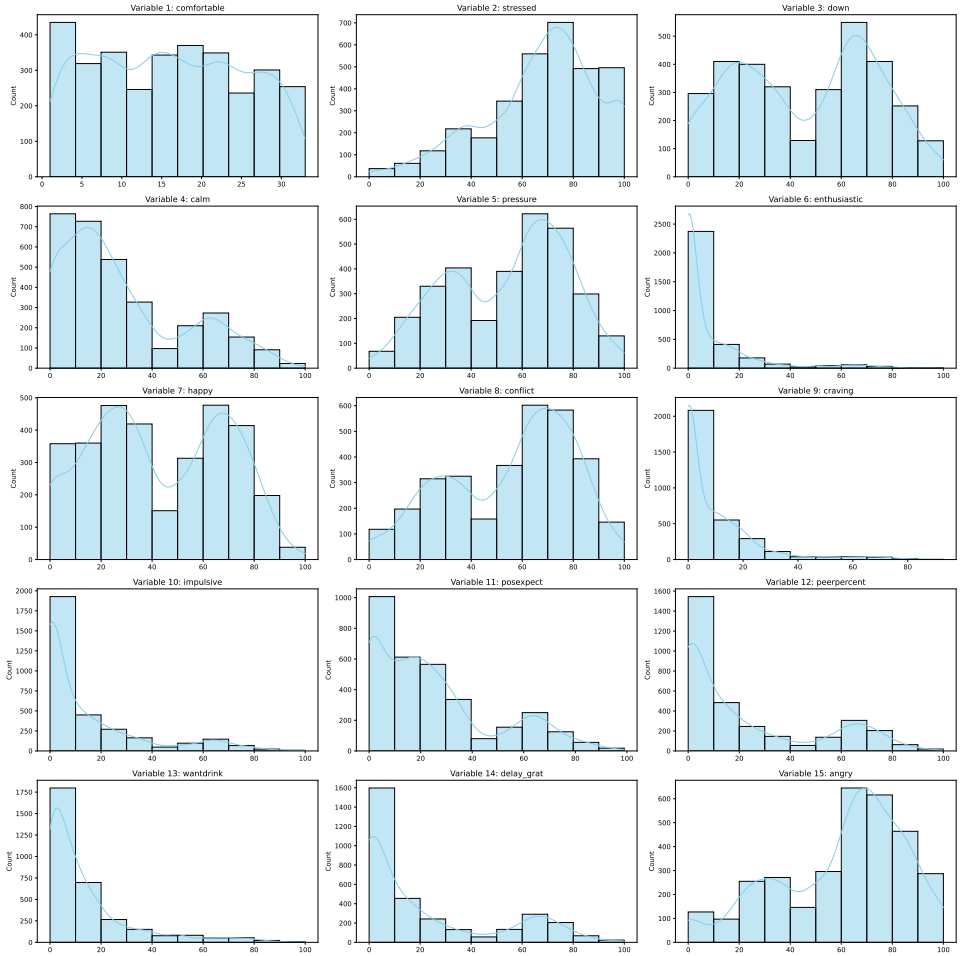


Figure 2.13: AlcoholDrink Dataset: Histograms of the frequency and spread of each variable.

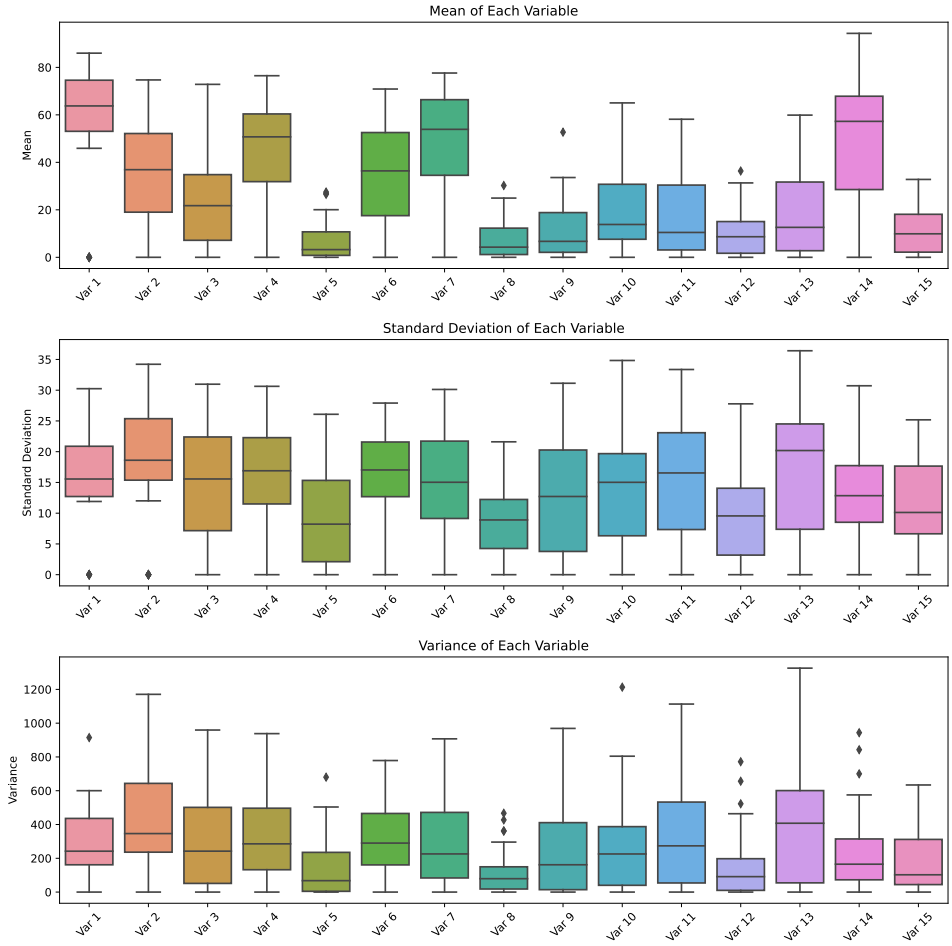


Figure 2.14: AlcoholDrink Dataset: Distributions regarding 3 statistical properties (mean, standard deviation and variance) of each variable.

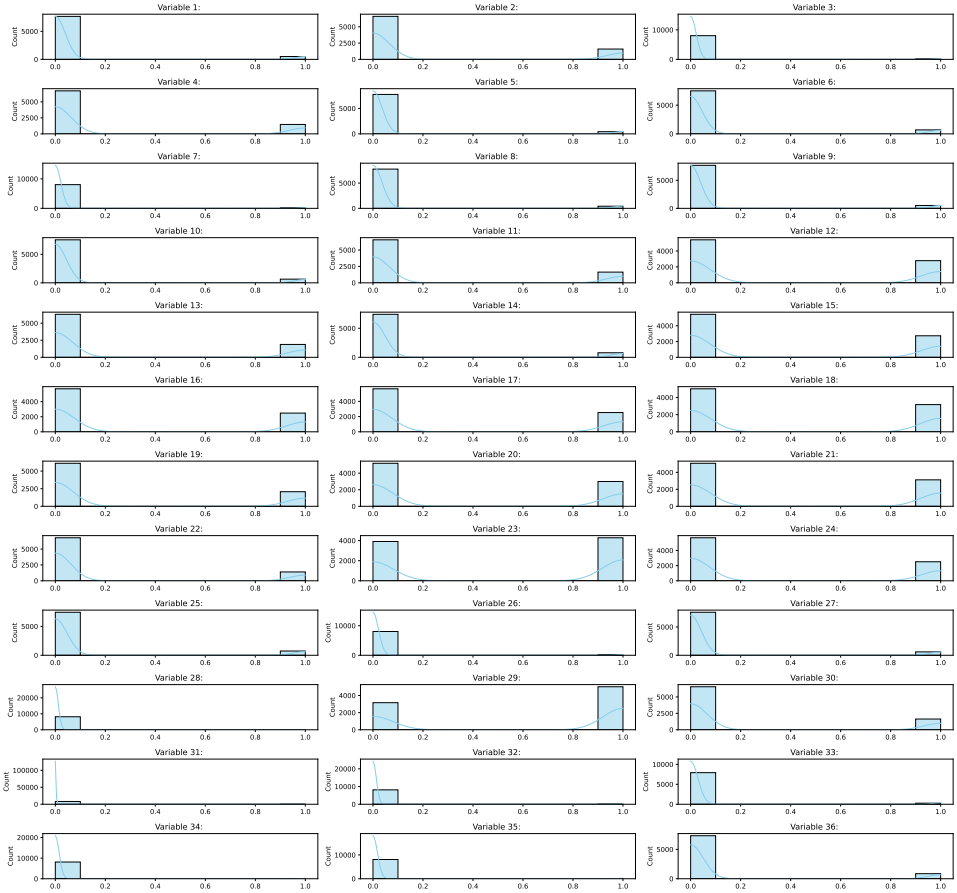


Figure 2.15: ThinkSlim2 Dataset: Histograms of the frequency and spread of each variable. The categorical variable has been removed for consistency.

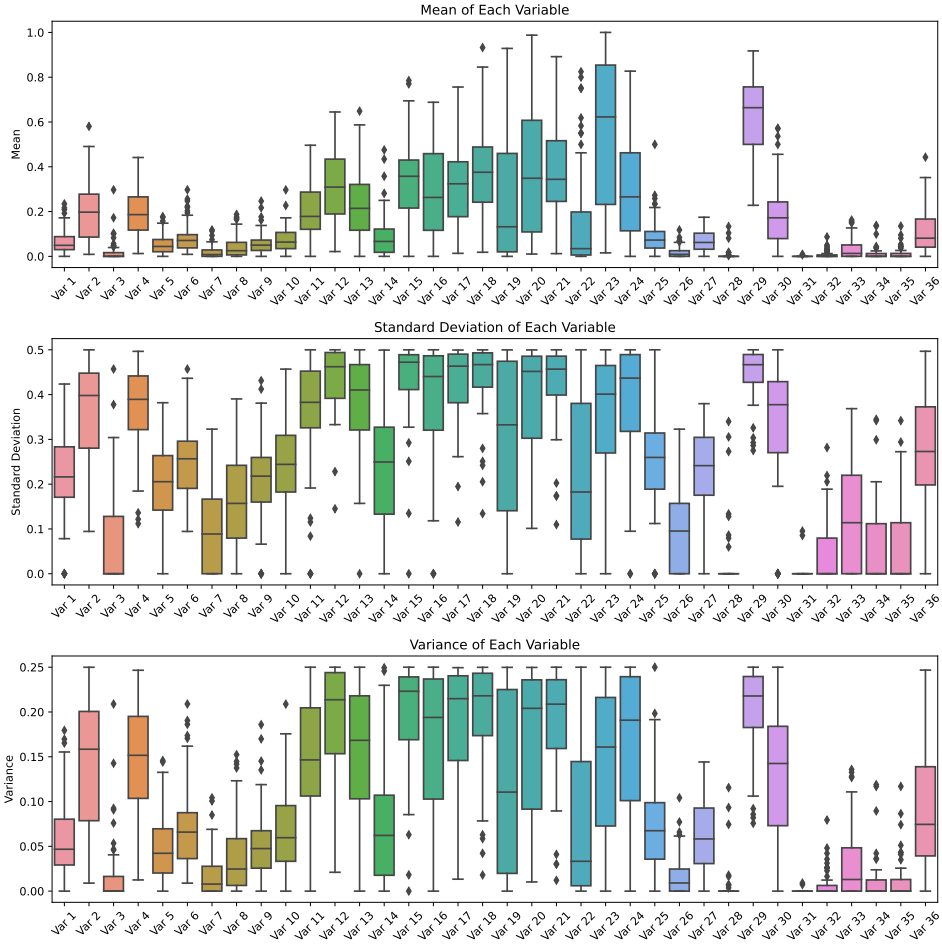


Figure 2.16: ThinkSlim2 Dataset: Distributions regarding 3 statistical properties (mean, standard deviation and variance) of each variable. The categorical variable has been removed for consistency.

3

COMPARE IDIOGRAPHIC AND NOMOTHETIC APPROACHES

While previous research on EMA data for mental disorders was mainly focused on network models and linear individual regression-based approaches, this chapter goes a step further by exploring the use of non-linear ML models in EMA classification problems. ML models can enhance the ability to accurately predict the occurrence of different psychopathology-related variables or events by recognizing complicated patterns between variables in data. To evaluate the efficacy of non-linear models, relying on ensembles of trees, their performance is compared to linear models using imbalanced synthetic and real-world EMA datasets. Moreover, apart from personalized approaches, nomothetic or group-based prediction models, which integrate data from more than one individual, are examined. Such approaches are also likely to offer an enhanced performance compared to individual or personalized models.

Parts of this chapter have been published in

- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Using explainable boosting machine to compare idiographic and nomothetic approaches for ecological momentary assessment data". In: International Symposium on Intelligent Data Analysis. Springer. 2022, pp. 199–211

3.1. INTRODUCTION

Recent technological and methodological advancements in EMA have significantly renewed the research interest in psychology and psychiatry. Particularly, through EMA, a large amount of personalized data has become available, providing the means for further exploring mental disorders [99]. With this large data availability, there has been a significant focus on developing statistical methods to model psychopathology [57]. Some practical applications of such models could be to predict the course of illness, determine treatment response or develop tailored psychiatric interventions [2].

Based on the literature, EMA time-series data have been mostly analyzed by applying a multivariate regression-based approach [57, 90]. More specifically, the most popular class of time-series models is the Vector Autoregressive (VAR) model that aims at estimating the dynamical interactions between all the measured variables (i.e., network structures) [46]. However, the fact that these models can only estimate linear statistical relationships can be a significant challenge for mental disorders, where the involved interactions are likely to be quite complex. When many symptoms or variables are involved in the course, these are more prone to interact in a non-linear fashion with each other. Thus, linear models seem insufficient to uncover the possible non-linear interactions and describe precisely the real complex nature of mental disorders.

A promising approach that can learn such complex and higher-order interactions of symptoms involves leveraging non-linear machine learning (ML) models [60]. ML models can enhance the ability to accurately predict the occurrence of different EMA variables or events by recognizing complicated patterns or relations between variables in existing data.

This chapter addresses three key research objectives, identified as RQ1, RQ2, and RQ3 in Section 1.7. It explores a spectrum of predictive approaches, from idiographic (personalized or individual) approaches under RQ1, through nomothetic (group-based) approaches partially covering RQ2, to a more advanced integrative approach partially addressing RQ3, that balances personalized and group-based strategies. First, according to the idiographic approach, personalized models are typically applied, as there are possibly different underlying mechanisms that drive future behavior in each individual. Thus, different non-linear interpretable models are evaluated in terms of performance to test whether they are superior to baseline linear models. Second, we should acknowledge that shared influences among different individuals may provide a complementary predictive utility. Therefore, prediction models are applied in a nomothetic approach showing that integrating data of more than one individual in a single model could also accurately predict future outcomes at an individual level [112]. The third approach balances the strengths of both idiographic and nomothetic approaches through a more advanced two-step process. Specifically, this method is designed to incorporate the insights gained from the nomothetic model (step 1) into the personalized models of each individual (step 2). By examining

these methodologies, this chapter contributes to developing more reliable models that could facilitate a better understanding of individual behaviors and interactions, at both personalized and group levels.

3.2. METHODOLOGY

This section provides an overview of the approaches used in this chapter. It explores a range of predictive approaches, starting with idiographic (personalized or individual) methods, moving through nomothetic (group-based) approaches, and progressing to a more advanced integrative approach, which balances personalized and group-based strategies.

3.2.1. IDIOGRAPHIC (PERSONALIZED OR INDIVIDUAL) APPROACH

Based on the fact that mental disorders can be modeled as a complex system, we assume that the course of illness and EMA patterns differ remarkably across individuals [113]. Most individuals suffering from the same disorder are likely to exhibit different symptoms, so different mechanisms possibly influence and drive future behavior [114]. Therefore, it is proposed that each individual should be examined separately using personalized prediction models [45], as illustrated in Figure 3.1.

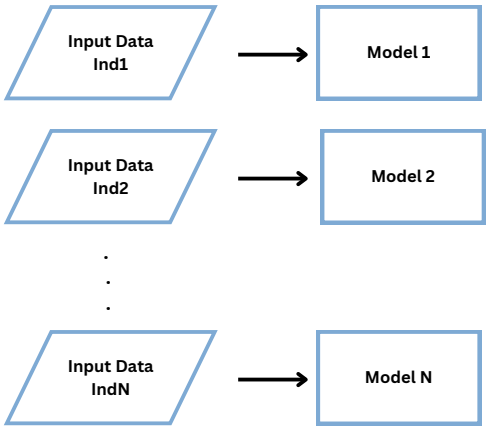


Figure 3.1: Idiographic Approach: The data from each individual (e.g. Ind1) is used to train a model (Model 1).

As already discussed in Section 2, starting from the widely used linear models, the progression to more sophisticated models allows for non-linear representations of features and interactions of features [105]. A flexible solution was given by developing the Generalized Additive Models plus Interactions (GA^2Ms), enhancing model complexity, accuracy and interpretability [106]. In this work, a fast implementation of the GA^2Ms algorithm is used, called Explainable Boosting Models (EBMs),

which is part of Microsoft's open-source Python package, called InterpretML [106]. Because EBMs is a relatively novel method, its performance is evaluated by comparing it to other full-complexity ML models, such as Extreme Gradient Boosting (XGBoost), Gradient Boosting Trees (GradBoost) and Random Forest (RF) [115]. Afterwards, non-linear models are also compared to linear models, such as Logistic Regression (LogReg) and Support Vector Machines (SVM), using a linear kernel.

3.2.2. NOMOTHETIC (GROUP-LEVEL) APPROACHES

Although personalized models capture the unique patterns of each individual, commonalities among different individuals may provide complementary predictive utility [116]. Thus, group-level prediction studies are also likely to offer an enhanced individual performance. Especially, in the case of more advanced ML models, incorporating more data could be of more help, compared to the traditional linear models. This approach could have a clear advantage in uncovering potential complex hidden relationships between variables. More specifically, two nomothetic approaches are investigated, the using-all-data (using_all) and the more advanced Knowledge Distillation (KD). These are described as follows.

NOMOTHETIC APPROACH: USING-ALL-DATA (USING_ALL)

The most trivial way of integrating data of more than one individual in a model is to concatenate the data points of all individuals together in a single dataset. The augmented or aggregated dataset is then used to construct a group-based model. Specifically, when all data from individuals within the same data collection effort is used, this approach is referred to as a population model. Such models produce generalizable predictions that can be relevant to a wider range of individuals. For example, a group-based model can be applied to new individuals who have not been included in the training process of the model. An additional scenario falling into that category refers to individuals for whom personalized modeling is impractical. This might occur due to a lack of adequate time points to train a robust personalized model. Therefore, a group-based approach provides valuable predictions and insights for these individuals, reaching a broader target population, with limited minority classes points or even total data points.

NOMOTHETIC APPROACH: KNOWLEDGE DISTILLATION (KD)

The second proposed approach is based on the Knowledge Distillation method, also known as the teacher-student framework [74]. In this approach, information extracted from a larger, more general (teacher) model is used to enhance a smaller, more specialized (student) model. More specifically, in our context, information from training a group-level (teacher) model using all data can be utilized in a personalized (student)

training concept. Thus, the KD method effectively builds on the group-level modeling process established by the using_all approach, allowing the specialized student models to benefit from the broader learning insights of the teacher model.

The approach of Knowledge Distillation was originally developed to fill the gap between the expressive power of the large models and the learnability of the smaller models in Neural Networks (NNs) [74]. While large NN models are known for their power and success in capturing complex patterns in data, these are often computationally expensive, overparametrized and too generalized to learn and extract insights regarding targeted parts of the data [117]. In practice, KD involves a 2-step training process, where a small NN is trained after incorporating additional information from a larger and more complex NN. Comprehensive overviews of this technique can be found in the survey papers [118, 119].

However, the aforementioned gap does not only exist in NNs but also in other machine learning methods, such as using the tree-based models described above [120, 121]. So, the distillation method using information extracted from larger models can be further exploited in non-NN models.

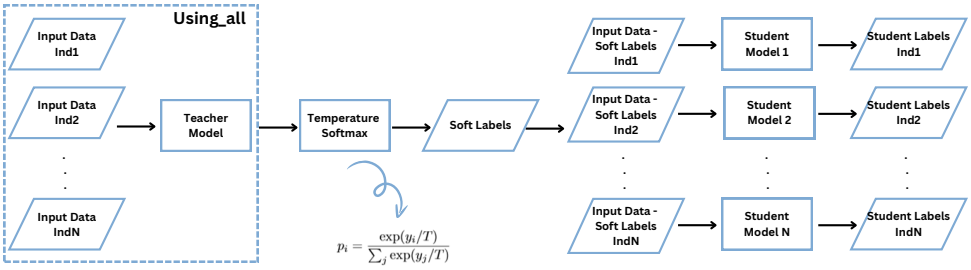


Figure 3.2: Overview of the proposed Knowledge Distillation method adapted to EMA, using group-level (using_all) models as teacher models and individual models as student models.

Inspired by the original concept, the proposed Knowledge Distillation method in our case is illustrated in Figure 3.2. The first part consists of the using_all approach, where the teacher model is trained on data from all individuals in a classification task. Following this, the outcome information from this model is used to train personalized student models for each individual separately. Instead of using the ground-truth (hard) outcome labels, the additionally-gained information is achieved through probabilities derived from the output of the softmax function on the teacher's logits (raw model's outputs). However, in classical ML models, we can work directly with probabilities instead of logits. Then, we adjust the probabilities by treating them as pseudo-logits, applying temperature softmax to smooth and calibrate the output, making the predictions less overconfident and more representative of true likelihoods. Specifically, the probabilities y_i of the training samples, referring to each output class i, are softened using a temperature softmax function. The smoothened

class probabilities p_i (soft labels) are calculated according to Equation 3.1 and Figure 3.3, where T is the temperature hyperparameter.

$$p_i = \frac{\exp(\frac{y_i}{T})}{\sum_j \exp(\frac{y_j}{T})} \tag{3.1}$$

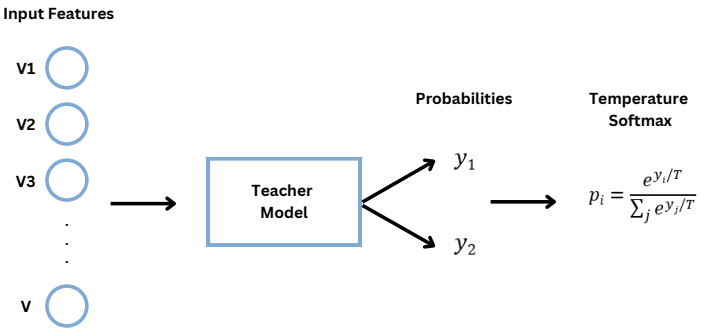


Figure 3.3: The proposed Knowledge Distillation method: After inputting each data sample to the teacher model, the extracted probabilities y_i are used to the temperature softmax function. The produced p_1 or p_2 are the labels (in the case of a binary classification) for the student models.

The T hyperparameter plays an important role in smoothing the distribution of the outputs, which is necessary to distill as much information as possible. When $T = 1$, p_i refers to the typical probabilities derived from a softmax function. However, in cases where the correct label is assigned with a very high (close to 1) probability, this does not provide much additional information than the ground-truth hard labels. To prevent such cases, a temperature $T > 1$ is applied [119]. An example of the difference between hard and soft outputs, regarding both $T = 1$ and $T > 1$, is given in Table 3.1.

This higher temperature setting effectively softens the probability distribution. However, selecting the appropriate degree of smoothing affects how soft the probabilities become. For example, a moderate increase in T, such as $T = 5$, could provide an informative distribution, retaining a reasonable difference in soft labels. On the contrary, a higher T, such as $T = 100$, smoothens the differences even more leading to almost equal probabilities. Therefore, the choice of T should be carefully calibrated.

Subsequently, during the second part of the process, the smoothened probabilities (or soft labels) are used as labels for the personalized student models. Compared to conventional personalized training that uses hard labels, distillation can provide additional useful information from other individuals to improve the personalized models.

Table 3.1: An example of how binary (hard) outputs are transformed to soft labels using different T values in temperature softmax. When $T = 1$, soft labels refer to typical probabilities, whereas when $T > 1$, to smoothened probabilities used in KD.

Hard Labels	Soft Labels	
	$T = 1$	$T > 1$
0	0.079	0.492
0	0.222	0.496
0	0.450	0.466
1	0.998	0.516
1	0.932	0.509
1	0.661	0.501

3.3. EXPERIMENTAL SETUP

3.3.1. EMA DATASETS

EMA data is organized in a hierarchical structure for each individual (discussed in Section 2.2), with observations collected multiple times a day for a predefined period of several weeks. The total number of observations as well as the collection period can be different among individuals because some may experience difficulties in following the schedule of the surveys. The characteristics of all datasets (presented in detail in Section 2.6) used in this chapter, following task-specific preprocessing for 1-lag or next time-point classification, are summarized in Table 3.2.

SYNTHETIC DATASETS

Due to a lack of access to large EMA datasets, we follow a simple method for generating random EMA datasets. This method addresses the challenge of limited data availability by creating randomized datasets that simulate real-world conditions. Each synthetic dataset is designed to consist of the feature vectors and labels for each simulated patient, aiming at a 2-class classification problem. It is also commonly noticed that medical-related EMA datasets, as well as the following examined real-world datasets, are characterized as imbalanced. This means that the majority of samples belongs only to one class, whereas much fewer to the other class. Thus, in this case, the ratio of samples assigned to the two classes is $0.7 : 0.3$ in the synthetic datasets as well.

Furthermore, the datasets must be created in a way to be structurally similar to the real EMA data. First, these must incorporate multivariate ordinal and categorical variables. This is a challenging issue, especially in high-dimensional datasets. The method for generating our feature vectors is based on sampling from a different random normal distribution for

Table 3.2: Characteristics of the examined datasets. For imbalance ratio and training/test sets, the mean and standard deviation values of all individuals are presented, after preprocessing.

Dataset	#Individuals	#Features	Imbalance Ratio	#Training Data	#Test Data
Synthetic	20, 50, 100	25, 60	2.33	35, 70, 210	15, 30, 90
AlcoholDrink	24	15	8.45(5.45)	72.83(12.02)	31.87(5.24)
ThinkSlim2	57	37	5.82(3.25)	86(38.72)	37.51(16.68)

each one. After sampling, these continuous values are transformed into ordinal features by applying an equal-width histogram binning, which divides each distribution into intervals of equal width. This process results in a random selection of six or two distinct ordinal values per feature. When six values are used, these resemble the ordinal scale of EMA questions, whereas categorical EMA questions are typically encoded as binary variables, represented by two distinct values.

It is also often necessary to impose some flexibility on the data variables, such as noise. Noise can be added to both output labels and feature vectors. In this setup, a small amount of noise is introduced by randomly reassigning 20% of the labels to samples and shuffling the values of 20% of the features. Additionally, various options for other characteristics of the synthetic datasets, such as the number of individuals, features, and samples, are evaluated.

DATASET: ALCOHOLDRINK

This first real-world dataset is the AlcoholDrink dataset, described in [108]. Regarding the output variable, the aim of this prediction was the occurrence or not of drinking events at the next time-point. So, a positive label was assigned to each sample when the number of alcoholic drinks at the next time-point was one or higher.

DATASET: THINKSLIM2

The second real-world dataset, ThinkSlim2, is larger and more challenging. It was obtained by a study described in more detail in [122], [109]. Regarding the output variable, the examined scenario was aiming at predicting the next healthy or unhealthy eating event. So, a healthy or unhealthy label was assigned to each sample according to the type of food consumed at the next time-point.

3.3.2. DATA PREPARATION

For each dataset, each participant's EMA data was prepared for analysis separately. These were assessed for the frequency of daily observations as well as the frequency and distribution of the outcome events. For

instance, in the case of the AlcoholDrink dataset, the counts of the outcome events (labels) are shown in Figure 3.4.

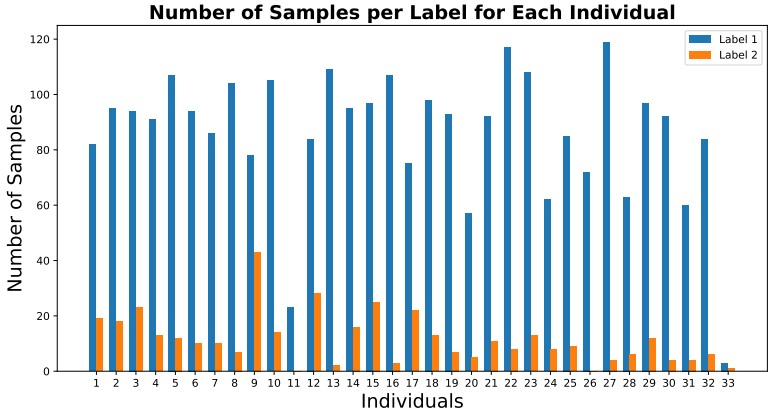


Figure 3.4: AlcoholDrink Dataset: Counts of outcome labels (Label 1 and Label 2) across Individuals.

Additionally, because of the final goal to predict (or classify) the next time-point event, consecutive data points had to be collected. For example, for each data point, if the following one (collected within the next 2 hours) was absent then we could not retrieve its prediction target and eventually it was also considered as missing. That way, some individuals were found to have so few outcome events of the minority class that subsequent cross-validation steps could not be conducted. So, these participants were also excluded from the final dataset. As a result, the number of retained individuals was 24 for the AlcoholDrink dataset having an average of 6.18 (std = 0.90) daily points and 57 for ThinkSlim2 with an average of 3.39 (std = 2.05) points.

As further seen in Table 3.2, data points of each individual were split sequentially at fixed time intervals into two datasets, a training and a test set, containing the first 70% and last 30% of the data points, respectively.

3.3.3. DATA ANALYSIS

IDIOGRAPHIC APPROACH

According to the idiographic approach, separate predictive models were applied to each individual, using various ML algorithms. The examined ML algorithms fall into the categories of linear or non-linear models. Regarding the linear models, Logistic Regression (LogReg) and SVM (using a linear kernel) were used, whereas for non-linear models EBMs, XGBoost, Gradient Boosting (GradBoost) and Random Forest (RF).

A necessary step is hyperparameter tuning, which frequently has a big impact on the model performance. In this setting, a time-series cross-

validation (CV) method, a variation of K-Fold designed for sequential data, was used, splitting the data into $k + 1$ folds. In each iteration, an increasing number of folds is used for training, with the next fold reserved for validation. This approach maintains the temporal order of the data while allowing for tuning of key hyperparameters in the tree-based methods, as illustrated in Figure 3.5.

The examined hyperparameters were different for each method. The number of pairwise interactions was important for EBMs. The learning rate, maximum depth, the minimum number of samples on a leaf, the gamma value and the fraction of the utilized features were examined in the case of XGBoost, whereas the learning rate and maximum depth when building the Gradient Boosting trees. In the case of Random Forest, the number of estimators, maximum depth and minimum number of leaf nodes were considered. All these combinations were exhaustively explored for each case using Grid Search and the one resulting in the best cross-validation score was retained for the following analysis.

The metric score of interest was the area under the ROC curve (AUC), measuring the true-positive rate and false-positive rate for the model's predictions using a set of different probability thresholds. AUC score was chosen for the prediction of both classes to be taken into account equally, regardless of the number of samples these classes contained. In other words, the prediction of samples belonging to the majority class should not play a more important role than predicting samples of the minority class. Similarly, other macro-average metrics, such as precision, recall, or F1 score, can also be used to achieve a balanced evaluation.

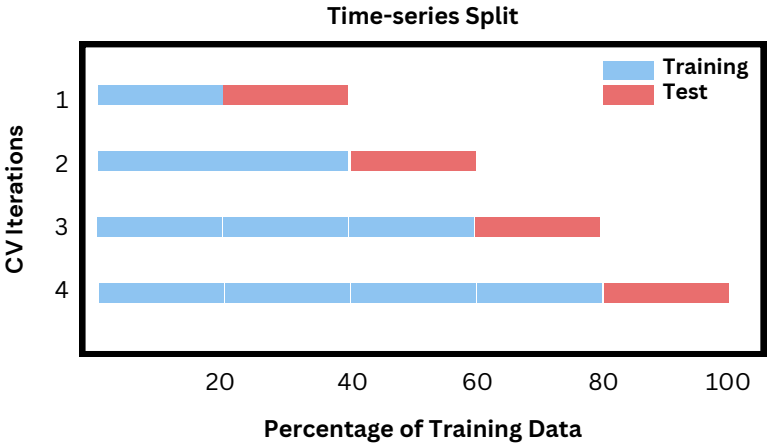


Figure 3.5: Time-series K-fold cross-validation (example for K=4).

NOMOTHETIC APPROACH

According to the nomothetic approach, the two methods described in Section 3.2.2 were investigated using the Explainable Boosting Machine

models (EBMs). EBMs were built using data from all individuals and then compared to the traditional personalized EBMs. In the first method, the training datasets from all individuals were concatenated into a single group-level dataset to train an EBM, referred to as EBM_all when applying the using_all approach. The number of interactions was fine-tuned to select the optimal value, as in the personalized models. The performance of this EBM_all model was evaluated separately on the testing set of each individual. The test sets are kept the same as in the personalized approach.

In the second method, information obtained from the first method (EBM_all used as teacher model) was further integrated into personalized EBMs. The class probabilities of the training samples were extracted and transformed to smoothed probabilities using a temperature softmax function, with the temperature value being selected from a range between 2 and 200. Thus, new datasets were created using the training samples of each individual and the extracted "probabilities" as a target label, instead of the initial hard labels $(0, 1)$. These new datasets created for each individual were used to train the student models, which are EBM regression models.

3.4. EXPERIMENTAL RESULTS

3.4.1. SYNTHETIC DATASET

IDIOGRAPHIC APPROACH

The initial step in evaluating the described methods was to create synthetic datasets. Using synthetic data, it is easier to understand the problem we have to solve and develop effective and efficient methods for that. To create the data, different values for the dataset's parameters, such as number of subjects, features and samples, were independently selected and investigated.

Synthetic datasets are first analyzed using a personalized approach. For each combination of the chosen parameters, personalized non-linear and linear models are applied to each individual of every dataset separately. After applying all personalized models, the mean and standard deviation values of the performance (AUC scores) across all created individuals are presented in Table 3.3. It is visible that the average AUC scores are greater when applying non-linear models. The extracted AUC results show that EBM models produce the best average scores in most datasets. However, even when RF or XGBoost show the best scores, their difference to EBMs is small. Moreover, EBMs achieve more accurate performance when trained on a large number of samples, such as 100 or 300.

NOMOTHETIC APPROACH

Subsequently, personalized EBMs are evaluated in comparison to the two nomothetic approaches described in Section 3.2.2, the using-all-data

Table 3.3: Performance of personalized models (EBM, XGBoost, Gradient Boosting, RF, SVM and Logistic Regression): the mean and standard deviation of the AUC scores are given for all synthetic datasets (each having a different number of users, features and samples). Numbers in **bold** indicate the highest mean AUC score for each dataset, while underlined numbers highlight cases, where EBMs achieve mean AUC scores that are close to the highest score, signifying competitive performance.

#Users	#Feat	#Samples	EBM	XGBoost	Grad	RF	SVM	LogReg
20	25	50	0.715 (0.149)	**0.747 (0.145)**	0.699 (0.179)	0.734 (0.168)	0.638 (0.185)	0.700 (0.149)
20	25	100	**0.736 (0.142)**	0.707 (0.127)	0.706 (0.132)	0.735 (0.130)	0.664 (0.130)	0.702 (0.087)
20	25	300	**0.695 (0.154)**	0.663 (0.148)	0.678 (0.133)	0.691 (0.147)	0.684 (0.163)	0.667 (0.157)
20	60	100	0.757 (0.147)	**0.762 (0.181)**	0.745 (0.153)	0.760 (0.142)	0.620 (0.147)	0.634 (0.143)
20	60	300	**0.761 (0.127)**	0.752 (0.121)	0.749 (0.107)	0.747 (0.127)	0.672 (0.105)	0.685 (0.113)
50	25	50	**0.736 (0.170)**	0.722 (0.170)	0.668 (0.157)	0.711 (0.155)	0.634 (0.188)	0.657 (0.173)
50	25	100	0.718 (0.128)	0.718 (0.133)	0.706 (0.128)	**0.726 (0.121)**	0.655 (0.145)	0.690 (0.132)
50	25	300	0.750 (0.111)	0.739 (0.108)	0.741 (0.107)	**0.751 (0.111)**	0.739 (0.123)	0.744 (0.121)
50	60	100	0.680 (0.154)	**0.684 (0.148)**	0.675 (0.136)	0.667 (0.148)	0.558 (0.150)	0.603 (0.136)
50	60	300	**0.764 (0.101)**	0.755 (0.105)	0.749 (0.103)	0.757 (0.101)	0.685 (0.101)	0.701 (0.102)
100	25	50	0.688 (0.179)	0.685 (0.158)	0.670 (0.172)	**0.695 (0.148)**	0.572 (0.193)	0.629 (0.177)
100	25	100	0.675 (0.147)	0.676 (0.144)	0.671 (0.144)	**0.690 (0.147)**	0.613 (0.133)	0.618 (0.131)
100	25	300	0.751 (0.110)	0.742 (0.101)	0.744 (0.104)	**0.757 (0.109)**	0.748 (0.109)	0.748 (0.110)
100	60	100	**0.737 (0.131)**	0.711 (0.134)	0.718 (0.122)	0.696 (0.122)	0.600 (0.131)	0.634 (0.122)
100	60	300	**0.722 (0.131)**	0.709 (0.128)	0.710 (0.117)	0.710 (0.126)	0.665 (0.091)	0.668 (0.112)

EBMs (EBM_all) and knowledge distillation (KD) method. In the case of knowledge distillation, different values for the temperature parameter are evaluated, ranging from 1 to 100. After applying all examined methods, the mean and standard deviation values of the AUC scores produced by each method for each synthetic dataset are presented in Table 3.4.

In the majority of the examined datasets, it is apparent that using personalized EBMs leads to worse performance than when either of the nomothetic methods is applied. More specifically, EBM_all gives the best results compared to the distillation method in all but three datasets, whereas in one of these, both methods achieved the same score. It is also interesting to mention that their difference in mean AUC score is quite large in certain datasets. This is the case in datasets with a small number of samples, such as when characteristics ({users, features, samples}) are {20, 25, 50}, {50, 25, 50}, {100, 25, 50}, {50, 60, 100} and {100, 60, 100}. Therefore, it is important to highlight that collecting sufficient data from each user can benefit the knowledge distillation process.

Furthermore, such comparative results could reveal insights regarding the role of temperature T in the predictive performance of KD. After exploring various values, from $T = 1$ to $T = 100$, a trend was apparent showing that $T = 100$ yields the highest AUC results, with a small difference to the overall best EBM_all. Therefore, increasing T to 100 leads to a softer probability distribution, which seems to facilitate a more effective transfer of information. Also, the difference in the effects highlights the importance of the temperature setting in achieving optimal perfor-

Table 3.4: Performance of the two nomothetic methods (EBM_all and KD): the mean and standard deviation of the AUC scores are given for all synthetic datasets (each having a different number of users, features and samples). Numbers in **bold** indicate the highest mean AUC score for each dataset, while underlined numbers indicate cases where distillation outperforms personalized EBMs.

#User	#Feat	#Samples	EBM	EBM_all	KD ($T = 1$)	KD ($T = 5$)	KD ($T = 100$)
20	25	50	0.715 (0.149)	**0.804 (0.151)**	0.753 (0.178)	0.768 (0.185)	0.776 (0.178)
20	25	100	0.736 (0.142)	**0.758 (0.162)**	0.739 (0.134)	0.735 (0.139)	0.753 (0.148)
20	25	300	0.695 (0.154)	0.691 (0.172)	**0.698 (0.167)**	0.694 (0.171)	0.690 (0.179)
20	60	100	0.757 (0.147)	**0.813 (0.111)**	0.786 (0.092)	0.779 (0.096)	0.795 (0.097)
20	60	300	0.761 (0.127)	**0.762 (0.119)**	0.757 (0.111)	0.756 (0.113)	**0.762 (0.119)**
50	25	50	0.736 (0.170)	**0.756 (0.183)**	0.707 (0.169)	0.719 (0.170)	0.731 (0.166)
50	25	100	0.718 (0.128)	**0.747 (0.146)**	0.713 (0.162)	0.720 (0.164)	0.733 (0.160)
50	25	300	0.750 (0.111)	**0.773 (0.133)**	0.762 (0.134)	0.769 (0.135)	0.769 (0.135)
50	60	100	0.680 (0.154)	**0.735 (0.140)**	0.689 (0.144)	0.686 (0.147)	0.700 (0.151)
50	60	300	0.764 (0.101)	**0.783 (0.120)**	0.751 (0.119)	0.755 (0.122)	0.766 (0.123)
100	25	50	0.688 (0.179)	**0.767 (0.175)**	0.720 (0.167)	0.725 (0.171)	0.736 (0.166)
100	25	100	0.675 (0.147)	0.723 (0.144)	0.719 (0.138)	0.720 (0.135)	**0.726 (0.141)**
100	25	300	0.751 (0.110)	**0.769 (0.121)**	0.767 (0.120)	0.765 (0.119)	0.764 (0.121)
100	60	100	0.737 (0.131)	**0.761 (0.140)**	0.712 (0.150)	0.721 (0.147)	0.738 (0.148)
100	60	300	0.722 (0.131)	**0.736 (0.142)**	0.724 (0.133)	0.720 (0.132)	0.729 (0.139)

mance.

3.4.2. DATASET: ALCOHOLDRINK

IDIOGRAPHIC APPROACH

First, the total number of 24 individuals is analyzed using a personalized approach. After applying all different ML models, the results of the personalized predictive models on the testing sets indicated that the produced results highly vary across individuals. For instance, some individuals had quite high AUC results, whereas others' results were at chance level.

To compare the different ML models, we show some of the statistical properties of all AUC scores, using the box and whisker plots in Figure 3.6. In this figure, we present the performance of EBMs compared to the full-complexity ML models as well as the performance of non-linear models compared to the traditionally used linear ones. Regarding the first comparison, the AUC distribution for EBMs is comparable to the ones of the other non-linear models. Apart from RF, which shows a slightly better overall performance, all statistical properties of the EBM scores reached higher values than the other three models. The median AUC score for EBM is around 0.81, only a bit lower than XGBoost (0.83). It can also be noticed that the minimum value of EBM performance was the highest among ML models, indicating a smaller variation among individuals in the case of EBMs.

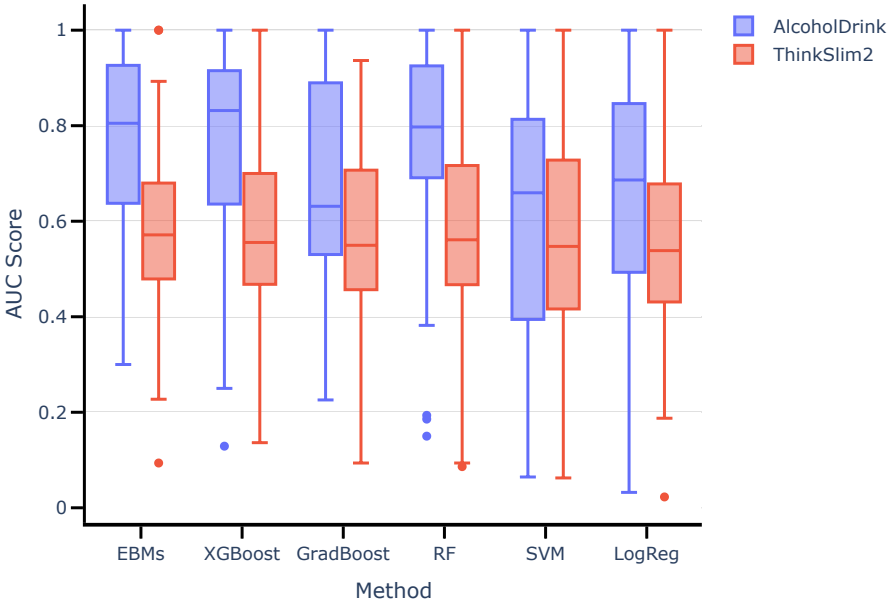


Figure 3.6: AUC performance of all non-linear, including EBMs, XGBoost, Gradient Boosting (GradBoost) and Random Forest (RF), and linear models, including Logistic Regression (LogReg) and SVM.

Regarding the second comparison, a distinction between the linear and non-linear models is visible. All statistical properties of the AUC scores are lower in the case of linear models. These findings highlight the ability of non-linear ML models to enhance the predictive performance of the traditionally applied linear ones.

NOMOTHETIC APPROACH

In the nomothetic approach, data from all individuals are pooled into a single dataset and modeled collectively by one EBM (EBM_all), or further exploited in a personalized way (KD). To facilitate comparison, box and whisker plots are utilized and presented (as before) in Figure 3.7.

Using a nomothetic approach, the AUC distribution of the KD method is improved compared to that of personalized EBMs. This shows more consistent performance scores across individuals, apart from 4 outliers. Regarding the EBM_all method, its AUC distribution is more spread out, with lower 25th percentile and minimum values compared to personalized EBMs and KD. However, the upper half of its distribution is comparable to the respective part of the distributions obtained through the other cases. Subsequently, by comparing the median values of both approaches, we see that there is a slight distinction between them, where personalized EBMs reach the level of 0.80, whereas around 0.76 and

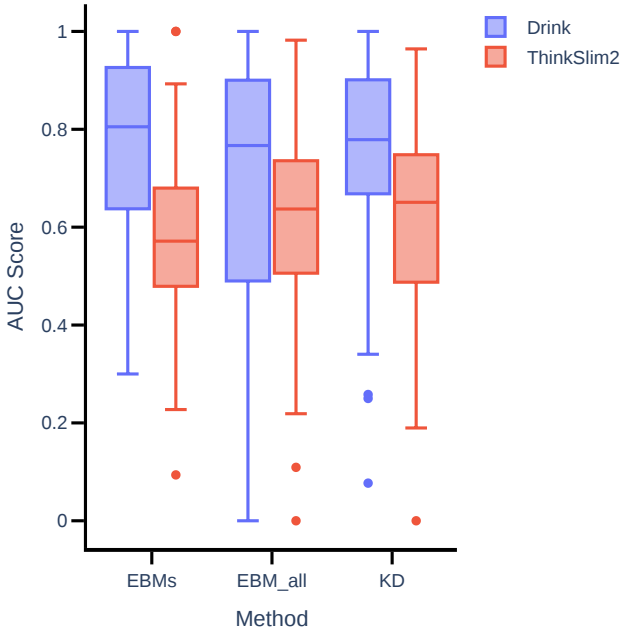


Figure 3.7: Comparing the performance of personalized EBMs to the two nomothetic approaches (EBM_all and KD).

0.79 for the EBM_all and KD methods, respectively. In contrast to the results on synthetic datasets, we see that in a more realistic dataset, the knowledge distillation method can lead to improved results compared to EBM_all.

3.4.3. DATASET: THINKSLIM2

IDIOGRAPHIC APPROACH

Similar to the previous dataset, the performance of 57 personalized predictive models is first evaluated. As the produced results highly varied across individuals, their performance is assessed here through box and whisker plots. Figure 3.6 presents the AUC scores of all different ML methods. According to AUC scores, all models' distributions are comparable to each other, having quite a large range. All methods show similar poor performance, achieving a low median value of around 0.57 in the case of non-linear models, whereas around 0.54 for the linear ones. That could be due to the more complex and challenging structure of this dataset, containing a larger number of individuals as well as features, but not more data samples compared to the previous dataset. Another interesting aspect of this experiment is that some AUC values are quite close to zero (for all setups). This means that probabilities produced by all models for these individuals lead to a flipped prediction label for almost all testing points.

NOMOTHETIC APPROACH

Finally, personalized EBMs were compared to both nomothetic approaches, EBM_all and KD. The results of all methods, in terms of AUC scores, are presented in Figure 3.7. The median, along with the 25th and 75th percentiles, are similar for both KD and EBM_all and are higher than the respective values for the personalized EBMs. The mean relative AUC increase for KD and EBM_all compared to EBMs is at 17% and 14%, respectively. It is also worth mentioning that one individual has an AUC score equal to 0. This means that the probabilities produced by both EBM_all and KD methods for this individual do not map the class labels correctly, maybe because they are different than the rest of the population. In challenging problems, like the one represented by the ThinkSlim2 dataset, where personalized non-linear models do not perform well, both nomothetic approaches are likely to achieve a slightly improved performance.

3.5. DISCUSSION

After the detailed presentation of experiments, this section presents a comprehensive view of the findings, evaluating the performance of idiographic and nomothetic models for EMA data modeling, particularly in predicting individual-level next time-point outcomes. Additionally, the unique challenges and considerations in EMA modeling are discussed.

3.5.1. IDIOGRAPHIC AND NOMOTHETIC APPROACHES

The idiographic approach demonstrates that personalized models, especially non-linear ones, can improve predictive accuracy by capturing unique patterns for each individual. This was particularly evident with complex ML algorithms, such as EBMs, which incorporate both linear and non-linear interactions, enhancing performance over traditional linear models. Among the non-linear models tested, Random Forest showed superior predictive accuracy across various synthetic datasets, while EBMs showed improvements as sample sizes increased, indicating that they perform better when more data is available.

In the examined real-world datasets, non-linear models maintained consistent performance, especially in the AlcoholDrink dataset, where EBMs and other non-linear methods had narrower and more consistent AUC score distributions than linear models. In the case of the ThinkSlim2 dataset, evaluating all personalized models was not that straightforward because their AUC score distributions were marginally improved in non-linear models. That could be due to the more complex and challenging structure of this dataset, containing a larger number of individuals as well as features, but not more data samples compared to the previous dataset.

Although the nomothetic approach is less tailored than idiographic models, it brings its own advantages. By training on aggregated data

across all individuals, these models can capture shared behavioral patterns, potentially making them suitable for individuals with sparse data. This generalizable approach could be beneficial for future application to new individuals or those with limited data, as they are not customized for specific individuals but instead provide insight into general EMA trends. Initially, we found that using-all-data models would achieve the highest AUC in the synthetic datasets, as these datasets are expected to lack the diversity and complexity typically seen in real-world data. This relative homogeneity allows the aggregated models to capture the underlying patterns more effectively across all individuals, leading to better overall performance. Regarding real-world datasets, applying these nomothetic methods to real-world datasets highlighted additional challenges. In particular, the more complex ThinkSlim2 dataset showed clearer improvements with nomothetic approaches, as Knowledge Distillation led to the highest AUC score. This improvement is more evident compared to the AlcoholDrink dataset, where individual models already achieved relatively high AUC scores (around 0.8), limiting the potential gains from aggregated data. The complexity and variability within ThinkSlim2 appear to better showcase the strengths of nomothetic over the idiographic methods, where general patterns could better reflect individual next time-point outcomes.

3.5.2. CHALLENGES OF MODELING EMA DATA

Studying the two aforementioned real-world datasets and noticing the variation in individual results in Figures 3.6 and 3.7 highlights the importance of collecting good-quality EMA data. It is challenging enough to map the complex nature of psychological behavior to a limited set of measured variables. EMA data collection is a tedious task, trying to capture multiple observations on subjective variables, meaning variables that rely on self-reports and individual perception, during an intensive period. Thus, EMA datasets may contain unclear or arbitrary responses due to user interpretation or variability in reporting, as well as missing values.

Moreover, label annotation is another challenging task in specific datasets. A clear example is the second examined real-world dataset, ThinkSlim2. Regarding the HU (healthy vs. unhealthy) output, the current goal is to predict any healthy or unhealthy events, by characterizing the type of food consumed each time. Although the labeling was based on the Dutch typical diet (as described in [123]), it remains a subjective task that relies on personal interpretation, making it difficult to describe the underlying phenomena. This challenge is evident in the poor performance of even the personalized predictive models, shown in Figure 3.6.

Furthermore, missing data is a significant problem of real-world EMA datasets that cannot be controlled during a study. Even though several individuals initially participate in a study, some may not produce

enough data for analysis (especially if one needs to take into account the temporal nature of the data). This issue was evident in Figure 3.4 for the AlcoholDrink dataset. Particularly, the sufficiency of data points depends on each individual's overall compliance throughout the entire data collection period, as well as their daily compliance. The most common approach to dealing with missing data is to delete them while keeping only the complete sets of data. However, this method relies on the assumption that the missing observations are missing completely at random (MCAR), which possibly is not always the case.

3.6. CONCLUSION

This chapter highlights the importance of exploiting the wealth of EMA data through more advanced ML models compared to linear ones. Non-linear vs. linear and idiographic vs. nomothetic approaches were investigated for classifying a target variable at a next time-point on different datasets.

The results showed great consistency for the idiographic approach, showing that non-linear models yield an enhanced performance on both synthetic and real-world data. Subsequently, regarding the nomothetic approaches, no clear trends were observed in the results of all datasets. Although the EBM_all method appeared to perform best for synthetic datasets, that is not the case for real-world datasets. Overall, the proposed knowledge distillation method could be recognized as the most beneficial method for improving the performance of personalized models. However, the performance differences between idiographic and nomothetic approaches were not found to be statistically significant.

In the next chapters, the focus stays on nomothetic approaches, with the goal of refining the selection of individuals as input to models. Given the considerable individual variability, it becomes apparent that instead of integrating all available individuals, focusing on those with similar characteristics may yield more effective results. Therefore, the question arises is how we can effectively identify and group similar individuals through clustering methods. Specifically, in the next chapter, Chapter 4, clustering based on raw time-series data is explored.

4

GROUP-BASED APPROACHES THROUGH CLUSTERING TIME-SERIES DATA

While nomothetic approaches with broad data integration offer extensive insights, the diversity between individual data and patterns in large datasets can sometimes obscure modeling the unique characteristics of each individual. To address this, strategically selecting meaningful groups of individuals can help optimize input information, enhancing our understanding of underlying processes at both individual and group levels. Such grouping can be obtained by clustering. Clustering is an unsupervised machine learning approach used to identify natural groupings within data based on similarity, without the need for labeled outcomes.

Specifically, this chapter examines the performance of various clustering approaches for grouping individuals based on the similarity of their raw time-series data patterns. Clustering is an unsupervised task and the true underlying groups are generally unknown, evaluating results can be challenging. Therefore, initially, simulated irregular time-series data, resembling EMA, are used to validate the performance of several methods under different clustering-related choices, such as the distance metric, with subsequent application to real EMA data.

Parts of this chapter have been published in

- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Clustering individuals based on multivariate EMA time-series data". In: The Annual Meeting of the Psychometric Society. Springer. 2022, pp. 161–171
- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Evaluating multivariate time-series clustering using simulated ecological momentary assessment data". In: *Machine Learning with Applications* 14 (2023), p. 100512

4.1. INTRODUCTION

Building on our exploration of nomothetic approaches from the previous chapter, where we discussed the integration of data from all available individuals, the focus shifts to building models on a subset of the whole population. Particularly, a more advanced strategy of only selecting homogeneous groups of similar individuals, as input to a model, is essential. Thus, an important question is whether meaningful groups of similar individuals could be uncovered, to subsequently facilitate building better models describing each individual. An answer can be found by grouping individuals, obtained by clustering.

Clustering algorithms aim to partition unlabeled data into homogeneous groups based on similarities among data points. In medical and psychological research, this type of analysis is highly valuable for multiple reasons. First, clustering enables a better understanding of shared characteristics and profiles within subgroups, which is essential for identifying common traits, behaviors, and potential risk factors among individuals. Such insights are critical for better understanding mental disorders and revealing the underlying processes that could inform targeted therapeutic strategies.

Additionally, clustering can support predictive modeling when personalized models are infeasible due to data limitations. As discussed in Chapter 3, in situations with insufficient data collected for a particular individual, training an effective personalized model may be impossible. Clustering offers a solution by allowing models to be trained on data from similar individuals, increasing the data pool and providing more reliable predictions for each subgroup. Moreover, clustering could help improve the performance of predictive/forecasting models [126]. By inputting more similar data, data of the clustering-uncovered similar individuals, to train a predictive model, the model is likely to produce more accurate outcomes than when using only personalized data. An increased accuracy contributes to more reliable representations of the overall processes.

In the context of EMA, clustering has clear benefits, but also poses many challenges, as discussed in [124]. Although applying clustering has already been studied for time-series [127–129], the question remains whether it is feasible to uncover meaningful clusters when there are no ground truth labels. Due to the unsupervised nature of the problem and the number of possible predefined parameters that come with every clustering algorithm, it is quite difficult to evaluate the validity and reliability of the results. The most significant reason is that there is no definitive answer about the true or optimal number and composition of groups. Each method, according to its objective function and parameters, aims to separate data in the most appropriate way, possibly leading to a different separation into groups. Thus, each method's setup yields a different grouping by capturing different data aspects or characteristics.

The wide range of available clustering-related choices demands ways to examine and validate clustering approaches for different scenarios,

such as heterogeneous datasets. A way to examine these choices is through simulations, using artificially generated datasets. Therefore, a large-scale EMA simulation study was conducted. Besides covering various scenarios regarding datasets' parameters (e.g., number of individuals, variables, missing data), another advantage of simulations is the existence of true labels. Labels will facilitate examining whether each method is reliable and accurate. Thus, this chapter aims to first validate the performance of several methods under different clustering-related choices, such as distance metrics, using simulated multivariate time-series data. The data are created in a way to resemble the complex structure of a real-world EMA dataset, considering its special characteristics, such as noisy and irregular data. Subsequently, because ground truth labels are not always available, or do not even exist, in real-world scenarios, the evaluation of the reliability and validity of clustering results is further investigated through distance-based and distance-free measures. Finally, some example simulated datasets are examined, showing in more detail all the necessary steps and comparisons when applying clustering for real-world applications on EMA data involving clustering approaches.

Summarizing, this chapter addresses the need to assess and validate clustering methods for diverse scenarios, particularly with heterogeneous datasets, to ensure accurate and reliable groupings of individuals. This aligns with RQ4, which examines different clustering approaches based on raw data, and RQ5, which focuses on evaluating the effectiveness of these clustering methods. To achieve this, a large-scale simulation study was conducted, using artificially generated datasets of well-shaped patterns, that ideally could resemble emotion-related data. Finally, the chapter provides a detailed analysis regarding the impact of different datasets' parameters on clustering performance, leading to valuable recommendations for future application of clustering on real-world EMA data.

4.2. RELATED WORK

Time-series clustering has been studied a lot lately, with some handful reviews found in [127–129]. Most studies focus on clustering based on the raw time-series data, exploring different choices for the clustering method, distance metric and evaluation. Considering that all well-known clustering algorithms can be adapted to time-series data, these methods range from traditional machine learning techniques [130], such as k-means and hierarchical clustering, to neural network-based approaches [131], such as Autoencoders [132] and Self-Organizing Maps (SOMs) [133], which learn representations before clustering. Consequently, a key challenge in time-series clustering lies in selecting the right distance metric [134]. Thus, most research studies have focused on finding a good representation of time-series similarities and integrating it into clustering algorithms. Typically, distance measures are based on the concept

of intensity distance or shape resemblance [135]. Despite the wide application of intensity-based measures, such as Euclidean distance, considering two individuals similar if their variables' intensity at each time point is close, these do not take into account shape information. Thus, two versions of the same pattern shifted in time cannot be considered similar. To deal with such common issues in time-series, recently, shape-based distance metrics have been widely applied, trying first to optimally align the data.

The field of time-series clustering has advanced due to the effectiveness of shape-based techniques, such as Dynamic Time Warping (DTW), known for aligning time-series data that may vary in speed. Building on the success of shape-based clustering, various DTW variations have been proposed. These include applying restrictions to DTW [136], softening optimal distance paths using softDTW [137], or enhancing the focus on local time-series structures with methods like shapeDTW [138]. Other studies exploring different shape-based information ([135, 139, 140]), propose the use of the longest common subsequence (LCSS), cross-correlation and Fréchet distance, respectively.

However, most studies have handled univariate time-series data. The added value of the current chapter is the multi-level structure of EMA data, including several multivariate time-series data [141]. In the case of multivariate time-series, kernel-based data representations have been proposed [142]. Kernels based on DTW, such as Global Alignment Kernel (GAK), were used [137]. Moreover, in [81], another time-series cluster kernel (TCK) was proposed, based on Gaussian mixture models (GMMs).

Specifically for EMA data, only limited research has been conducted as far as clustering is concerned. In [66], clustering EMA data into similar meaningful groups or clusters is proposed. However, clustering was not implemented, leaving a gap that is covered in this chapter. Other than this, a different goal focusing on clustering EMA variables or items was investigated in [143, 144]. In such cases, clustering was used to organize an individual's symptomatology into homogeneous categories of symptoms, rather than to group different individuals, as in the current chapter.

4.3. CLUSTERING EMA DATA

This section provides an overview of all the necessary steps for performing a clustering analysis on EMA data and evaluating the validity and reliability of the solution. Taking into account the heterogeneity of clustering methods available, this chapter focuses mainly on distance-based methods. Such methods rely heavily on the way of estimating distances between the data points of the EMA time-series. Thus, we first explore the most appropriate distance metrics, followed by clustering algorithms and evaluation measures.

4.3.1. DISTANCE METRIC FOR EMA DATA

A thorough understanding of the characteristics of the EMA data can facilitate the selection of the most appropriate distance metric [140]. These have been already identified in Chapter 2 (Section 2.2.1), thereby we can move on to the choice of a distance metric. Calculating distances provides insight into the data elements, in this case, individuals, that are more similar and need to be grouped together. There are various distance metrics, each reflecting a different kind of similarity between time-series [139, 140]. As a result, applying different clustering analyses to the same dataset can yield markedly different results.

The most traditional distance metric used in clustering tasks is the Euclidean distance. It calculates the distance between the time-series of two individuals x_1, x_2 using the formula defined in the following Equation 4.1, focusing on intensity difference or change. The only requirement is that both time-series are of the same length ($T_1 = T_2 = T$).

$$d_{EUC}(x_1, x_2) = \sqrt{\sum_{v}^{V}\sum_{t}^{T} ||x_{1,v,t} - x_{2,v,t}||^2} \tag{4.1}$$

In the case of EMA data, this requirement is usually violated because of the occurrence of missing values [145]. Differences in the number of missing values among individuals result in MTS of varying lengths.

To tackle this issue, another distance metric can be used. Dynamic Time Warping (DTW) is the most widely used metric for time-series data [146]. It has also become a state-of-the-art metric because of its high accuracy and its application in the case of variable-length time-series data [137, 147]. By stretching or compressing time series along the time axis, DTW aims to find the best shape-based alignment of these [140]. This way, it also accounts for differences in points' time intervals due to missing values, but at the same time, outliers or noise do not significantly affect it. In practice, this is achieved by comparing all possible alignment paths and finally getting the one leading to the minimum distance, denoted as π^*. An example of the optimal path between time-series is depicted in Figure 4.1. This is also shown in the following Equation 4.2, which gives the distance between two time-series x_1, x_2. In detail, the alignment path $\pi^* =< p_1, p_2, .., p_K >$ consists of K elements, where each one is represented by an index pair (t_i, t_j) capturing the time-point index of the series x_1 and x_2, respectively. Then, the distance D_{p_k} at each $p_k = (t_i, t_j)$ path element is defined by Equation 4.3, using a kernel ϕ which is the squared Euclidean distance.

$$\begin{aligned} DTW(x_1, x_2) &= \min_{\pi^*} D_{\pi^*}(x_1, x_2) \\ &= \min_{\pi^* =< p_1, p_2, .., p_K >} \sqrt{\sum_{k=1}^{K} D_{p_k}(x_1, x_2)} \end{aligned} \tag{4.2}$$

$$D_{p_k}(x_1, x_2) = D_{t_i,t_j}(x_1, x_2) = D(x_{1,1..V,t_i}, x_{2,1..V,t_j}) \\ = \phi(x_{1,1..V,t_i}, x_{2,1..V,t_j}) + DD \tag{4.3}$$

where

$$\phi(x_{1,1..V,t_i}, x_{2,1..V,t_j}) = \sqrt{\sum_{v=1}^{V} ||x_{1,v,t_i} - x_{2,v,t_j}||^2}$$

$$DD = \min \begin{cases} D(x_{1,1..V,t_{i-1}}, x_{2,1..V,t_{j-1}}) \\ D(x_{1,1..V,t_{i-1}}, x_{2,1..V,t_j}) \\ D(x_{1,1..V,t_i}, x_{2,1..V,t_{j-1}}) \end{cases}$$

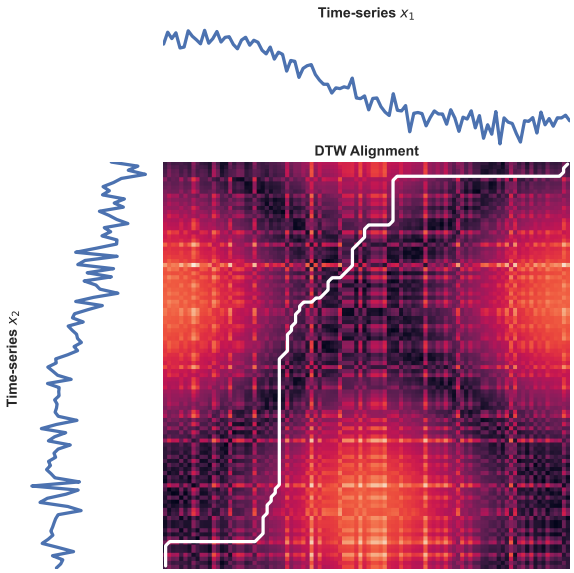


Figure 4.1: An example of the DTW alignment between two time-series, x_1 and x_2, with the optimal alignment path π^* shown as a white line. It illustrates how DTW aligns similar patterns, such as steady (flat) parts and downward slopes, between the two time-series, which may represent the same variable measures for two different individuals.

Because of its success, many variations of DTW were developed, such as subsequence DTW and softDTW [137, 139]. Among these, the Global Alignment Kernel (GAK) is an interesting extension of softDTW (calculated in Equation 4.4 below), which also has all inherent advantages of kernels, see [85].

Practically, GAK can be used to estimate data similarity according to the Equation 4.5. Although it still uses a non-positive kernel ϕ, it develops a "seemingly" positive definite kernel. This is achieved through the exponentiation of ϕ. Specifically, exponentiation helps transform it into a form that behaves similarly to a positive definite kernel. As also shown in Equation 4.5, k_{GAK} incorporates a gamma hyperparameter controlling the softDTW smoothing, where $\gamma = 2\sigma^2$.

$$\begin{aligned} softDTW(x_1, x_2) &= soft\_min_\pi D_\pi(x_1, x_2) \\ &= -\gamma \log \sum_\pi e^{\frac{-\phi(x_1, x_2)}{\gamma}} \end{aligned} \tag{4.4}$$

$$\begin{aligned} k_{GAK}(x_1, x_2) &= e^{\frac{-softDTW(x_1, x_2)}{\gamma}} = e^{\frac{-\gamma \log \sum_\pi e^{\frac{-\phi(x_1, x_2)}{\gamma}}}{\gamma}} \\ &= \sum_\pi e^{\frac{-\phi(x_1, x_2)}{\gamma}} \end{aligned} \tag{4.5}$$

4.3.2. ADJUSTING CLUSTERING METHODS FOR EMA DATA

Knowing about the heterogeneity of clustering methods, this chapter is limited to raw-based algorithms [140]. Their goal is - by using information from the raw time-series data - to better separate the most dissimilar data elements, or individuals in our case. Methods belonging to both hard and soft or fuzzy clustering categories are described.

HARD CLUSTERING

All the well-known hard clustering methods, such as k-means and hierarchical clustering can be used on time-series data [127–129]. Two main challenges arise: i) how to integrate the appropriate distance measure, and ii) how to calculate the centroid (center) of a cluster in case it is needed. While the first is addressed differently for each clustering method, through their own objective function, regarding the second, time-series centroid calculation is mainly based on barycenter averaging, as proposed in [148]. Centroid calculation requires averaging time-series, which, in the case of DTW, is not that straightforward, relying on shape information.

According to the literature on clustering analysis, the vast majority of studies use the simplest k-means algorithm, which starts with k random initial cluster centers and it updates these with respect to the objective function, which is the clustering error [149]. In other words, it finds local optimal solutions by minimizing the intra-cluster distances d (e.g., d_{EUC}, DTW) between each cluster center c_j and the individuals x_i belonging to this cluster, C_j. Its objective J is summarized in Equation 4.6, showing how the selected distance metric can be incorporated. In the case

of DTW, where $d = DTW$, distances are calculated according to Equation 4.2, while the centroid through DTW barycenter averaging [148].

$$J = \sum_{j}^{k} \sum_{i \epsilon C_j} d(x_i, c_j) \tag{4.6}$$

However, there are two main disadvantages. The dependence on the initial cluster centers and the requirement of linearly separable cluster centers can have a great impact on k-means performance. To address the second weak point of k-means, kernel k-means was developed. In principle, kernel k-means is a generalization of the standard k-means algorithm, where the input data have already been mapped to a higher feature space through a non-linear kernel [149]. Here, the GAK kernel k_{GAK} is used to transform the data and calculate their similarities. This is incorporated into J using two ways. According to the objective function defined in Equation 4.6, k_{GAK} is input into d after transforming similarity to distance through $d = d_{GAK} = 1 - k_{GAK}$, or by aiming to maximizing J, instead of minimizing, when keeping $d = k_{GAK}$.

One of the issues with kernels is that we cannot have access to cluster centers in the original feature space. This is because using a kernel function, we take advantage of the kernel trick, which means that we get the inner products of input data in the feature space without explicitly knowing the transformation ϕ.

Another type of hard-clustering approach is the agglomerative hierarchical clustering algorithm (HC). HC is based on the step-wise integration of single individuals into clusters. The whole process starts with each individual representing a separate cluster. Then, the most similar clusters are grouped together until all belong to a single cluster. When clusters contain multiple elements, each element in one cluster must be compared with every element in the other cluster. The distance between clusters is then defined based on the chosen linkage method: single linkage (minimum distance), complete linkage (maximum distance), or average linkage (average distance). The distance scores across all the grouped clusters can be easily represented graphically by a dendrogram. The highest distance indicates the optimal number of clusters. Again, choosing between d_{EUC}, DTW, k_{GAK} for a distance/similarity metric is possible.

FUZZY CLUSTERING

In the field of psychopathology, grouping individuals using a hard clustering method is not always a realistic scenario, for example, due to comorbidity (see Section 1.2.2), individuals could meet the criteria of more than one diagnosis of a mental disorder. This suggests that it is theoretically plausible that individuals belong to more than one group [2]. So, a fuzzy clustering approach might be more appropriate [150].

The most widely used fuzzy algorithm is Fuzzy c-means (FCM) [150, 151]. Similarly to k-means, it tries to optimize the clustering error, with

respect to cluster centers as well as the membership matrix. At each iteration, the membership degree is calculated according to the distance/similarity of each cluster center. As the FCM objective function, shown in Equation 4.7, is optimized, each individual x_i gets a higher membership $e_{i,j}$ (controlled by the fuzziness parameter m) to the cluster whose center c_j is more similar to its actual TS. All membership degrees of each individual belonging to each cluster must be summed to 1. Similarly to k-means, in the case of DTW, $d = DTW$ and J need to be minimized, whereas for GAK, $d = k_{GAK}$ and the J are maximized.

$$J = \sum_{j}^{k} \sum_{i \epsilon C_j} e_{i,j}^{m} d(x_i, c_j) \tag{4.7}$$

As an alternative to finding centroids between individuals, Fuzzy k-medoids (FKM) uses the most representative individual of each cluster as its center. Compared to artificially-created centroids, which are barycenters of time-series [148], in this case, clusters are represented by individuals already existing in the dataset. Practically, in Equation 4.7, instead of c_j, x_j is used. Regarding the distance metric d, all above mentioned can be applied.

4.3.3. EVALUATION MEASURES

A key challenge in clustering lies in evaluating the reliability and validity of the clustering solution. We should acknowledge that as in most cases, there are no optimal or correct results. Each method, according to its objective function, parameters and approach to find similarities between the data points, aims to partition data in the most suitable way. This can result in different clustering outcomes, each with varying group formations that may be considered optimal within the context of that specific method.

To overcome this issue and be able to identify a "good" clustering result, evaluation is usually based on clustering quality. In general, individuals belonging to the same cluster need to be close (cohesion), and each cluster needs to be well-separated from the other clusters (separation). Both cohesion and separation rely on the proximity between individuals, which is estimated through a similarity or distance metric. Several distance-based evaluation measures compare different clustering solutions by taking into account both intra-cluster and inter-cluster similarities [152]. However, most of these, such as Inertia and Davies-Bouldin Index, are centroid-based measures, meaning that they are strongly dependent on artificially-extracted centroids. Centroids, or barycenters in MTS, are hard to estimate, especially in irregularly spaced EMA data. This places an additional error factor in the evaluation procedure. So, measures based on between-individuals distances are considered the most appropriate approach.

A popular distance-based evaluation measure is using Silhouette coefficients [152]. According to the formula in Equation 4.8, for each individual i belonging to a cluster C_i, it compares the average distance (a_i) across all w individuals belonging to the same cluster (where $w = 1..W_i$ and W_i is the total number of individuals belonging to C_i) to the distance b_i across all z individuals of the closest cluster (where $z = 1..Z_i$ and Z_i is the total number of individuals belonging to the closest to C_i cluster). To find the closest cluster, similarities among all individuals in a cluster are taken into account. So, it's quite straightforward to interpret the clustering results. Its values range from -1 to 1, where high values indicate a good clustering, values close to -1 a bad clustering (inaccurate grouping), and values close to 0 a meaningless clustering (random grouping).

$$Sil = \frac{\sum_i^N \frac{b_i - a_i}{\max(b_i, a_i)}}{N}, \text{where} \qquad (4.8)$$
$$a_i = \frac{\sum_{w \epsilon C_i} d(x_i, x_w)}{W_i} \text{ and } b_i = \frac{\sum_{z \epsilon \text{ closest } C_i} d(x_i, x_z)}{Z_i}$$

An alternative approach to evaluate clusters is through distance-free methods, where distances across individuals are not taken into account for their calculation. An example of a promising distance-free metric is clustering stability [153]. Because of initialization issues, running a clustering algorithm several times on the same dataset may lead to different results. For instance, there are multiple ways to separate N individuals into k clusters, all finally giving different clusters of individuals. It is apparent that when clusters are very different, clustering should be considered unstable. In other words, even in cases where the same optimal number of clusters k is always found, a different separation of individuals also affects the quality of clustering. To evaluate clustering stability, it is needed to run the clustering algorithm several times and compare the matching of individuals' cluster assignments. The matching distance can be calculated using Mutual Information across all pairs of produced labels. The result represents the clustering stability index [153].

4.4. SIMULATIONS FRAMEWORK

To assess the performance of each clustering method, several simulated datasets were generated to resemble EMA data. Each MTS dataset was designed in a controlled way by keeping its parameters (e.g., number of clusters) in specific ranges. By keeping some parameters constant, we can check the influence of others in clustering performance. Also, in such a controlled setup, the true clustering labels of each dataset are known. These can be exploited in the evaluation process, after comparing them to the produced partition labels of each method, to check the methods' validity. However, because in real-world cases the true labels are not available, other evaluation measures are also examined. These include

some distance-based measures, such as Silhouette coefficients, as well as distance-free measures, such as labels' stability.

4.4.1. SIMULATED SCENARIOS

After carefully considering the special characteristics of an EMA dataset and its complex structure (given in Section 2.2.1), the simulation procedure was formulated. The simulated datasets were designed to represent diverse cases of MTS data, resembling real-world EMA scenarios to some extent. The datasets' heterogeneity is reflected by the number of parameters and the chosen ranges of their values. All changing parameters are summarized below:

- Number of clusters [2, 3, 4, 5]
- Population size [20, 50, 100, 250]
- Number of features [2, 5, 10]
- Percentage of noisy features [0, 0.2, 0.5, 0.8]
- Percentage of missing data [0, 0.1, 0.2]

For example, the number of investigated clusters was $2, 3, 4$ and 5. Thus, each simulated dataset is eventually characterized by the combination of all chosen parameters' values.

During the generation process of each MTS dataset, first, the values for the parameters were all set, regarding the number of clusters, population size (total number of individuals), number of features (or variables) and percentage of noisy features and missing data. Then, individuals were equally split into the defined number of clusters. For example, if the population size is set to 20 and the number of clusters to 2, there will be 10 individuals in each cluster. Afterwards, each cluster was formed by selecting all feature patterns (MTS) of all cluster's individuals, that is the equations used to generate each time-series feature. This level of complexity occurs because individuals need to be represented by time-series of high dimensionality, but also be part of distinct clusters. Thus, the MTS of an individual feature space must be similar for all in the same cluster, meaning that they are all characterized by the same random combination of features (or time-series patterns). Then, to differentiate individuals within a cluster, auto-correlated Gaussian noise (distribution's variance is 0.2) was added to each of them. To illustrate this, Figure 4.2a shows an example of 3 clusters, each including 5 individuals characterized by 1 feature/variable. On the contrary, separate clusters were characterized by a different combination of features. In other words, all features were different, generated by different equations. Regarding these equations, they were formed in a way to provide some particular or basic patterns for time-series, plausible to resembling emotional behavior, captured in EMA items. The basic patterns are represented by a pool

of trends, including linear and non-linear trends, as well-expected in real EMA data. For instance, these included upwards, downwards, constant as well as sinusoidal with different periods and phases, or combinations of these [154], and generated for 100 time-points. Figure 4.2b shows an example of some of the basic patterns used, while the rest were created in a similar way. Overall, the combination of features was randomly selected, but controlled in a way that all individuals of the same cluster had the same combination of features, which was different from other clusters.

As already mentioned, each time-series feature was generated for 100 time-points. However, following the fact that missing points typically occur in real-world EMA data, the percentage of missing data also varied across the simulated datasets. By changing the maximum percentage of missing points, each individual was represented by a different number of time-points, regardless of the cluster they belonged. Each time, the number was randomly selected with respect to the maximum percentage of missing data. For example, when examining a maximum percentage of missing data of 20%, $P_m = 0.2$, all individuals' data points were selected from a range of $[80, 100]$.

Other parameters that were investigated are the total number of features ($2, 5,$ and 10) and the percentage of noisy features L_n, checking for $0, 0.2, 0.5,$ and 0.8. The latter parameter indicates the percentage of total features, which are not generated following one of the basic patterns and are just drawn from a Gaussian distribution. In such a setup, the basic patterns are considered non-noisy features. For instance, in the case of the combination of 2 clusters, 10 features and $L_n = 0.8$ or 80% noisy features, for each cluster, 2 features are selected from the pool of basic patterns, while 8 are random Gaussian noise. These 2 non-noisy basic features selected for the 2 clusters must always be different in order to better distinguish the clusters.

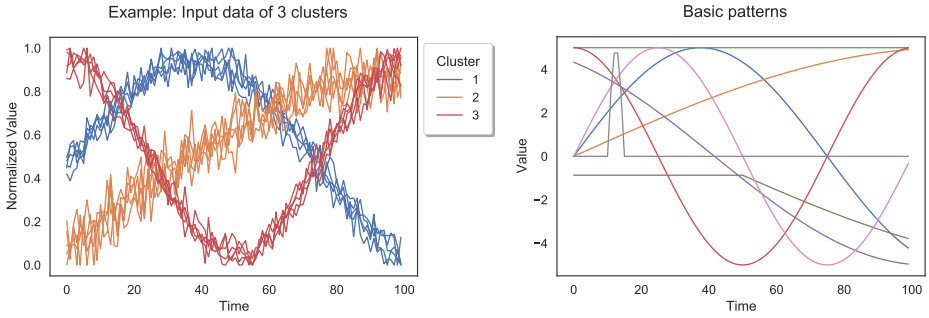
Overall, for each dataset (combination of different parameters examined), the generation process is repeated 10 times, leading to 10 different example datasets. This is necessary to ensure that clustering is not affected by any specific feature combination. Therefore, 10 examples of each dataset lead to a total number of 5760 datasets, examined within each configuration with a varying percentage of missing data points.

4.5. SIMULATIONS RESULTS

To assess the performance of the above-described methods along with the different clustering choices[1], these were tested using the generated simulated datasets. All clustering-related choices that need to be examined as well as the ways of their evaluation are summarized in Table 4.1.

Methods that rely on their initial cluster centroid definition, such as *km*,

[1] All distance metrics were applied using the toolbox [147].



(a) Example of 3 clusters (differentiated by color), each including 5 individuals.

(b) Pool of basic patterns (foundational shapes) from which individual time-series are constructed.

Figure 4.2: Examples of the generated simulation patterns used to create synthetic time-series data for clustering evaluation.

Table 4.1: All the examined clustering parameters regarding methods, distance metric and evaluation.

Method	Distance Metric	k	Evaluation
k-means (*km*)	DTW	2	True Labels (AMI)
Fuzzy c-means (*FCM*)	GAK	3	Silhouette
Fuzzy k-medoids (*FKM*)		4	Stability
Hierarchical (*HC*)		5	
		6	

FCM, and *FKM* are applied 10 times to each dataset. At each iteration, initial cluster centroids are randomly generated and updated according to each algorithm. Then, the derived 10 sets of labels are used to calculate stability, while the average result for Adjusted Mutual Information (AMI) with the true labels and Silhouette coefficients is recorded. For *HC* methods, only individual distances are taken into account. So, they are not dependent on any initialization, and stability does not need to be calculated.

All methods are applied to every simulated dataset. A simulated dataset is characterized by a unique combination of all the controlled parameters with varying values (see Section 4.4.1). For every combination of parameters-values, we generate 10 different example datasets to better represent each combination because of the random selection of features in a cluster. For each example dataset, the patterns of a cluster's features $(2, 5, \text{or } 10)$ are randomly chosen from the pool of basic feature shapes.

4.5.1. SUMMARY RESULTS

EVALUATION SUMMARY USING TRUE LABELS

First, the overall performance of all clustering methods is tested against the true labels of each dataset. The AMI between the true and predicted labels is calculated and depicted in Figure 4.3. This figure is divided into 4 subplots, each representing the performance of all methods' performance on datasets with a different number of true clusters $(2, 3, 4, 5)$. Note that all different datasets are summarized in this plot. For example, there is no distinction between datasets with and without noisy features.

In each subplot, the AMI distributions corresponding to true k values (equal to the number of true clusters) show the highest median AMI score across all methods. Moreover, the overall distribution of these true k value is elevated compared to different k values. By zooming in on the first subplot, all statistical properties of the blue distributions ($k = 2$) are the highest for all methods. However, the color of a distribution may not always be visible, if the distribution is flat (minimal spread). In some cases, it can be even represented by a line with some outliers. For example, this pattern is mostly observed for FKM_{GAK} and HC_{GAK}, indicating their robustness. However, the existence of some outliers is highlighted in all subplots, showing cases where performance decreases. Outliers with low values are prevalent in methods where the DTW distance metric is used (as opposed to kernel-based methods). As it will be proven later in the analysis, lower clustering performance is connected to a high proportion of noisy features when non-kernel methods are applied.

Having access to the true underlying clusters gives us the opportunity to actually validate the examined clustering methods. Therefore, through simulations, comparing the predicted and true labels of individuals led to the confirmation of most methods' validity. This means that in a simulated scenario, each method could not only identify the right number of clusters, but also the correct partitioning of individuals between clusters. This is important information for validating the performance of all clustering methods and the robustness in cases where values are close to the highest possible score of 1. However, this is not the case when a wide range in the AMI distribution or even some outliers are observed. More specifically, there are methods, such as FCM_{DTW} and FKM_d, whose AMI distributions are quite spread, deviating a lot from 1. These are, then, considered less robust methods, not being capable of always identifying the correct separation of individuals into clusters. Moreover, it should be noted that there are some outlier examples where DTW-based clustering methods could not uncover the true clusters and need to be investigated. Further analysis should be then performed on the influence that the datasets' changing parameters might have.

EVALUATION SUMMARY USING SILHOUETTE ANALYSIS

Since ground-truth labels are not usually available in real-world datasets, different evaluation measures need to be additionally assessed for their

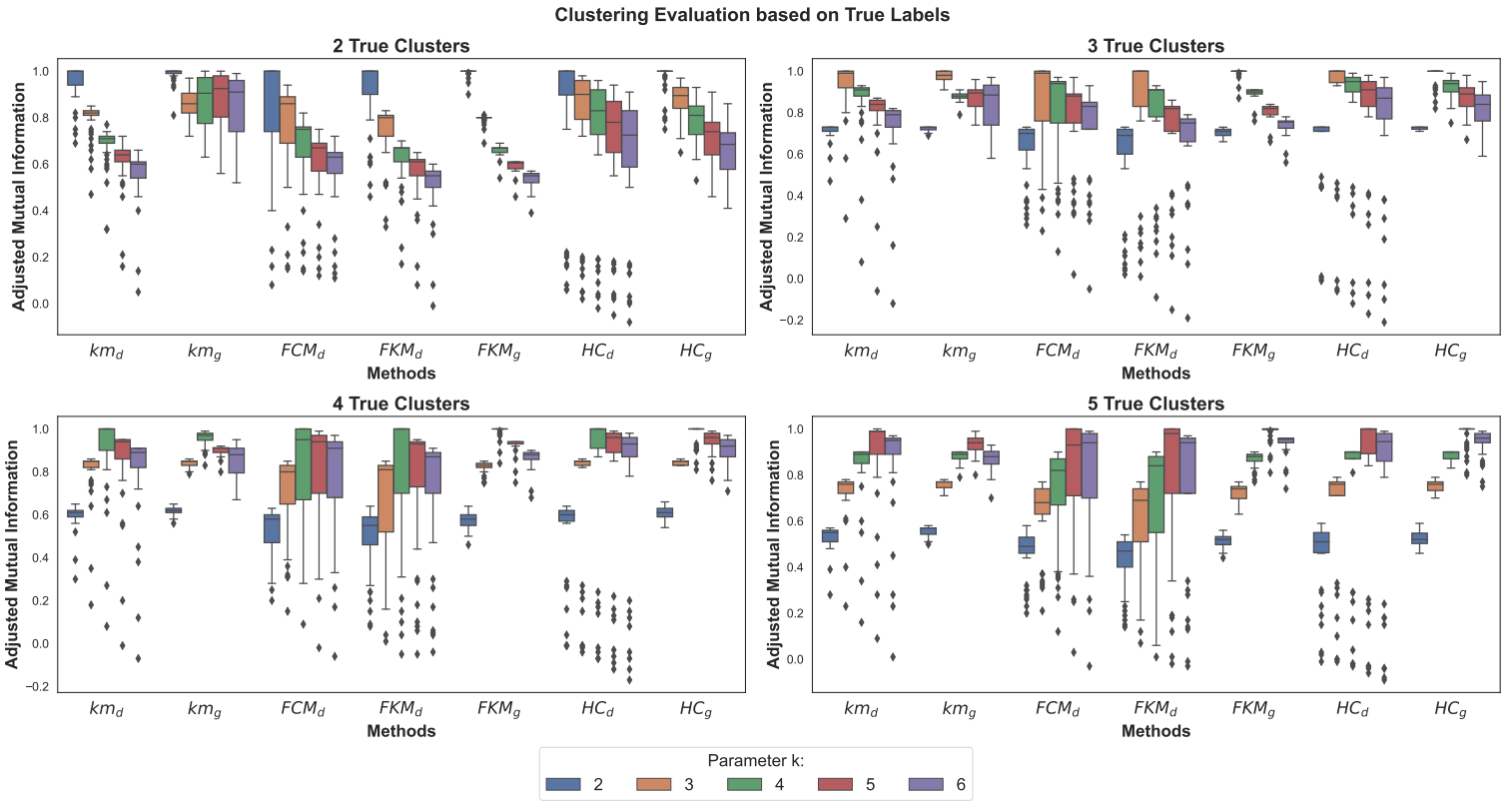


Figure 4.3: Overall performance of all clustering methods tested against the true labels of each simulated dataset. Each subplot presents the performance only on datasets of 2, 3, 4, and 5 true clusters, respectively. In each subplot, a boxplot of the average AMI scores is presented for all methods with different k hyperparameters applied to each simulation setup of datasets. Different colors indicate the different values of parameter k.

ability to identify the correct k cluster value. Silhouette coefficient and stability are two widely used alternative measures to evaluate clustering. Their results are shown in Figures 4.4 and 4.5, respectively. Again, each figure has 4 subplots, with each subplot reflecting the datasets with a specific true number of clusters ($2, 3, 4, 5$ clusters).

For Silhouette analysis, the distributions of the coefficients are increased on average when the value of k is equal to the true number of clusters of the examined dataset. This means that even without having access to the true labels, Silhouette analyses can also provide evidence regarding the correct value of k for all clustering methods. However, the Silhouette coefficients are not as high as the AMI values, compared to when true labels were known, which is expected because they rely on internal data structure, without reference to the true labels. If methods using DTW distance are applied, distributions are quite widespread, reaching an average value of around $0.6 - 0.7$ and a maximum of 0.8. Such values are considered relatively average, if compared to the highest possible value, which is 1. So, even without reaching the highest score, by comparing the Silhouette coefficients for different k values, it results to the right number of clusters.

Interestingly, it is observed that the boxplots of the kernel-based methods are much higher (average $0.8 - 0.9$) with a smaller range, compared to the non-kernel methods. After further investigation, the lower values of non-kernel methods are derived from datasets with a high proportion of noisy features. These values can be very low in some datasets, reaching close to zero. This effect also confirms the conclusion of Figure 4.3, where kernel-based methods led to higher AMI scores and efficiently grouped individuals in all possible scenarios. Nevertheless, it is not yet apparent if and why a kernel method clearly outperforms the others, which needs further investigation.

When considering a distance-free approach of evaluation, the distributions of the stability indexes also seem powerful enough to identify the true underlying labels of different datasets. Similar to the previous plots, all statistical properties of the distributions are higher when the correct value of k is picked, also confirmed by the true number of clusters. Compared to the Silhouette distributions, stability can reach higher values, averagely close to 1 when the true k is used. So, stability can be applied complementary to finding the best k value for all methods. It is also clear that FKM_{GAK} achieves in all cases the most stable clustering results. Note that *HC* methods are not part of this plot, as they are not dependent on any initialization parameters, and therefore produce a single partitioning. The *HC* methods are only based on individuals' distances that can be calculated before applying the algorithm.

IMPACT OF PARAMETERS: NOISE

As a next step, the impact of different dataset parameters is investigated, such as the proportion of noisy features, number of features and

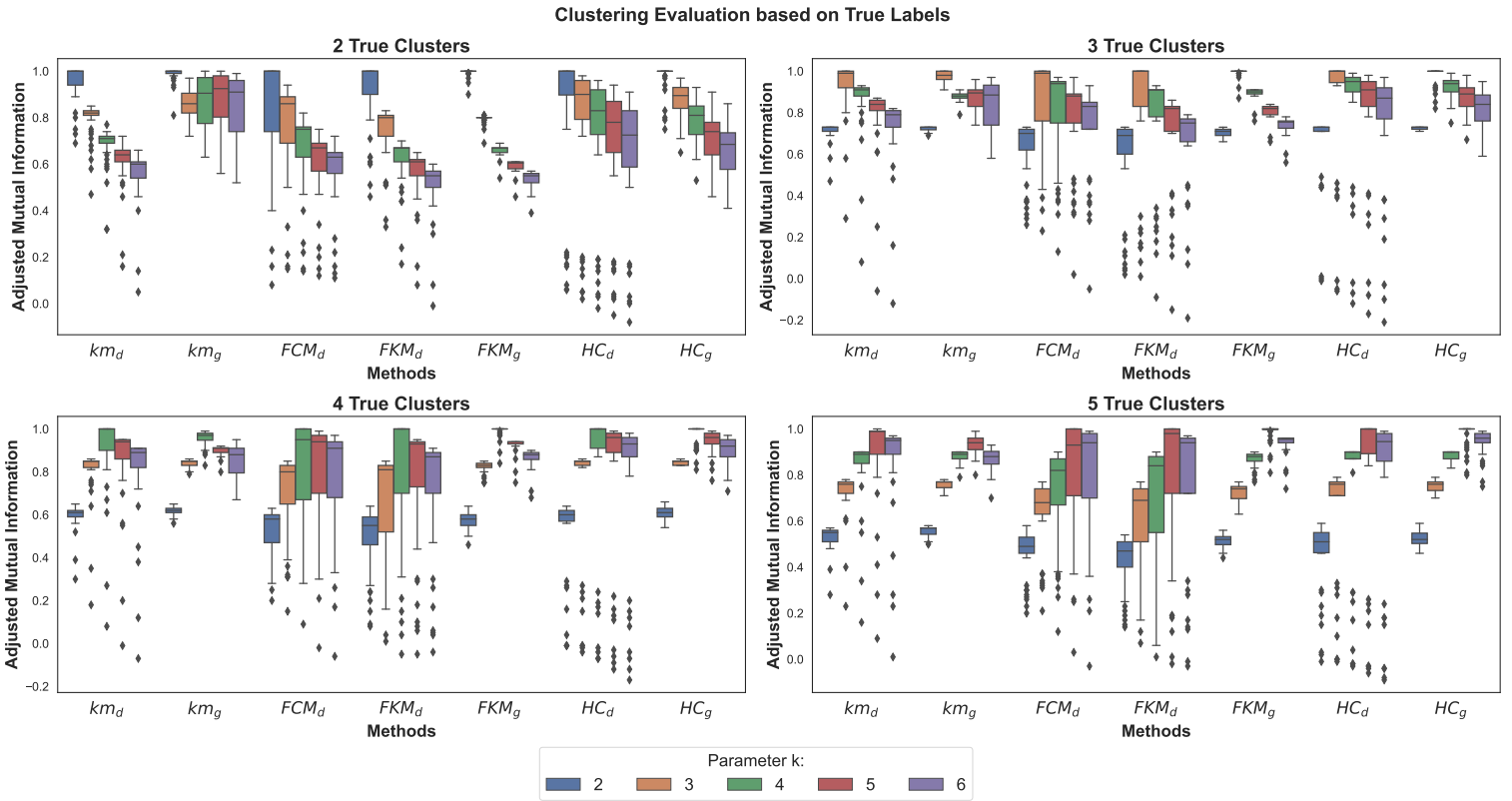


Figure 4.4: Overall performance of all clustering methods assessed through Silhouette coefficients of each simulated dataset. Each subplot presents the performance only on datasets of 2, 3, 4 and 5 true clusters, respectively. In each subplot, a boxplot of the average AMI scores is presented, for all methods with different k hyperparameters applied to each simulation setup of datasets. Different colors indicate the different values of parameter k.

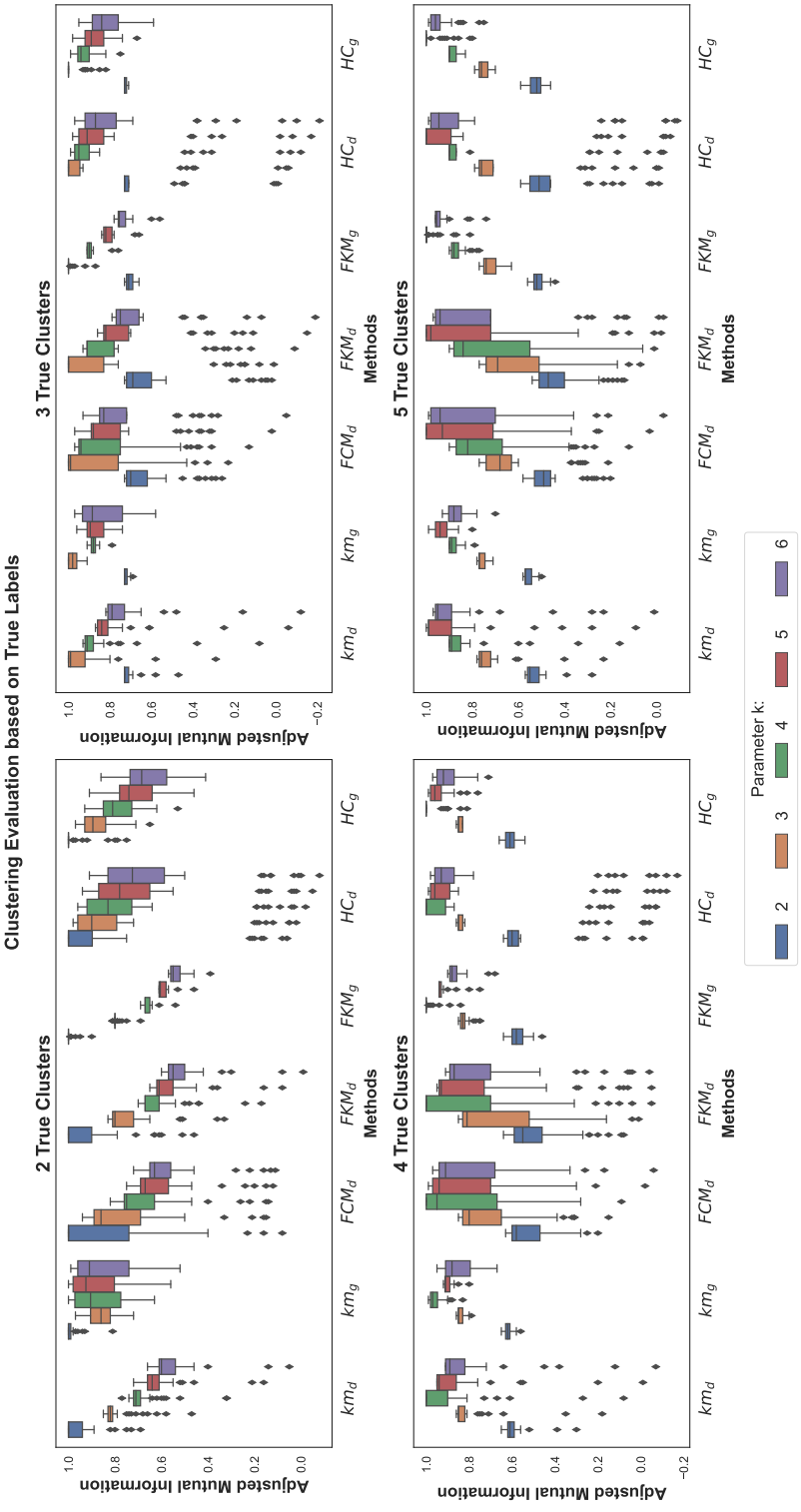


Figure 4.5: Overall performance of all clustering methods assessed through Stability of each simulated dataset. Each subplot presents the performance only on datasets of 2, 3, 4 and 5 true clusters, respectively. In each subplot, a boxplot of the average AMI scores is presented, for all methods with different k hyperparameters applied to each simulation setup of datasets. Different colors indicate the different values of parameter k.

missing data as well as population size. Figure 4.6 shows the impact of the proportion or level of noisy features (L_n set to $0, 0.2, 0.5$ or 0.8) on clustering performance, as measured by AMI. The presented figure is a bit more complex than before, where rows and columns represent each of the clustering methods and the true number of clusters, respectively. To limit this figure's complexity, only the cases of 2 and 3 true clusters are shown. As shown in Figure 4.6, AMI decreases as the proportion of noisy features increases, with the largest difference in datasets with $L_n = 0.8$. For example, when examining the red boxplots ($L_n = 0.8$) of HC_{DTW}, AMI reaches values close to zero, showing poor performance, but also the distinction between the values of the k parameter is not clear. Similar patterns appear for km_{DTW}, FCM, and FKM_{DTW}.

Contrary to $L_n = 0.8$, the difference gap is smaller for $L_n = 0.5$, and is negligible for 0.2 noisy features. Interestingly, in the case of kernel-based methods, the performance drop is not that significant, even for $L_n = 0.8$. This indicates that existing relationships, like similarities, between individuals in very noisy datasets can be more reliably extracted after applying a GAK kernel, than just using DTW distances.

Similarly, when checking the impact on Silhouette and Stability, the same patterns are apparent (see Figures 4.18 and 4.19 in the Supplementary Material of this chapter). Again, the gradual decrease in both measures is obvious as the proportion of noisy features increases, with the most significant drop visible at $L_n = 0.8$. However, the range of the distributions is different between kernel and non-kernel methods, where the latter ones are more widespread and averagely decreased.

All in all, Silhouette analysis is capable of efficiently retrieving the true number of clusters of all datasets when using kernel methods, even with a high proportion of noisy features. Similar effects are also observed for clustering evaluation through stability.

IMPACT OF PARAMETERS: NUMBER OF VARIABLES

Having already identified the irregularities when dealing with datasets with high proportions of noisy features, we further focus on examining cases with noise $L_n = 0.8$. For these datasets, the influence of the number of variables on clustering performance is investigated using Silhouette and Stability, presented in Figures 4.7 and 4.8, respectively. Due to space limitations, this analysis aims to check if the potential impact can be captured in case there is no access to the true labels (impact on AMI can be similarly checked). Both figures present multiple subplots, separating all methods and the number of true clusters. Once more, only the GAK kernel-based methods seem to show a good performance after finding the right number of clusters. Additionally, when the number of variables is higher, 10 compared to 5, clustering performance appears slightly increased for the majority of the presented cases. For kernel-based methods, this could be interpreted as a need to transform high-dimensional data, because an alternative data representation might bet-

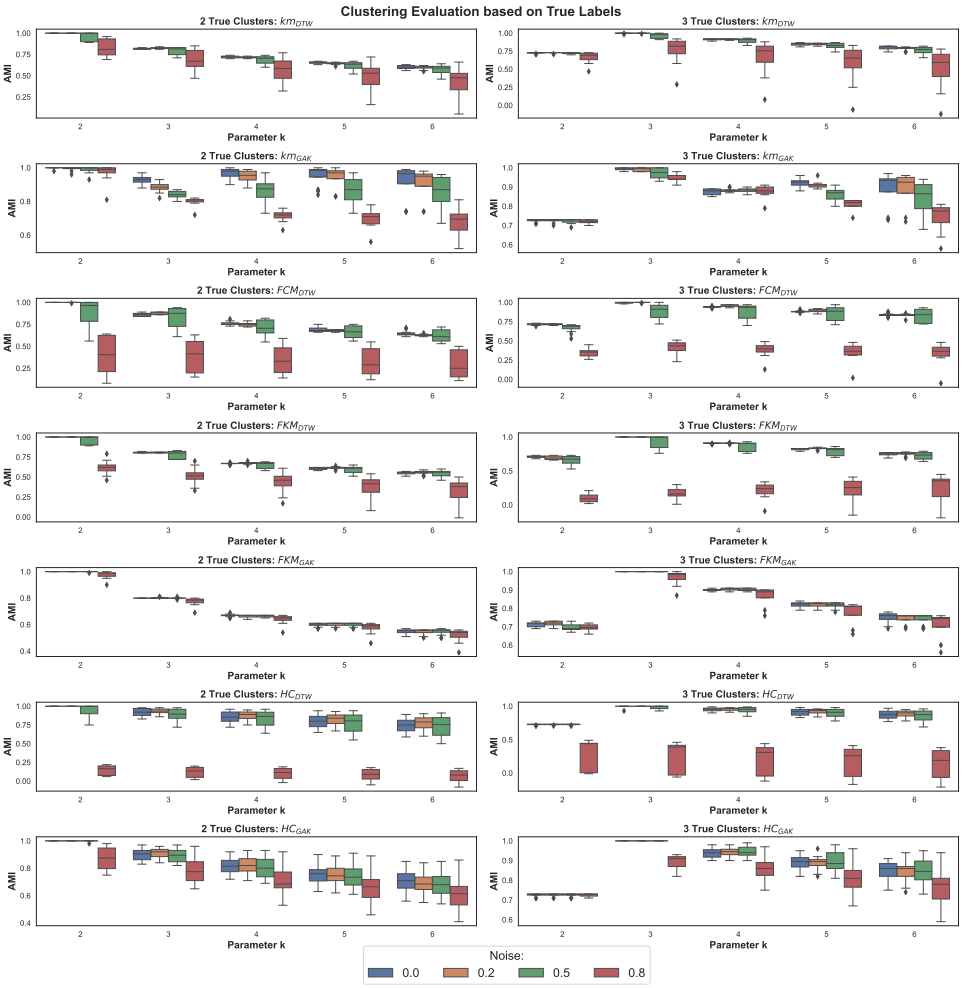


Figure 4.6: Influence of noise L_n (represented by a different color) on the overall performance of all clustering methods assessed through AMI. There are 7 by 2 subplots, where the rows represent each of the clustering methods, while the columns the true number of clusters (2 and 3 clusters are shown).

ter more capture its patterns. However, the fact that this also holds for non-kernel methods makes this interpretation a bit more complicated, because no kernel transformation is involved. Therefore, it is possible that the effect is not due to the high number of features but to another reason, such as the number of meaningful ones.

Furthermore, in both plots, but more clearly when considering Silhouette coefficients, non-kernel methods tend to consistently group individuals into 2 clusters. This occurs even when more clusters may be necessary to accurately represent the true underlying patterns, but only 2 are

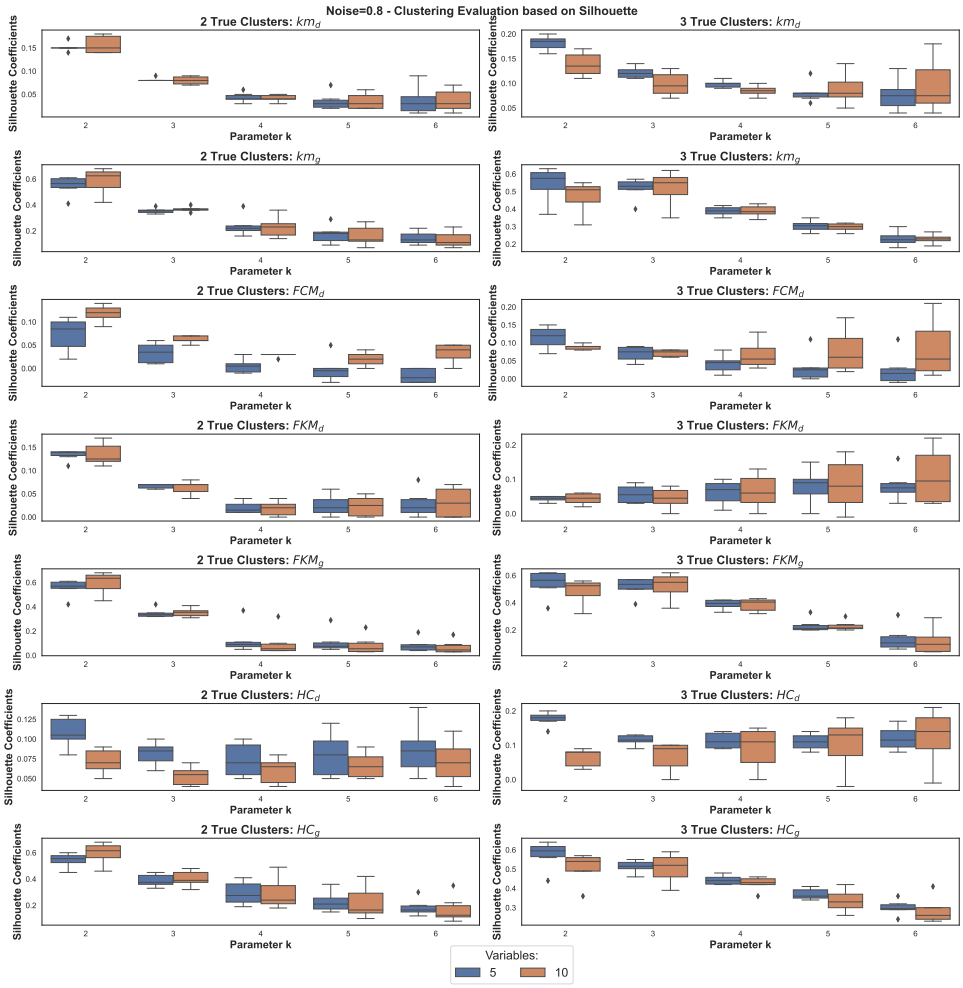


Figure 4.7: $L_n = 0.8$: Impact of variables' number on the overall performance of all clustering methods assessed through Silhouette. There are 7 by 2 subplots, where the rows represent each of the clustering methods, while the columns the true number of clusters (2 and 3 clusters are shown).

commonly extracted.

IMPACT OF PARAMETERS: MISSING DATA

Similarly, the impact of the missing data percentage of a dataset was explored. Given that in real-world EMA datasets, irregular time series with missing values commonly arise, it was crucial to investigate its effect on clustering performance. Clustering validity was first tested using AMI, but also relative to noise, as proven quite an influential factor. For low

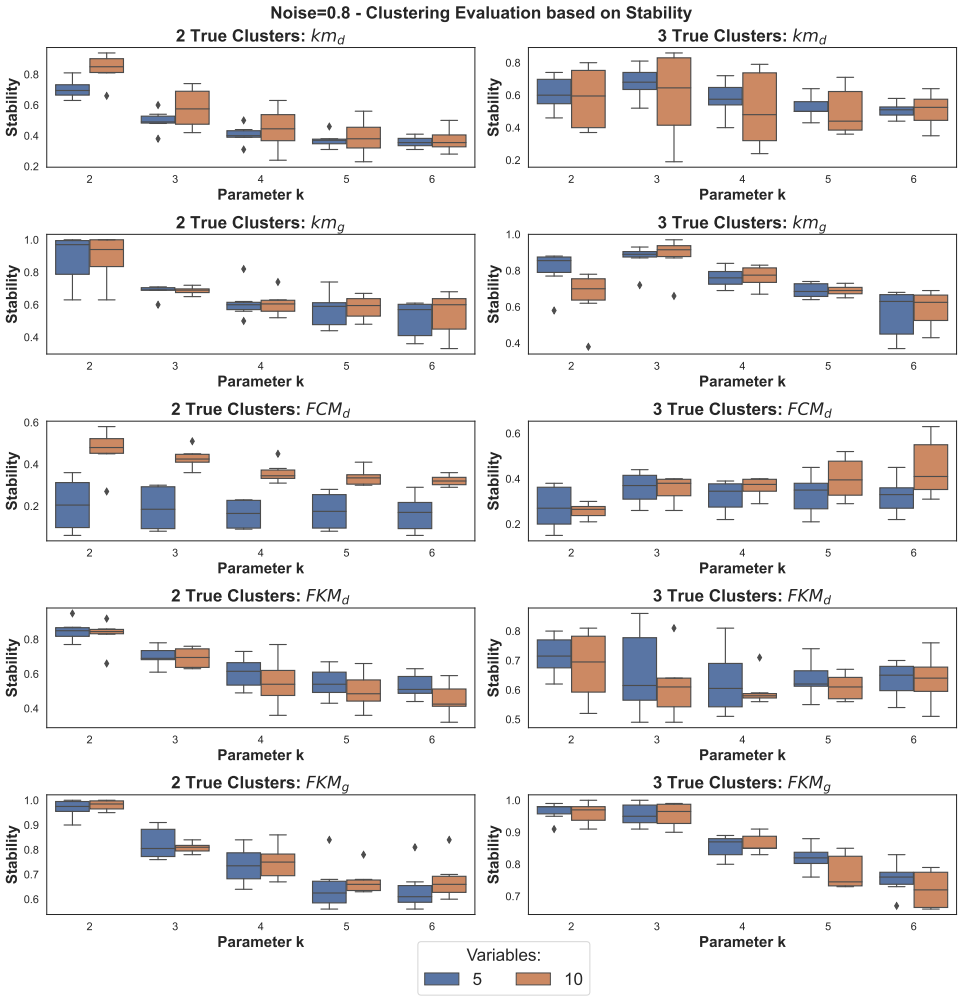


Figure 4.8: $L_n = 0.8$: Impact of variables' number on the overall performance of all clustering methods assessed through Stability. There are 3 by 2 subplots, where the rows represent each of the clustering methods, while the columns the true number of clusters (2 and 3 clusters are shown).

proportions of noise, the presence of missing data did not affect clustering performance, only a slight drop was observed for $P_m = 0.2$, in a few cases using the GAK distance metric. However, when $L_n = 0.8$, the impact is depicted in Figure 4.9. As demonstrated before, noise did not significantly affect the kernel-based methods. This is confirmed again by showing that the right clusters are mostly uncovered, and always outperforming the DTW-based methods. However, the effect of missing data is not negligible. While, for $P_m = 0$ or $P_m = 0.1$, the AMI scores are close

to 1, when $P_m = 0.2$, the scores drop to 0.8 for both FKM_{GAK} and HC_{GAK}, and even lower to 0.6 for k-means. Thus, $P_m = 0.2$ along with high noise proportion starts affecting the good performance of the GAK distance metric. However, this refers to the average AMI score across multiple datasets, and as further analyzed, this score is heavily influenced by the dataset itself and the particular points that were excluded in each one. For instance, in case a lot of data points, which were randomly removed, were important to characterize its feature pattern, then it naturally gets more challenging to discover its true cluster.

Nevertheless, for non-kernel-based clustering methods, noise had been already found to have a significant influence. Here, it is observed that this influence is apparent even when $P_m = 0$, meaning that there are no missing values in data, where reaches a maximum score of 0.8. Consequently, a stronger decrease was expected for $P_m = 0.2$. Nevertheless, the opposite trend was found. In other words, AMI was equal or higher in cases with more missing data, for example, when $P_m = 0.2$ compared to $P_m = 0.1$. This was probably achieved because removing data points from a noisy feature does not take any useful information out, but it may become less noisy. However, in all these cases, even the improved ones, the yielded scores were not higher than 0.5, so still the clustering is not considered successful.

Clustering performance should also be assessed through Silhouette coefficients and Stability to check whether the same patterns could be derived without having access to the true labels. The results of the Silhouette analysis are given in Figure 4.20 of the Supplementary Material of this chapter. The true clusters can be always uncovered in case of kernel-based methods, with scores decreasing as P_m increases, but not lower than 0.5. Likewise, stability analysis yields similar findings. Therefore, kernel-based clustering methods can reliably find true clustering even in the case of noisy datasets with irregular time-series. As expected, increasing the percentage of missing points affects the performance, but depends a lot on which data points are removed along with their significance in characterizing the feature pattern.

IMPACT OF PARAMETERS: POPULATION SIZE

Finally, we experimented with the number of participants in the dataset. Based on all evaluation measures, all methods seem to accurately perform on average, without a clear distinction between the varying population sizes. The same patterns appear in boxplots of datasets with 20 or 250 individuals. Therefore, the total number of individuals in a dataset does not influence clustering performance. Detailed figures presenting the influence of population size on clustering performance are provided in the Supplementary Material of this chapter (Figure 4.21).

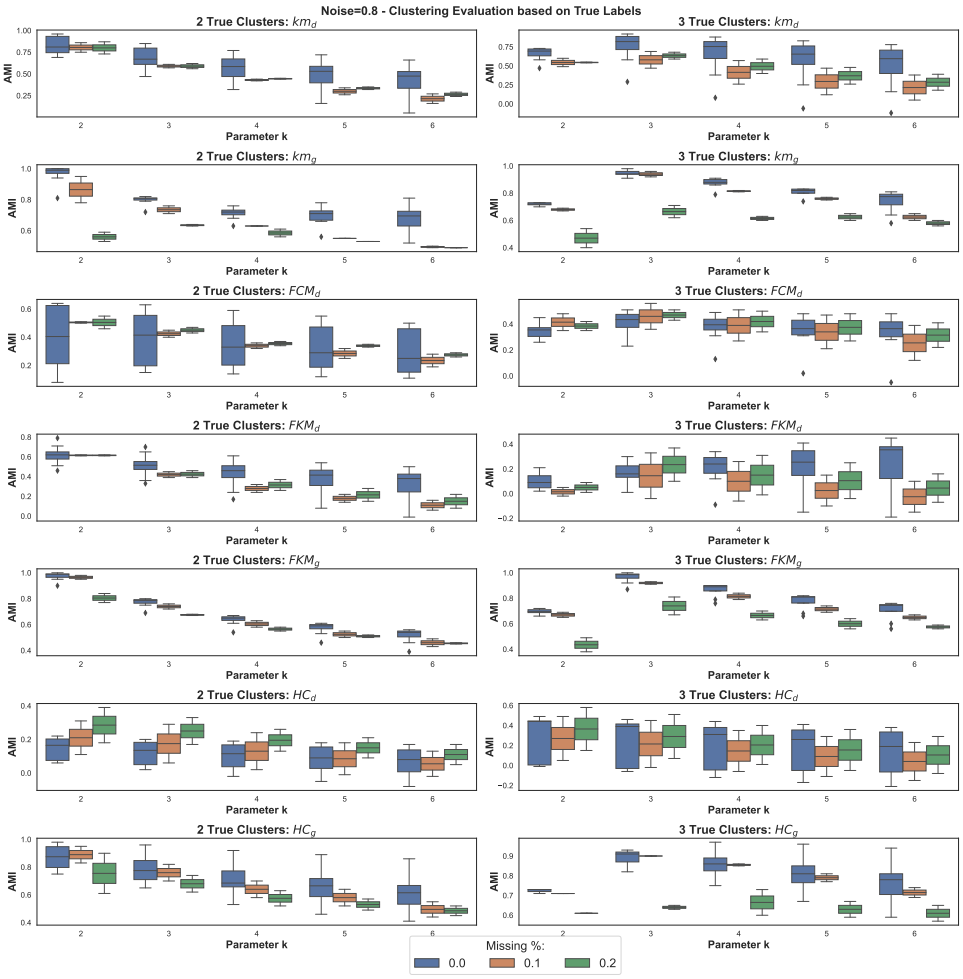


Figure 4.9: Influence of the percentage of missing data points P_m (represented by a different color) on the overall performance of all clustering methods assessed through AMI. There are 7 by 2 subplots, where the rows represent each of the clustering methods, while the columns the true number of clusters (2 and 3 clusters are shown).

4.5.2. APPLICATION ON A SIMULATED SCENARIO

We further experimented with some particular example datasets to analyze in more detail how the above-examined evaluation measures can be used when dealing with a new dataset, but also to highlight the differences in results between some clustering choices. As described in Table 4.2, the focus was on two simulated scenarios without missing data, covering different proportions of noise, $L_n = 0$ and $L_n = 0.8$. The selected example datasets represent a case of 20 individuals belonging to 4 clus-

ters. Each individual was generated having 10 features, with varying percentages of non-noisy features. Relevant to our analysis, in these examples, 20 individuals are sequentially split into clusters, that is, the first 5 were designed to belong to the first cluster, the next 5 to the second cluster, etc.

Table 4.2: The characteristics of the two simulated scenarios examined for Section 4.5.2.

	Clusters	**Population**	**Features**	**Noisy Features**
Scenario 1	4	20	10	0
Scenario 2	4	20	10	0.8

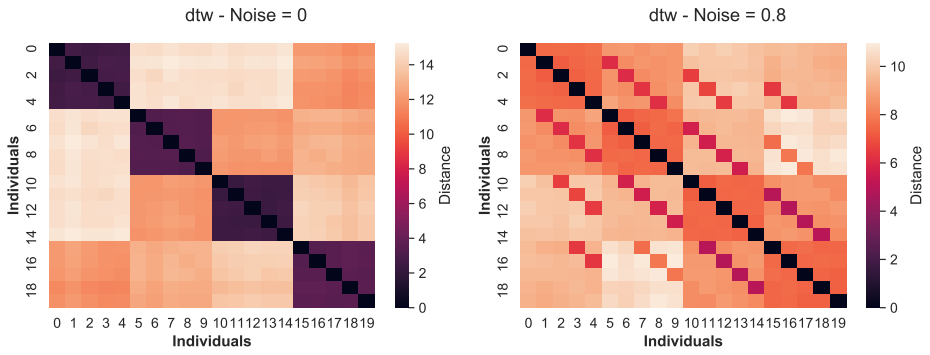


Figure 4.10: DTW distance matrices across all 20 individuals, for different noise levels, $L_n = 0$ and $L_n = 0.8$.

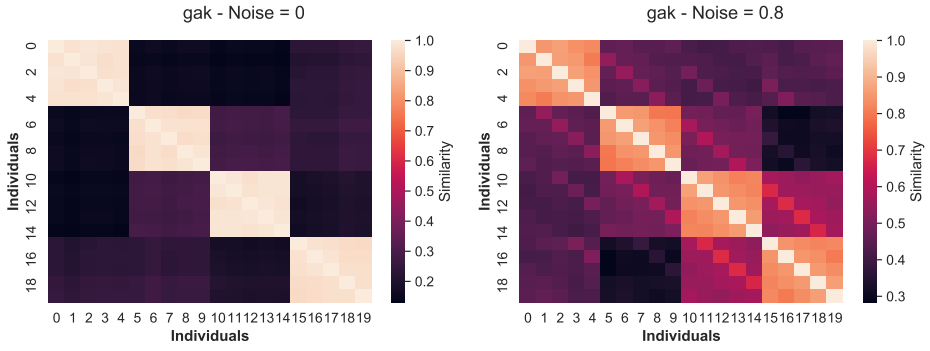


Figure 4.11: GAK similarity matrices across all 20 individuals, for different noise levels, $L_n = 0$ and $L_n = 0.8$.

The first step is choosing the most appropriate distance metric. Here, both DTW and GAK are separately tested. It is important to emphasize

that DTW estimates distances, whereas GAK similarities. The produced distance and similarity matrices of all 20 individuals for DTW and GAK are shown in Figures 4.10 and 4.11, respectively. Both result in 20 by 20 symmetric matrices. Across all pairs of individuals, the distance/similarity values reflect which individuals are close to each other. In other words, small distances or high similarity reflect members of the same cluster.

At first glance, although 3 plots mostly follow a specific pattern uncovering the 4 clusters, a peculiar pattern, resembling randomness, emerges for DTW and high noise ($L_n = 0.8$). This observation likely explains the poor performance of non-kernel methods in the summary results. Interestingly, the GAK kernel does not seem to be affected a lot by high noise. Only within-cluster similarities slightly decrease, and between-cluster similarities slightly increase as the proportion of noisy features increases.

Subsequently, all different clustering methods are applied to the examined datasets, and clustering results are evaluated. The performance evaluation is conducted in the same three ways for both datasets, as illustrated in Figures 4.12 and 4.13. As shown in the summary results, for $L_n = 0$, all methods are able to find the correct number of clusters. This is confirmed by all three evaluation measures. Only in the case of Silhouette coefficients, non-kernel clustering methods lead to lower values compared to kernel-based methods. However, these lower coefficients reach a value close to 0.7 for $k = 4$, which still indicates relatively strong clustering performance, given that the maximum possible value is 1.

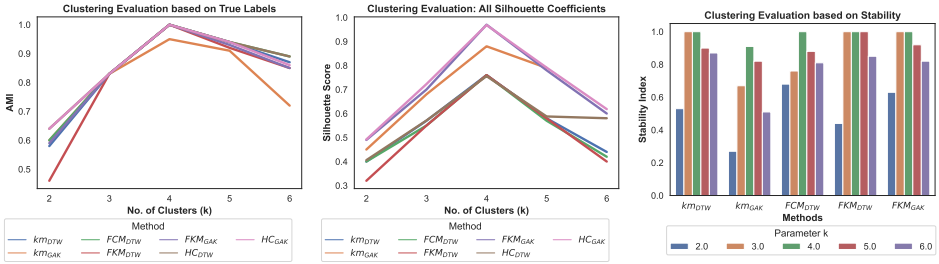


Figure 4.12: $L_n = 0$: Clustering evaluation through true labels, Silhouette scores and Stability index.

Additionally, some of these measures, such as Silhouette coefficients, can provide further insights into clustering quality. According to the Equation 4.8, the total Silhouette coefficient is estimated as an average of all individuals' Silhouette coefficients. Therefore, the Silhouette coefficients of each cluster can be particularly informative. This can verify whether a clustering analysis has produced a set of meaningful clusters. For example, the presence of clusters with low Silhouette coefficients (e.g., lower than 0.2) suggests that some clusters may not be well-separated, indicating the need to adjust the number of clusters. Figure 4.14a illustrates the Silhouette coefficients of each method, calculated

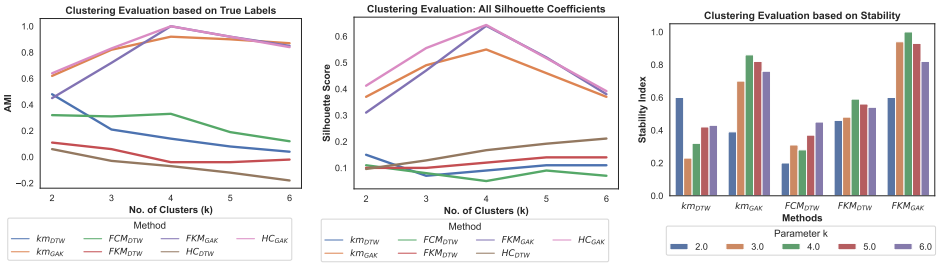


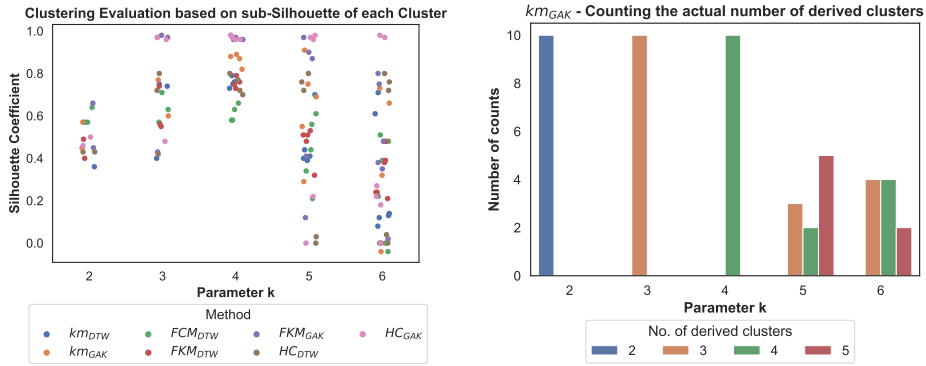
Figure 4.13: $L_n = 0.8$: Clustering evaluation through true labels, Silhouette scores and Stability index.

across various methods and different numbers of cluster k. Each method is represented by a distinct color, and the same number of points (corresponding to k) appears for each method. As k increases, it becomes evident that the coefficients for some clusters begin to drop, approaching 0. This trend indicates that, beyond $k = 5$, additional clusters do not add meaningful structure and instead result in poorly defined clusters. These findings strongly suggest that a smaller number of clusters is more appropriate for capturing the underlying structure of the data.

Furthermore, regarding stability, another aspect that could be analyzed is the actual number of derived clusters for every value of the k parameter. Even though the input of a clustering method is k, this works as an upper bound for some algorithms, such as km_{GAK}. Therefore, it does not mean that it always extracts k clusters, but if necessary, it could also extract fewer than k clusters. In Figure 4.14b, the distribution of the total number of clusters is examined across 10 iterations for all methods and input values of k. The different colors represent the number of total counts. The results reveal that the algorithm consistently extracts the requested number of clusters for lower values of k. However, for higher values ($k = 5$ and $k = 6$) the algorithm occasionally produces fewer clusters than specified. For instance, with $k = 6$, the algorithm most frequently extracts 3 or 4 clusters. These findings suggest that the algorithm naturally adjusts to the underlying structure of the data.

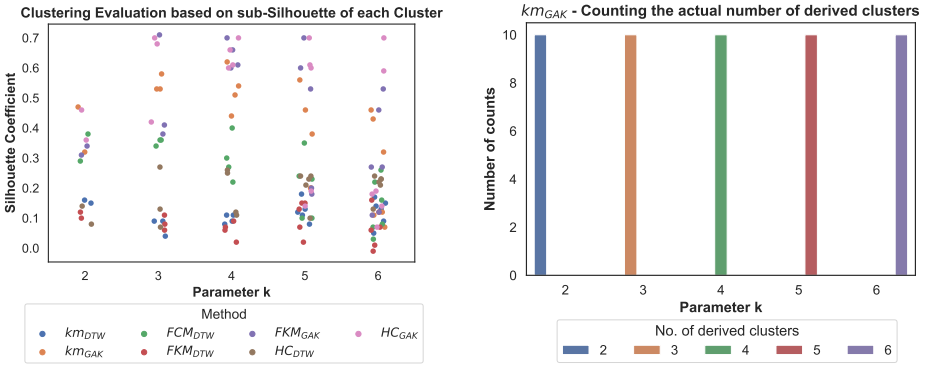
The same analysis can be conducted for the second dataset with $L_n = 0.8$. The three initially introduced evaluation measures are presented in Figure 4.13, while the two additional aspects of Silhouette and Stability are shown in Figures 4.15a and 4.15b, respectively. When considering the first three measures, only the kernel-based clustering methods found the correct number of clusters. Although the AMI values are close to 1, the obtained Silhouette coefficients are in the range of $0.5 - 0.7$, with FKM_{GAK} and HC_{GAK} achieving the highest values. Similarly, km_{DTW} and FKM_{GAK} reveal the highest stability, where the latter one reaches 1 for $k = 4$.

Regarding the second set of evaluation measures, clusters' Silhouette analysis and Stability-derived counts are shown in Figures 4.15a and



(a) (b)

Figure 4.14: $L_n = 0$: (a) Clustering evaluation through Silhouette coefficients for individual clusters, examined for various clustering methods and number of clusters k. (b) Clustering evaluation through the distribution of the actual number of clusters derived across 10 iterations for different values of k.



(a) (b)

Figure 4.15: $L_n = 0.8$: (a) Clustering evaluation through Silhouette coefficients for individual clusters, examined for various clustering methods and number of clusters k. (b) Clustering evaluation through the distribution of the actual number of clusters derived across 10 iterations for different values of k.

4.15b, respectively. Compared to $L_n = 0$, it is obvious that clusters' Silhouette coefficients start revealing meaningless clusters from $k = 2$ for some methods. By checking the colors of those clusters, we observe that they were produced by non-kernel methods. Clusters' coefficients of kernel-based methods start decreasing after setting $k > 4$. Finally, regarding the actual counts of clusters across iterations of each method,

no useful information can be extracted. Although this measure could not confirm the true number of clusters, all other figures verified that kernel-based methods can efficiently uncover the true underlying groups in data.

4.5.3. APPLICATION ON A REAL-WORLD DATASET: NSMD

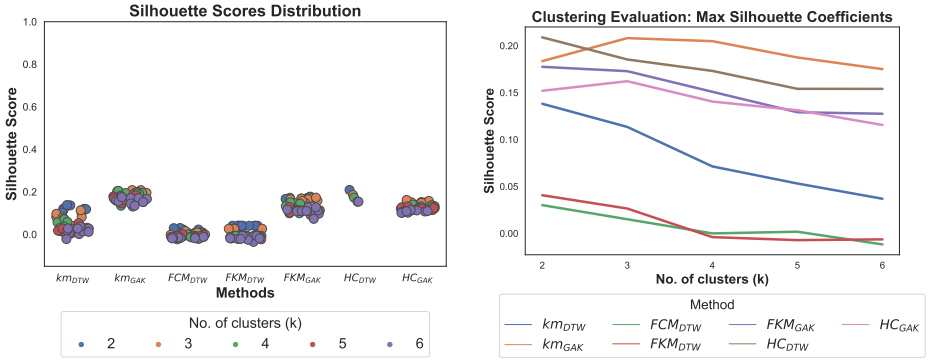
In this section, an additional real-world dataset is utilized to demonstrate all the clustering-related decisions discussed in the previous sections. The dataset used is the NSMD dataset, a real-world dataset described in Chapter 2, Section 2.6. Clustering is performed on the 187 individuals based on their 12-variable time-series. Following, clustering results are evaluated by assessing both cluster quality and stability.

First, clustering is performed using all 7 algorithms, k-means (km_{DTW}, km_{GAK}), FCM (FCM_{DTW}), FKM (FKM_{DTW}, FKM_{GAK}), and HC (HC_{DTW}, HC_{GAK}). Both distance metrics, DTW and GAK, are examined, except for FCM where only the DTW is used. This is because extracting cluster centroids in the original dimensions becomes challenging with kernel-based methods due to kernalization. The γ hyperparameter of the GAK kernel relies on the given data and it is calculated as the average of the median of all distances [85]. Then, the groups derived from each clustering method are evaluated in terms of the Silhouette coefficient and stability, as the true number of underlying clusters is not known. This evaluation helps determine both the optimal number of clusters and the quality of the clusters obtained.

Regarding the Silhouette analysis, the overall results are presented in Figure 4.16a, while the maximum values are highlighted in Figure 4.16b. Notably, the k-means method with a GAK kernel, extracting 3 clusters, gives the highest score at 0.21. This score remains quite constant across different clustering repetitions, potentially also leading to high stability. A similar Silhouette score is produced by HC_{DTW} with $k = 2$. Following these, the next best options are given by FKM_{GAK} and HC_{GAK}, both using a small number of clusters. Therefore, these findings suggest that when kernel-based methods are utilized, the quality of the retrieved clusters improves, showing that kernel transformations are necessary to better represent the complex structure of EMA data.

Apart from these, the remaining clustering methods show lower Silhouette scores, with FCM_{DTW} and FKM_{DTW} approaching zero. Very low or negative Silhouette scores are typically interpreted as not-so-meaning clustering results. In such cases, individuals within clusters may not be significantly closer to each other than to individuals in other clusters, reducing the interpretability of these clustering solutions.

Next, we assess the stability of the clustering-derived groups by examining the consistency of the Silhouette scores as well as the Stability Index. Stability is often questioned because of the random initialization effect in many clustering algorithms. Particularly, the Silhouette scores distribution and the Stability index were computed across 10 iterations



(a) (b)

Figure 4.16: Overall clustering evaluation for all methods through Silhouette scores: (a) Distribution of Silhouette scores over several iterations. (b) Maximum Silhouette scores.

(or repetitions) of each algorithm, as presented in Figures 4.16a and 4.17, respectively. For this analysis, HC_{DTW} was not included as it is independent of initialization issues, while for HC_{GAK}, only the parameter γ was varied. The impact of this variation proved negligible, with stability consistently close to 1.

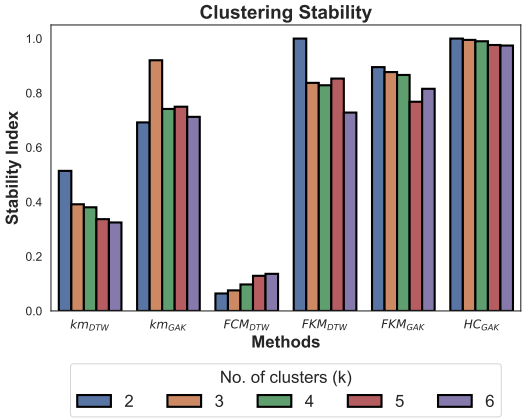


Figure 4.17: Overall clustering evaluation for all methods: Stability Index.

Besides HC, the most stable clustering result is produced by FKM, whereas the least stable by FCM_{DTW} and km_{DTW}. A high Stability index shows that group assignments remain relatively consistent across repetitions. An interesting case is km_{GAK} with $k = 3$, which achieves a score approximating 0.92, which is quite higher compared to the other methods. As already discussed, this finding is also reflected in Figure 4.16a, giving the highest Silhouette score. Given the agreement between both

evaluation measures, this particular grouping is further investigated in Chapters 6 and 7 of this dissertation.

Summarizing, from a methodological perspective, a variety of algorithmic choices, distance metrics, and evaluation methods are available, each leading to a different result. The variety in these choices underscores the inherent flexibility and complexity in clustering analysis. While it is important that a consistent identification of the optimal number of clusters across methods is identified, this does not necessarily imply that individuals are assigned to groups in the same way. This variability arises because different clustering methods and distance metrics can interpret the structure and similarities within the data differently, leading to unique group assignments for the same dataset. This is further reflected in the current evaluation when varying stability results are obtained. Additionally, consistency between different evaluation measures, such as the Stability Index and Silhouette score, is crucial. An agreement between these metrics provides greater confidence in the reliability of the clustering result regarding the grouping of individuals.

4.6. DISCUSSION - RECOMMENDATIONS

Based on our experiments, we can derive conclusions and make preliminary recommendations for choosing a clustering method when grouping individuals in EMA studies. Although our analysis is based on artificially generated datasets, which are less complex than real-world EMA datasets, they can still give valuable insights for real-world applications and future studies, since the scenarios explored were very extensive.

4.6.1. DIFFERENCE IN PERFORMANCE OF CLUSTERING METHODS

Through simulated datasets, this chapter aimed to assess the performance of several clustering methods using different distance metrics and other parameters. First, clustering performance was validated against predefined (true) labels. Overall, a good clustering performance was confirmed for all methods when there was no or a low level of noise in the data, in terms of the percentage of noisy features, also in the presence of missing data. However, when the proportion of noise increased to 0.8, results showed substantial differences between clustering methods. Although GAK kernel-based methods still yielded good results, performance significantly dropped for non-kernel methods, that is, methods that are based on DTW distance. At high noise levels, increasing the percentage of missing data also impacted performance, slightly decreasing the results of kernel methods. Next, similar patterns of results were obtained when examining clustering performance through Silhouette coefficients and Stability.

No significant differences in performance were observed between the various kernel-based methods. According to all three evaluation measures, the results of k-means were only slightly decreased compared to

the other FKM and HC. Between the last two, no distinction was observed, even if one belongs to the hard clustering methods and the other to the soft methods.

4.6.2. CHOOSING THE MOST APPROPRIATE CLUSTERING-RELATED PARAMETERS

As already stated, choosing the most appropriate clustering hyperparameters plays a key role as they can heavily affect clustering performance. According to the findings of our analysis, some preliminary directions can be set to support more efficient parameter selection.

Because clustering is an unsupervised task, it is necessary to compare the performance of different clustering methods to opt for the most suitable approach for a particular dataset. It is common for a clustering method not to be universally applicable to all types of datasets. So, a thorough comparison is necessary between different methods for datasets with different characteristics. To evaluate all these, Silhouette and Stability analysis can be of great value to help identifying some "optimal" choices. Therefore, each combination of methods and parameters should be compared and assessed based on both evaluation measures - Silhouette and Stability analyses. Although the best-case scenario is for both measures to agree, there are datasets for which the two "best" choices can differ. Then, the actual Silhouette and Stability values should be examined, because their level also plays an important role. Even if these evaluation measures agree on a specific value of k, the retrieved scores need to be quite high to lead to a reasonable grouping. For example, a clustering whose Silhouette coefficient is close to 0 reflects a not-so-meaningful grouping. However, according to the simulation results, Silhouette cannot reach values higher than $0.7 - 0.8$ if datasets include a substantial number of noisy features, or even lower (averagely around 0.5) in case of 20% missing data points.

4.6.3. EFFICIENT APPROACHES TO REAL-WORLD EMA DATASETS

Real-world EMA datasets are complex-structured data, nesting temporal information of multiple items within each individual. Although this was captured to some extent in simulated datasets, the patterns of the items (features) collected are not expected to have distinctly clear shapes. It is more common for the real EMA items to more closely resemble the noisy features examined in our simulations. That is, based on our simulation design, a real-world EMA dataset should be represented by a simulation dataset with a high proportion of noisy features. Moreover, missing data is a prevalent issue in real-world datasets. Through data pre-processing, individuals' compliance is always checked so that those with a high percentage of missing points are omitted. However, small percentages are usually acceptable. Therefore, irregular time-series data, having up to 20% missing data, were also thoroughly explored.

In similar investigated scenarios, a GAK kernel transformation appears to be a promising first step when cluster-analyzing EMA data. This type of data transformation is needed to better represent the data structure that can be exploited by clustering methods at a following step. Therefore, GAK kernel-based clustering methods are considered the most appropriate approach to partition the EMA data of several individuals.

4.7. CONCLUSION

This chapter investigates various clustering methods and clustering-related parameters by analyzing data from simulations. The simulations cover different scenarios that mimic real-world cases, involving multiple individuals, noisy features and/or irregular time-series data. Experiments showed that all methods achieved a good performance when applied to datasets with few or no noisy features. However, for datasets containing 80% noisy features, with or without missing data, only GAK kernel-based methods provided a good clustering result. This result indicates that employing alternative representations for EMA data has great potential to better capture its unique characteristics and underlying patterns.

Moreover, when the true clusters are unknown, clustering evaluation through Silhouette coefficients and stability provides valuable insights into the effectiveness and consistency of the examined clustering approaches. Nevertheless, it is important to note that in most cases there is no definitive answer regarding the best clustering method, as different methods aim to achieve effective groupings based on their respective algorithmic processes. Therefore, it remains crucial to deliberately pick the appropriate clustering method, parameters and distance metric for every new dataset and clustering problem.

In the next chapter, the focus remains on clustering approaches, where we further investigate alternative methodologies. Beyond clustering applied directly to raw time-series data, clustering can rely on various individual information or characteristics. For instance, key characteristics of individuals could derive from their personalized models. Specifically, utilizing the underlying models to represent each individual allows for clustering based on the parameters of these models rather than the raw data itself. This approach abstracts the data into essential model parameters, reducing sparsity and enabling the identification of groups that are homogeneous in their underlying statistical properties. This is particularly useful for complex and dynamic datasets, such as EMA. Moreover, it is interesting to investigate how clustering-derived group-based information could be further exploited in the modeling process. It is likely that clustering information could help in improving the performance of personalized models by providing insights regarding specific group characteristics. Thus, Chapter 5 will further evaluate clustering efficacy within the context of a downstream predictive/forecasting task.

4.8. SUPPLEMENTARY MATERIAL

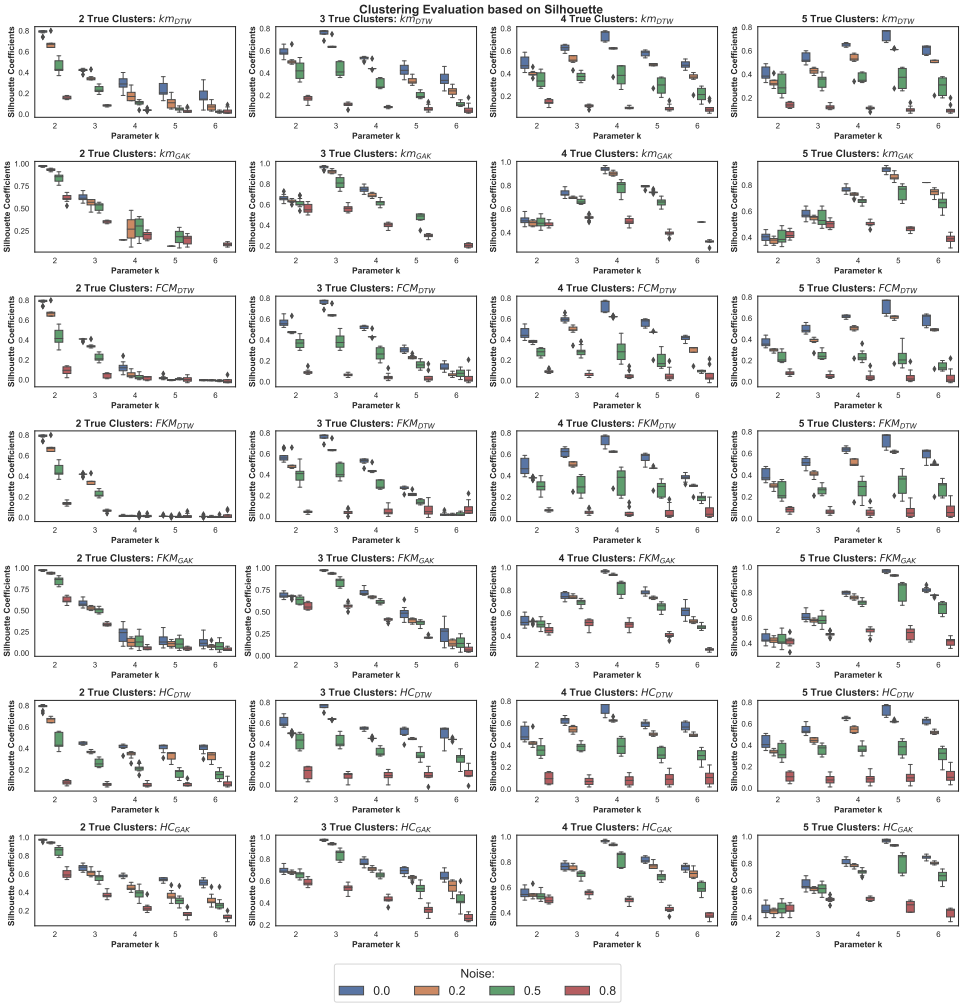


Figure 4.18: Influence of noise L_n (represented by a different color) on the overall performance of all clustering methods assessed through Silhouette coefficients. There are 7 by 4 subplots, where the rows represent each of the clustering methods, while the columns are the true number of clusters.

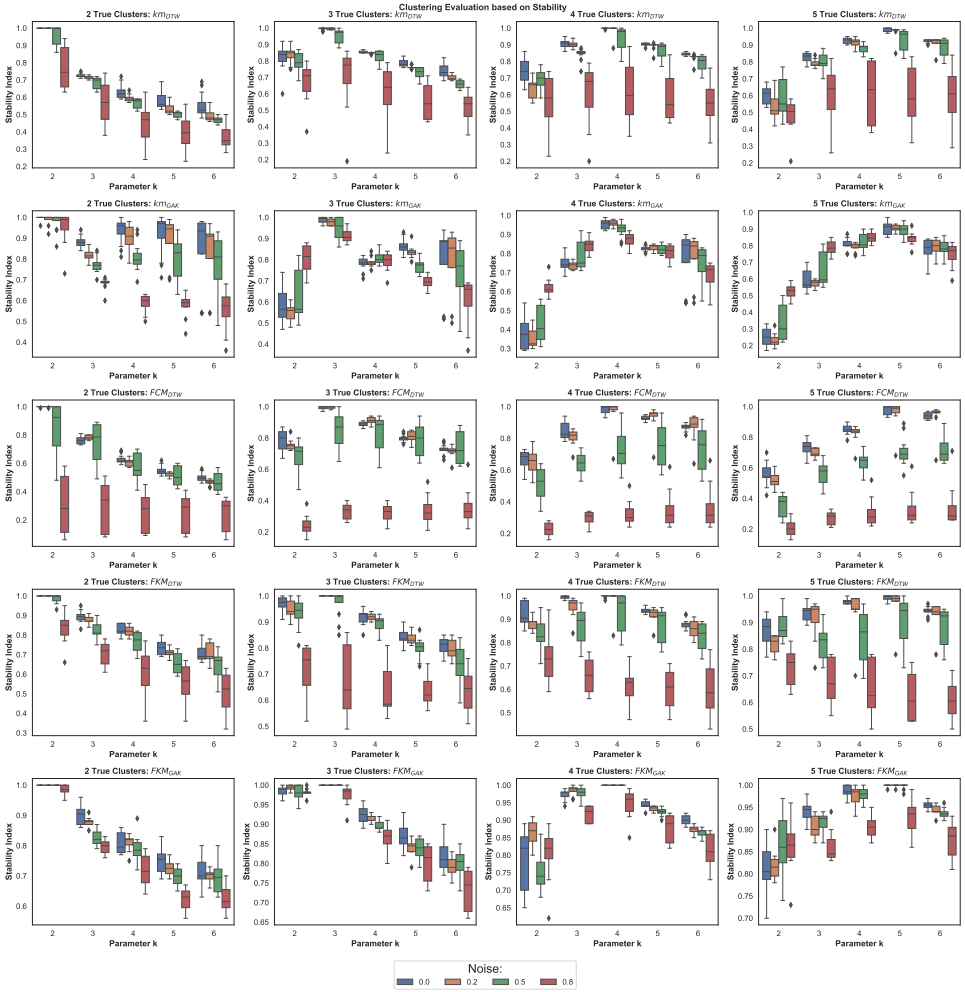


Figure 4.19: Influence of noise L_n (represented by a different color) on the overall performance of all clustering methods assessed through Stability. There are 5 by 4 subplots, where the rows represent each of the clustering methods, and the columns the true number of clusters.

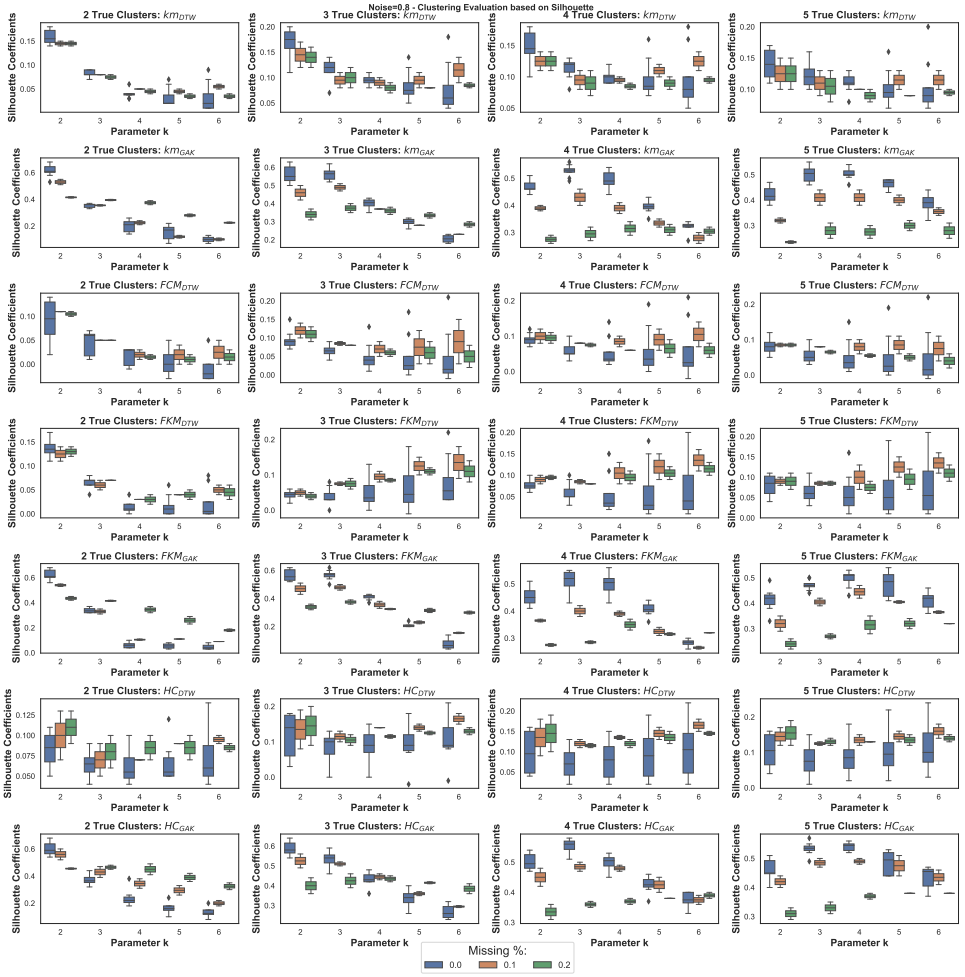


Figure 4.20: Influence of maximum percentage of missing data P_m (represented by a different color) on the overall performance of all clustering methods assessed through Silhouette coefficients. There are 5 by 4 subplots, where the rows represent each of the clustering methods, and the columns the true number of clusters.

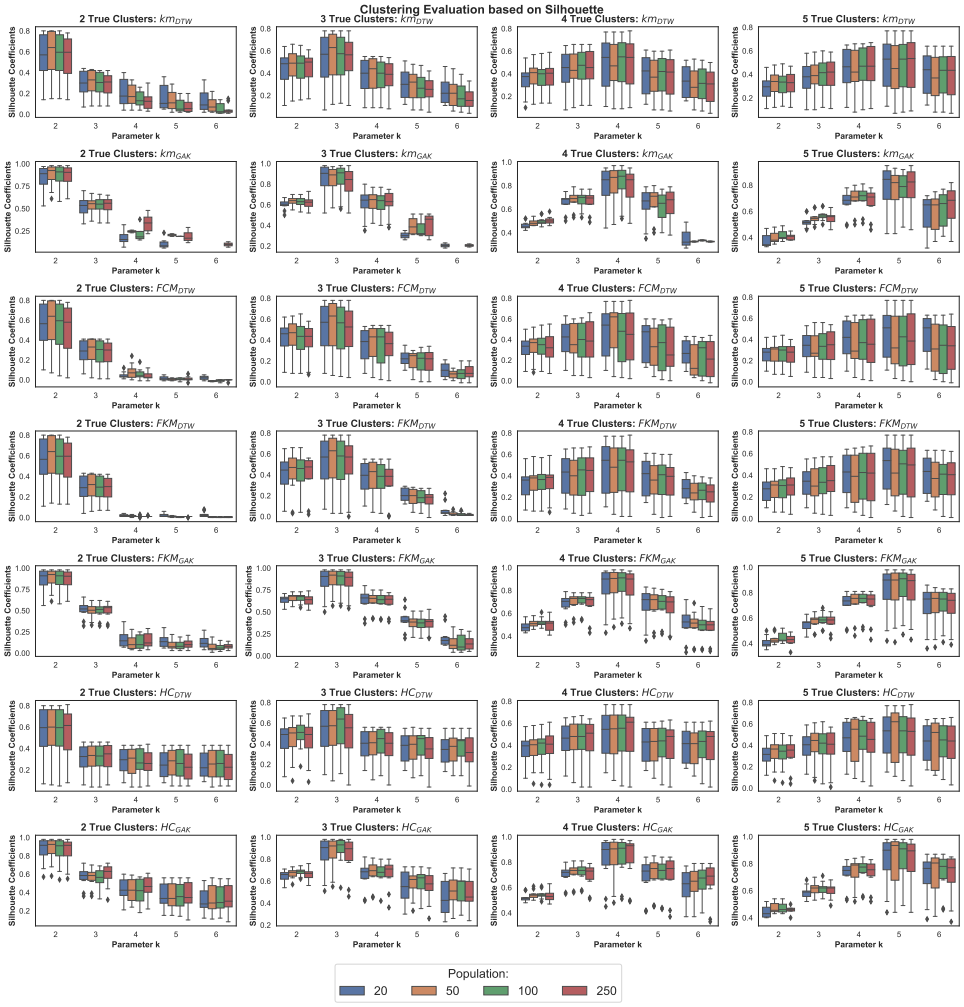


Figure 4.21: Influence of population (represented by a different color) on the overall performance of all clustering methods assessed through Stability. There are 7 by 4 subplots, where the rows represent each of the clustering methods, while the columns are the true number of clusters.

5

GROUP-BASED APPROACHES THROUGH MODEL-BASED CLUSTERING

Having established the potential of clustering individuals using EMA data to gain a better understanding of mental disorders, particularly regarding individuals' commonalities and differences, this chapter focuses on using a different representation basis for clustering. Beyond applying clustering based on raw time-series data, this can be investigated using alternative sources of information, such as model-derived information, that capture distinct aspects of individuals' profiles. More specifically, model-based clustering approaches utilizing information derived from personalized models, such as model parameters, to represent each individual are examined to group the most similar individuals. Subsequently, the clustering results are assessed through group-based predictive models. It is hypothesized that supplementing personalized or individual models with additional information on similar individuals is likely to enhance the predictive performance as well as the description of each individual through derived group information.

5.1. INTRODUCTION

Advancements in technology using smartphones and sensors have offered new opportunities in collecting more and more EMA data for longer time periods for each individual. However, due to factors such as lengthy

Parts of this chapter have been published in

- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Model-based Clustering of Individuals' Ecological Momentary Assessment Time-series Data for Improving Forecasting Performance". In: BNAIC/ BeNeLearn 2023: Joint International Scientific Conferences on AI and Machine Learning. 2023

collection periods or frequent sampling, missing data remains a common issue [156]. Insufficient data leads to the fact that too little data is not well-representative of individual patterns, and consequently, it is sometimes not enough to train accurate and reliable personalized models.

Although it is known that every individual is unique and exhibits their own symptoms and behaviors, it is likely that shared patterns can also be found in groups of people. A way to approach this issue could be through nomothetic approaches, providing information collected by other individuals during the same EMA study. However, instead of using all available data, refining this approach by selecting only those individuals with similar EMA profiles and underlying patterns. By focusing on modeling these relevant groups separately, it is possible to get great insights for better describing and understanding the profiles of single or groups of individuals, uncovering hidden structures as well as building more accurate predictive models for short-term changes in EMA variables.

As already introduced in Chapter 4, finding similar patterns among elements, or individuals in the current setting, when the true grouping is not available, could be uncovered by clustering [124]. Clustering has been studied a lot, with a great interest in time-series data in recent years [127, 157]. Although most straightforward and popular clustering approaches use raw time-series data and further research the most appropriate similarity/distance measure, other types of representational information can also be used to characterize each individual. For example, model-derived information, such as model parameters, could also be utilized, reflecting another promising clustering approach, that is model-based clustering [75].

In model-based clustering, since each individual is described by a personalized predictive model, the objective is to identify groups of similar models that, in turn, represent similar groups of individuals. In this case, the raw time-series data is still used, to build the prediction models, not directly for clustering. Then, clustering is applied to information derived from these models. This model-specific information may include different characteristics of each personalized model. For instance, for linear models, it can be the extracted coefficients of the trained models. Thus, identifying similar sets of coefficients could be useful for uncovering similar individuals.

This chapter aims to investigate the use of model-based characteristics or information for clustering high-dimensional time-series EMA data, through two different approaches. First, model-derived parameters, such as coefficients or feature importance of personalized models, are exploited for applying clustering [75]. Second, since one of the clustering goals is to improve the forecasting performance, performance could be also considered as an alternative information used to optimize clustering [158]. To evaluate both clustering approaches, all clustering scenarios are first assessed on some intrinsic evaluation measures, such as Silhouette coefficients and stability. Second, clustering is evaluated through the performance of the clustering-derived group models. For perfor-

mance evaluation, clustering methods are also compared to three baseline scenarios, such as personalized, using-all-data, and random group-based approaches [111, 112].

5.2. RELATED WORK

Time-series clustering has been studied a lot lately, with some significant reviews detailed in [127, 157]. Building on the related work regarding time-series clustering presented in Chapter 4, Section 4.2, this section aims to further enrich the research field by exploring model-based clustering approaches specifically tailored to time-series, which is the emphasis of this work.

Beyond using the raw time-series information, different data representations can also play a key role when applying clustering [159]. These representations often include statistically derived features, dimensionality reduction techniques like Principal Component Analysis (PCA), or other transformation-based methods, all aimed at capturing the underlying time-series dynamics in a simplified yet meaningful way. Once these features are extracted, clustering techniques can then be applied to group similar data points effectively [160].

Similarly, model-derived features and characteristics can serve as an alternative approach to represent time-series data. This approach falls under the category of model-based clustering. Instead of relying solely on raw time-series data, this approach utilizes features extracted from predictive or generative models that encapsulate the key dynamics of the data. According to two review papers on time-series clustering, model-based approaches can be broadly categorized into two main groups: parameter-based and mixture-based methods [127, 157].

5.2.1. CLUSTERING BASED ON MODEL PARAMETERS

The first approach starts with the assumption that each individual's data can be reliably described by a model, so that it can be represented by the model's estimated parameters. As a result, the problem of finding the most similar individuals is translated into finding the most similar parameters among the different models. In this case, different types of parameters can be used, depending on the applied base model. For example, parameters can be easily extracted from linear models, such as the Autoregressive (AR) model or Autoregressive Integrated Moving Average (ARIMA) [161], where the autoregressive coefficients are then used. Various transformations of these can also be exploited by clustering [162]. Similarly, using probabilistic models, like Hidden Markov Model (HMM) [163], clustering is applied to the probability densities derived from individual HMM. In the case of non-linear models, another type of parameter can be produced. For instance, for Random Forest (RF), measures relying on feature importance values have been proposed [164, 165].

Recently, with advancements in deep learning, time-series embeddings derived from neural networks have become increasingly popular [131, 166]. Techniques such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and advanced Autoencoders can generate latent representations of time-series data, often referred to as "embeddings". These embeddings are more complex internal model representations, capturing both the temporal dynamics and local data structures [167]. For instance, in [168], the parameters derived from the output layer of an RNN model are utilized as the input for clustering, summarizing the entire time-series into a fixed-length feature vector while preserving its sequential dependencies.

5.2.2. MIXTURE-BASED CLUSTERING

The second category aims at recovering the optimal data partition by fitting a mixture of group models that better represent the whole set of individuals. More specifically, it is alleged that each individual is optimally described by a group model trained on a set of similar individuals, individuals belonging to the same cluster. In such settings, each group model would correspond to one cluster, with each individual's data contributing to the training of these group models. Commonly, the models used in this approach can be based on statistical and probabilistic methods. This category includes methods estimated by the Expectation-Maximization (EM) algorithm, such as Gaussian Mixture Models (GMM) and HMM. Using EM, the produced mixture of models always results to a soft-clustering solution. Additionally, hard-clustering is possible if, instead of EM, a version of k-means is used. Such an approach was recently studied in [75]. The proposed K-Models paradigm aims to fit a separate model for each cluster, tested on AR and ARIMA.

However, it is worth highlighting that all these approaches are not easily adaptable to multivariate time-series data of unequal lengths across multiple individuals. This presents a unique challenge, which this chapter addresses by tackling a clustering problem involving multivariate time-series of multiple individuals. Moreover, another goal that the current chapter explores is the application of both aforementioned approaches with more advanced non-linear models.

5.3. METHODOLOGY

This section starts by providing some details regarding the personalized forecasting models that play a key role in the proposed clustering procedure. Following, we present thoroughly the two model-based clustering approaches we focus on this work. While the first one is based on the parameters of already trained individual-forecasting models, the second is focusing on training group models as part of the clustering procedure with the goal to find the most representative group models for all individuals.

5.3.1. INTRODUCTION TO PERSONALIZED FORECASTING MODELS

Initially, we establish one of the main procedures in the examined approaches besides clustering: the training of forecasting models. In the current setting, the prediction task of the models used is forecasting. It is important to highlight the key role of the forecasting models since the whole clustering procedure and clustering evaluation depend on them.

Forecasting models aim to accurately predict the future of EMA responses. However, within EMA data, the task becomes more complicated by aiming for the 1-lag future values of all variables. Specifically, in a multivariate time-series setting with V variables, we need to build V independent models, each using the same input predictors, the set of all variables, and predicting one of the examined variables shifted in the future. This setup is illustrated in Figure 5.1. Consequently, when we refer to one individual or personalized forecasting model, this implies all V independent sub-models necessary for predicting all V variables in the future.

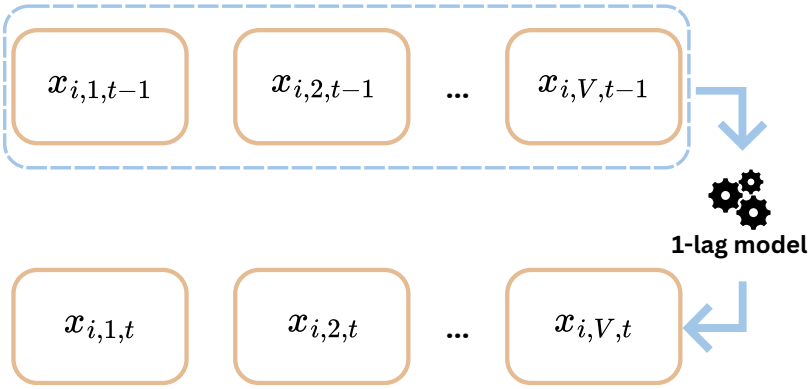


Figure 5.1: Personalized 1-lag Forecasting Model: The input includes all variables at time-point $t-1$, denoted as $x_{i,1..V,t-1}$, while the output corresponds to all variables at time-point t, denoted as $x_{i,1..V,t}$.

5.3.2. MODEL-BASED CLUSTERING APPROACHES

This section presents the two examined model-based clustering approaches. While the first approach focuses on clustering based on model parameters, the second aims at building a mixture of representative group (or cluster) models by clustering individuals using their predictive performance.

APPROACH I: PARAMETER-DRIVEN CLUSTERING (PDC)

In the first approach, which we refer to as Parameter-Driven Clustering (PDC), clustering is performed using model-derived parameters repre-

senting each individual. In particular, this approach considers the parameters extracted by the N personalized models as the input of clustering, as presented in Figure 5.2. In other words, it uses the parameters of trained personalized models to represent every single individual, assuming that these models can accurately describe them. For each individual, V variables need to be predicted. Given the EMA structure, where each variable can be predicted based on the values of all variables from the previous time-step, this is achieved through V separate models. So, the parameters of all V independent models need to be concatenated in order to better represent each individual.

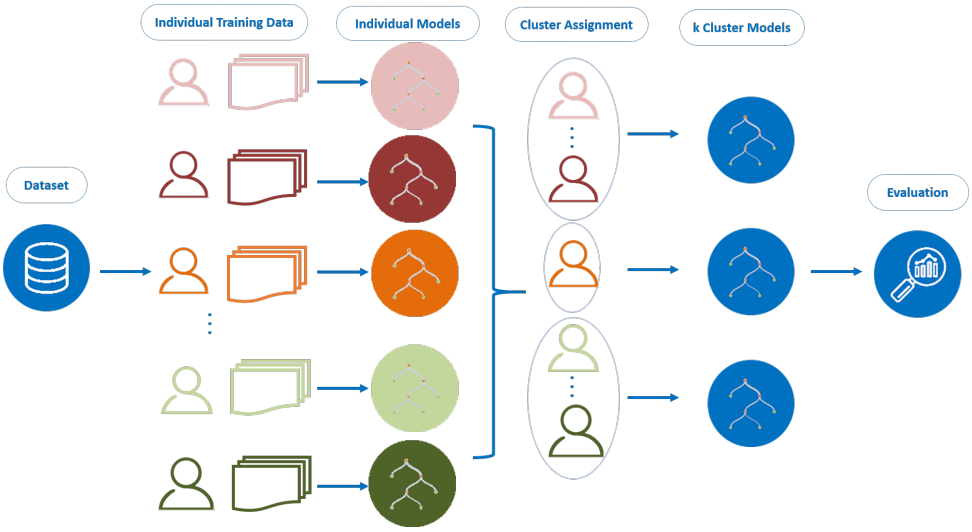


Figure 5.2: Clustering approach I: Parameter-Driven Clustering (PDC). This approach considers the parameters extracted by the N personalized models as the input of clustering.

The type of parameters depends on the chosen base model that is used. Here, different base models are applied, both linear and non-linear ones. In the case of linear models, the fitted coefficients can be easily extracted. Each coefficient indicates the influence of one variable in predicting the future values of another. So, to cover all combinations, finally $V \times V$ coefficients are used for representing one individual. Finally, the coefficients' matrix, that is the set of N by $V \times V$ parameters, is input to clustering.

For non-linear models, such coefficients are not inherently available. Thus, similarly, trying to quantify the influence of one variable on another, feature importance values can be used. In this scenario, the number of parameters for each individual has remained the same, $V \times V$, as well as the total parameters matrix, N by $V \times V$.

APPROACH II: PERFORMANCE-OPTIMIZED CLUSTERING (POC)

In the second approach, clustering is applied using different model-derived information related to predictive performance, that is assumed to alternatively describe the individuals. In this case, clustering aims at building representative global or cluster models, each consisting of similar individuals based on their forecasting performance [158]. Each cluster is optimized on the total test performance of its individuals, in terms of the Mean Squared Error (MSE) across all variables and time-points in their test set.

In particular, the procedure is quite similar to the original k-means algorithm, with the main difference found in the objective function. The goal now is to minimize the MSE errors on the test set of all individuals, instead of minimizing the within-cluster distance of all individuals to the centroid. In more detail, the different steps are summarized in Figure 5.3, and described as follows:

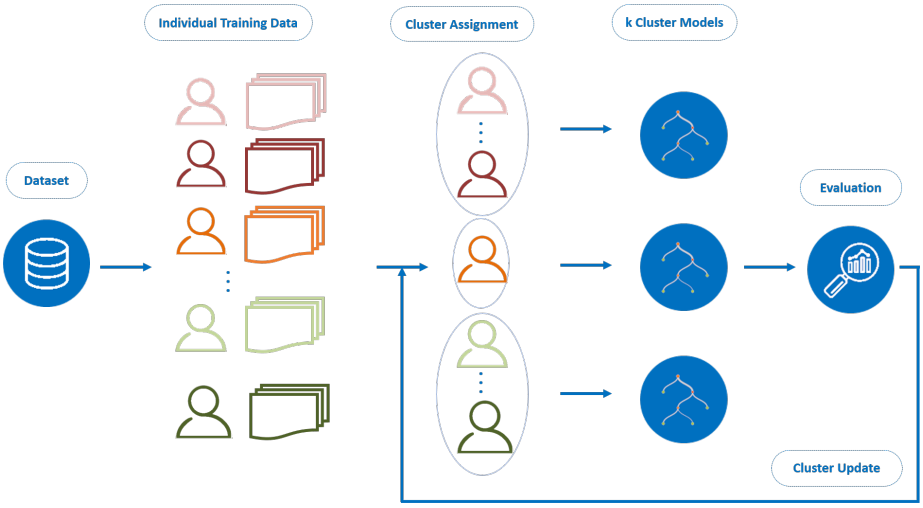


Figure 5.3: Clustering approach II: Performance-Optimized Clustering (POC). This approach is optimized for forecasting performance, ultimately aiming at building k representative cluster models.

- Initialization Step: The procedure starts with a random initialization, setting the centroids of the clusters. For faster convergence, we follow an approach similar to the k-means++ algorithm. More specifically, after the first centroid is randomly selected, the rest should be set in a way to be the most dissimilar to the first one. To find the most dissimilar individual to the centroid, it is assumed that its performance on a model based on this individual-centroid should be the worst. So, we start by building a model trained on the first centroid (centroid model) using the data of the first selected individual-

centroid. Then, this model is tested on all the remaining $N-1$ individuals and the one with the highest test MSE error is then selected as the individual-centroid of another cluster. If the predefined number of clusters (k) is greater than 2, this process is repeated until the centroids of all clusters have been found.

- Clusters Assignment Step: To assign the rest of the $N-k$ individuals to a cluster, again their performance is used as a measure. Instead of calculating the actual distance to the clusters' centroids, their performance is examined, in terms of MSE error, on the k centroid models. As a result, each individual is assigned to the cluster with the minimum error.

- Forward Step: Having assigned all individuals into clusters, the main procedure begins with the goal to optimize the objective function of the predictive algorithm. This is described in Equation 5.1, where the total test MSE error L_{MSE} corresponding to all N individuals need to be minimized, meaning the error of every single individual. To achieve this, the cluster models are built using the training data (the first part of a dataset) of the individuals belonging to each cluster. Then, all the cluster models are assessed on the test set (the last part of each dataset, where time-points range from 1 to T) of all individuals, predicting all V variables.

$$L_{MSE} = \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \sum_{v=1}^{V} (x_{i,v,t} - \hat{x}_{i,v,t})^2}{T \cdot V} \tag{5.1}$$

- Update Step: For each individual, its test performance is compared on all the cluster models. If the test MSE error is smaller in another cluster than the one already belonging to, the individual moves to the other cluster. This way, all the clusters' composition is updated.

The last two steps, forward and update, are repeated until there is a convergence, meaning that individuals are fixed in clusters, or the maximum number of iterations has been reached. When the clustering algorithm stops, the clusters' composition has been finalized, the group or cluster models are trained and the total minimum loss has been found. However, beyond the initialization step, clusters' centroids are not directly used or calculated.

5.4. EXPERIMENTS

In the following section, the results of the two model-based approaches are presented. First, both are tested using different base models, such as Explainable Boosting Machine (EBM) [106] introduced in Chapter 2, Section 2.4, where clustering and final forecasting performance are examined. Then, the final test performance of the clustering approaches is

evaluated on the last 30% of each individual data and compared against some baseline approaches. These include (1) the personalized, also referred to as the N-Clustering problem, (2) the using-all-data forecasting models, or 1-Clustering, and (3) group models where individuals were randomly assigned, or Random-Clustering.

5.4.1. EXPERIMENTAL SETUP

In the analysis, the following set of experiments is investigated on the real-world NSMD dataset, which is detailed in Section 2.6. To assess the clustering performance, three baseline scenarios are examined: N-Clustering (or personalized), 1-Clustering (all individuals belong to one group) and Random-Clustering (random-groups clustering), as follows:

- k-Clustering: During k-means clustering, the number of clusters (k) needs to be predefined. Here, the values of k are set in a range from 2 to 20. When k is set, we compare the results with a different base forecasting model, where here there is a set of 3 different models. The set includes one linear, the Vector Autoregressive (VAR) model, and 2 non-linear ones, Random Forest (RF) and Explainable Boosting Machine (EBM) [106]. For each combination of k and forecasting model, k-means clustering is then repeated for 10 iterations.

- N-Clustering: The total forecasting performance after clustering is compared with the case of having personalized models. So, for each individual, a separate model is trained based on their own data and assessed on its unseen test data. Having N personalized models for N individuals is identical to the case clustering using $k = N$. Thus, the concept of using personalized models is also called N-Clustering.

- 1-Clustering: In a similar manner, another baseline scenario under investigation involves the case of $k = 1$, which is also referred to as 1-Clustering. Then, it is assumed that all individuals belong to 1 cluster, so that only one model could describe them all after being trained on all data. For a fair comparison, its performance is tested separately on the individuals' test data.

- Random-Clustering: To ensure that the effect on clustering performance is not caused just by the fact that fewer individuals than all, but also more than one, are used in a model, some additional experiments are added. The case of randomly assigning individuals to different clusters is then examined. The same range of k values is again used and 10 iterations are executed for each experiment.

5.4.2. EVALUATION

After applying clustering, there are several ways to evaluate the derived results. Here, the evaluation is conducted through the quality of clus-

ters (intrinsic measures) as well as the performance of a downstream forecasting model.

- Clustering Evaluation: For clustering evaluation, either intrinsic or extrinsic measures [152] are mostly used. Because the latter requires obtaining ground truth labels, which are not available in our case, the focus of this work is on the intrinsic measures. The majority of the intrinsic measures are computed using the compactness of each cluster as well as how well different clusters are separated. Similar to Chapter 4, the well-applied Silhouette analysis is selected, since other measures (such as Davies-Boudlin [152]) are mostly based on clusters' centroids, which lack any natural explanation. Silhouette coefficients are given through Equation 5.2 (also in Chapter 4 in Equation 4.8), calculating the average score across all N individuals. For each individual i belonging to cluster C_i, the score compares the average similarity a_i across all individuals of the same cluster (w, where $w = 1..W_i$ and W_i is the total number of individuals of C_i) to the average similarity b_i of the individuals belonging to the closest to C_i cluster (z, where $z = 1..Z_i$ and Z_i is the total number of individuals of the closest to C_i cluster).

$$Sil = \frac{\sum_i^N \frac{b_i - a_i}{\max(b_i, a_i)}}{N}, \text{where} \qquad (5.2)$$
$$a_i = \frac{\sum_{w \epsilon C_i} d(x_i, x_w)}{W_i} \text{ and } b_i = \frac{\sum_{z \epsilon \text{ closest } C_i} d(x_i, x_z)}{Z_i}$$

  Another evaluation measure is the stability of the clustering results [153]. Through repeating clustering multiple times, the stability or agreement of individuals assignment into clusters can be assessed. More specifically, the Adjusted Mutual Information (AMI) between the individuals' cluster labels is used as stability index [169].

- Forecasting Performance Evaluation:

  An additional way to evaluate clustering performance is through the performance of the derived cluster model. Using the data of each of the clustering-derived groups, different cluster models can be trained to aim at forecasting the 1-lag future values of all variables, as described in Section 5.3.1. Then, these can be evaluated using the MSE on the test set of each individual. Particularly, each individual is evaluated on the MSE of their test set data (last 30% of its data) using the cluster model that belongs to. The total MSE of all individuals on their test sets is used as the performance indicator of each clustering method. Similarly, the performance of the baseline methods can be assessed, using the total MSE on the same test sets of all individuals.

5.4.3. RESULTS

The results of the conducted experiments are presented in this section. First, the two proposed clustering methods, PDC and POC, are assessed through different intrinsic evaluation measures. Afterwards, clustering is assessed through its downstream forecasting performance as well as in comparison to three baseline methods, 1-Clustering, N-Clustering and Random-Clustering.

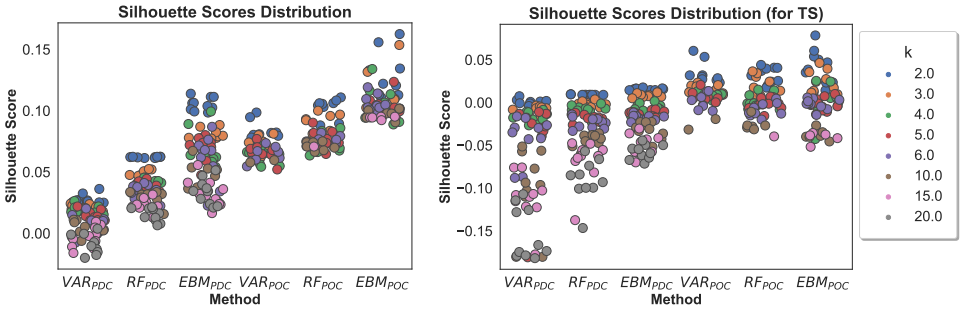
CLUSTERING EVALUATION

As already described in Section 5.3.2, for cluster analysis, or, as we also call it, k-Clustering, two model-based approaches are investigated, based on model parameters (PDC) and performance (POC), respectively. In both approaches, different choices are examined, such as the applied base models, which can be VAR, RF and EBM, while the number of clusters (k) can also take values from the set $\{2, 3, 4, 5, 6, 10, 15, 20\}$. Regarding PDC, the selected base model refers to both individual and cluster models. Then, for both approaches, all combinations of these choices represent the examined experiments, where each one is repeated 10 times.

All these experiments are evaluated using two intrinsic clustering evaluation measures, described in Section 5.4.2, Silhouette coefficients and Stability. First, the Silhouette coefficients of all the experiments of both approaches are depicted in Figure 5.4a. For each method (combination of base model - clustering approach) on the x-axis, all points of the same color represent 10 iterations of each experiment, repeated using a particular value of k. Between the two clustering approaches, the second one, POC, clearly reaches much higher scores than the first one. This holds for all different base models. In detail, EBM gives the highest maximum scores in both approaches, around 0.17 and 0.11, respectively, whereas the VAR models show the lowest maximum scores at 0.10 and 0.03, respectively. Although the second approach mostly outperforms the first one, the difference becomes less significant when using EBM with a low value for k. Thus, EBM models seem to better describe the complex EMA data, even in the form of feature importance values.

It is also noticed that increasing the number of clusters used, the produced Silhouette tends to decrease. Regarding the first approach, there are iterations for EBM indicating that $k = 2$ and $k = 4$ give the best clustering, whereas for the second approach, that occurs for $k = 2$ and $k = 3$.

In a similar way, according to the derived clustering grouping, the Silhouette coefficients calculated based on the DTW distance of the individual time-series data can also be calculated and shown in Figure 5.4b. The found pattern resembles the one of Figure 5.4a, although the Silhouette values are lower for all the experiments. In the first clustering approach, all Silhouette coefficients are below zero, indicating a meaningless clustering. However, referring to the second clustering approach, the Silhouette scores for low values of k are a bit higher, but not exceed-



(a) Based on the model-based parameters used.

(b) Based on the DTW distance of individuals time-series data.

Figure 5.4: The Silhouette coefficients of all experiments of both approaches (PDC and POC) are presented. The difference between the chosen values of k and base models is depicted. The same-colored points represent different iterations of the same experiment (combination of k value and base model).

5

ing 0.07, when using $k = 2$ and EBM.

Next, all clustering experiments of both approaches are assessed for their stability, in terms of cluster labels agreement across all 10 iterations. The experiments' stability, as expressed by the AMI scores, is presented in Figure 5.5. In most of the cases, it is obvious that the stability values are quite low. This is caused by the fact that the produced groups were quite different from each other. Based on the complexity of individual EMA patterns, it seems unlikely for different people to be always separated in a particular way. Thus, it is preferable to work with the group information each iteration separately extracts.

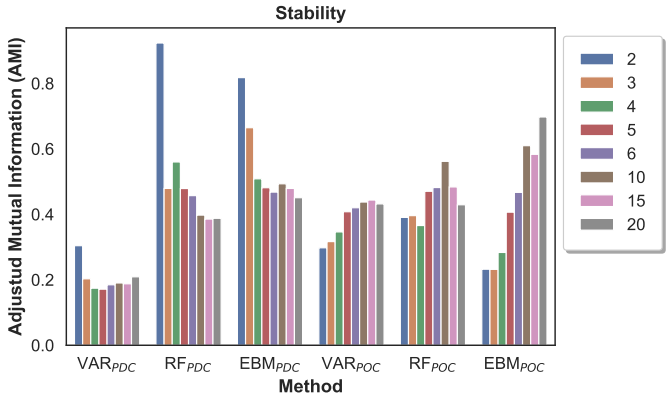


Figure 5.5: Stability of all experiments of both approaches (PDC, POC) is presented.

It is also important to note that, even though k-means requires a prior specification of the number of clusters (given k), the number of produced clusters could deviate from this. So, in the whole analysis, only the iterations of experiments that actually produce the given number of clusters are considered valid. While in PDC, clustering always produces the given number of clusters, for POC the number mostly deviates from the expected. The average number of clusters is displayed in Table 5.1. When the average produced number is smaller than the given one, it means that there are iterations with fewer extracted clusters. This is usually the case when increasing the chosen number of clusters. These are eventually excluded from the analysis of Silhouette coefficients and Stability.

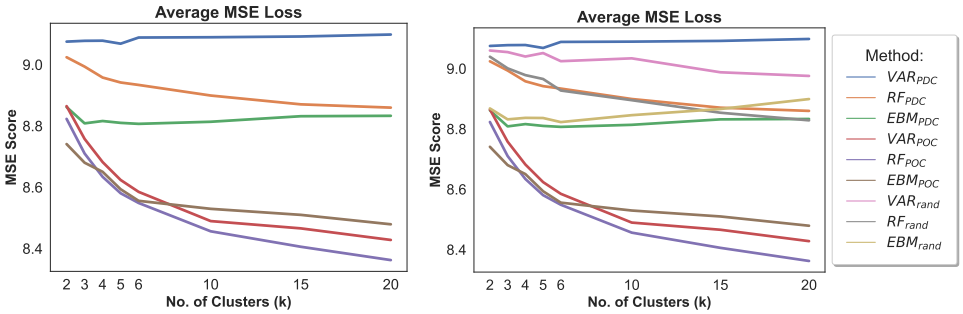
Table 5.1: POC: The average number of clusters, across 10 iterations, for all experiments.

k	2	3	4	5	6	10	15	20
VAR	2.0	3.0	4.0	4.8	5.8	9.2	12.6	13.6
RF	2.0	3.0	4.0	5.0	5.9	9.7	13.6	17.2
EBM	2.0	3.0	4.0	5.0	6.0	9.7	14.4	18.1

DOWNSTREAM FORECASTING PERFORMANCE

As a second step, clustering can be evaluated through the forecasting performance of the clustering-derived group models. First, MSE loss scores are summed across all individual test sets within each experiment to obtain a total MSE for each experiment. These total MSE scores are then averaged over 10 iterations and compared across experiments in the two proposed clustering approaches, as shown in in Figure 5.6a. A clear distinction is again obvious between the two clustering approaches, with POC leading to slightly lower loss scores, which is translated to a better performance.

In more detail, on the one hand, for the first approach, the scores do not vary a lot for the examined number of clusters, leading to a score approximately at 9.07, 9.02 and 8.86 in the cases of VAR, RF and EBM models, respectively. On the other hand, for the second approach, the scores are much lower and not that constant across the different values of k. While for low values of k, the loss indicates that EBM shows the best performance, this changes after $k = 6$, where RF, then, gives the best scores. Although that difference between the base model is not important, the difference across the k values is quite large. For all base models, it starts at around $8.7 - 8.9$ and ends at $8.3 - 8.5$. Thus, increasing the number of clusters, and consequently, that of the cluster models, seems to improve the performance. However, in that case, the number of the actual clusters found tends to be smaller than the



(a) The experiments of the two proposed clustering approaches (PDC and POC) are compared.

(b) The two clustering approaches (PDC and POC) are compared to the random individuals grouping (rand).

Figure 5.6: Total MSE loss of all individuals is given, averaged across all iterations.

5

given one. Therefore, there should be a trade-off between performance improvement and the capability of finding the given number of clusters.

As a general result, the overall loss scores are partly in agreement with the findings after the Silhouette analysis. Both show that the POC outperforms PDC, while clustering, using EBM as the base model, yields mostly the best performance. Nevertheless, the impact of different k values is not consistent. For instance, while performance is optimal at $k = 20$, the corresponding Silhouette coefficient reaches its lowest value, highlighting a potential trade-off between clustering quality and predictive performance.

Because of this disagreement in the analysis, it is essential to further reject the possibility of having enhanced performance results by just using more cluster models, instead of meaningful cluster models. In other words, we need to determine if this improvement could be a random effect of just splitting the individuals in more clusters, leading to an increased number of cluster models. To test this, we split individuals randomly into the same predefined range of possible clusters, without having applied any clustering technique. This refers to the Random-Clustering approach. Again, every experiment is conducted 10 times, and the average total loss scores are exhibited in Figure 5.6b. It is observed that the loss line plots of random grouping follow the patterns of PDC, whereas they are quite worse than the experiments of POC. Thus, it seems that the improved performance is not connected to a random effect of just using more cluster models.

Finally, the two proposed clustering approaches are compared to two baseline scenarios, namely N-Clustering and 1-Clustering, which correspond to the personalized and using-all-data models, respectively. For all these scenarios, the total loss scores of all individuals (again according

to Equation 5.1, averaged on all variables and test time-points for each individual) are given in Table 5.2. As for k-Clustering, both $k = 2$ and $k = 20$ clusters are considered, which, in POC, yield better performance compared to the baseline scenarios. According to the derived scores, the lowest loss is found when using 20-Clustering and RF, reaching 8.36. Among the baseline scenarios, the lowest MSE score is for the personalized EBM, with an MSE of 8.9, which is much better than that of N-Clustering. Whereas, for RF, the reverse trend becomes apparent, where N-Clustering outperforms 1-Clustering. However, in both scenarios, the differences seem almost negligibly small, with improvements reaching a maximum of 7.99% and 7.17% compared to the 1-Clustering and N-Clustering, respectively. Overall, k-Clustering using the POC approach is found to enhance the overall forecasting performance over the baseline approaches, which actually exhibit only slight differences in MSE.

5.5. DISCUSSION

According to the evaluation results, the superiority of the second approach, i.e., clustering based on downstream performance, is apparent over the first approach, which is based on model parameters. This was confirmed by both the Silhouette coefficients and forecasting performance. Regarding the Silhouette analysis, the fact that the produced scores did not exceed 0.17 shows that clustering cannot be characterized as meaningful. However, in real-world datasets, because of their structure complexity, values close to the maximum possible, that is 1, are not realistically expected. Thus, in this case, we should evaluate the scores in comparison to the produced values of all the rest of the examined experiments. Moreover, regarding performance, it is reasonable that lower errors occur in the second approach, as in principle the clustering approach is optimized on the same metric, total MSE.

According to the aforementioned evaluation measures, there was a disagreement in determining the optimal clustering parameters. For example, increasing the number of extracted clusters seems to give a lower performance error, whereas the opposite holds for the Silhouette coefficient. Thus, a trade-off analysis regarding the number of clusters is necessary to be further studied, also on the basis of the specific application domain.

Another critical point of clustering is its evaluation in terms of stability. For the majority of experiments, the estimated cluster labels of all individuals were quite different across the 10 running iterations, leading to low stability values. This means that the initialization step had a large impact on the resulting data partition, although this impact was not always observed in the test performance error. Thus, despite the low stability, similar values for performance loss were found. Utilizing an initialization method similar to kmeans++ only leads to locally optimal solutions with respect to the defined clustering optimization. Therefore, this problem of initialization bias needs to be further investigated in more datasets.

Table 5.2: MSE Loss of all the examined scenarios, 2- and 20-Clustering, as well as 1- and N-Clustering.

	VAR	RF	EBM
2-Clustering (PDC)	9.075	9.024	8.862
2-Clustering (POC)	8.865	8.823	8.741
20-Clustering (PDC)	9.098	8.860	8.833
20-Clustering (POC)	8.428	8.362	8.479
1-Clustering	9.072	9.089	8.904
N-Clustering	9.072	9.008	9.052

Finally, as a general remark, although using the model-based estimated parameters in a clustering task has been widely applied, in real-world problems, more complex data representations may be necessary. It is likely that high-dimensional, dynamic, and possibly noisy time-series data cannot be accurately described only with parameters derived from linear/non-linear equations. Thus, more complex feature representation should be learned using deep learning techniques [166, 168]. Specifically, using such sequential deep learning models, including a greater number of historical values than just one previous time-point, could also be beneficial in training, achieving better accuracy. Enhanced model accuracy often leads to better representations of the underlying data, which can potentially support more meaningful clustering results.

5.6. CONCLUSION

In this chapter, two model-based clustering approaches have been introduced, with a twofold goal (1) to group similar individuals using their EMA data, and (2) to improve the forecasting performance for all individuals. The two approaches utilize different model-derived information, one based on models' parameters (PDC), whereas the other on their overall performance in a forecasting task (POC). Throughout the chapter, various experiments were conducted to assess both approaches along with a set of other important clustering-related choices, such as the number of clusters and the base model used. The evaluation is, first, conducted using intrinsic evaluation measures (Silhouette coefficients, clustering stability) and then on the performance of a downstream forecasting scheme, where each cluster is described by a separate cluster model.

According to the evaluation results, besides stability which was quite low, the remaining measures indicated that the second clustering approach (POC) outperforms the former one (PDC). The POC approach of k-Clustering was also found to enhance the overall forecasting performance, compared to the baseline approaches, achieving a maximum

improvement of 7.99% over the 1-Clustering when using RF. Thus, the results demonstrated that the superiority of clustering performance is not a random effect arising from the fact that a mixture of models is used.

Overall, given that potential benefits have already been found when using clustering-derived information, more advanced integrating approaches need to be explored, expecting to further enhance model performance. Such approaches could offer a balance between clustering and personalized modeling, prioritizing individual data. For example, combining transfer learning methods with clustering presents a promising direction for enriching the modeling process by transferring knowledge from group-level to individual-level. This concept is later explored in Chapter 7 [46, 111].

Having explored a wide range of possible clustering-related methods and options within the context of EMA data in both Chapters 4 and 5, we recognize the importance of good evaluation measures to assess clustering effectiveness. Therefore, in the next chapter, Chapter 6, we shift our focus towards explainability as another potential evaluation measure for clustering. Such an approach could complement traditional evaluation metrics but also offer a critical view for a deeper understanding and more comprehensive validation of the clustering outcomes. For instance, by generating explanations, the structures and key factors of the derived clusters could be identified, offering insights into what differentiates all clusters. Such information could enhance the overall applicability and relevance of clustering, making it more useful for targeted analysis.

6

EXPLAINING EMA CLUSTERING BASED ON MULTIVARIATE TIME-SERIES

With the goal to broaden EMA analysis by incorporating group-based models, that rely on individuals exhibiting similar temporal EMA patterns or characteristics, various clustering methodologies were explored in Chapters 4 and 5. Identifying homogeneous groups of people is of great importance to eventually improve the modeling performance. While some evaluation measures, examining the quality of the derived clusters, have already been explored in the previous chapters, this task remains quite challenging. Therefore, this chapter further investigates additional measures for clustering evaluation. A key component of this evaluation is clustering explainability. To approach this, we propose an attention-based interpretable framework to identify the important timepoints and variables that play a primary role in distinguishing between clusters. Specifically, the goal is to examine ways to analyze, summarize, and interpret the attention weights as well as evaluate the patterns underlying the significant segments of the data that differentiate across clusters.

Parts of this chapter have been published in

- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention". In: Explainable Artificial Intelligence. Ed. by L. Longo, S. Lapuschkin, and C. Seifert. Cham: Springer Nature Switzerland, 2024, pp. 75–99. isbn: 978-3-031-63797-1

6.1. INTRODUCTION

Despite the observed EMA data heterogeneity among individuals, identifying similarities in their patterns can be also valuable, giving insights into more general mechanisms that are valid for particular subgroups. However, as highlighted throughout the previous chapters, it is quite difficult to ascertain if there are such subgroups for which commonalities hold and also how to discover them. While various clustering-based methodologies were investigated in Chapters 4 and 5, uncovering homogeneous groups of people is a complex task. Especially, in real EMA datasets, given that clustering is an unsupervised process and true EMA grouping is not commonly available, evaluating the quality of the derived clustering labels poses significant challenges. It is also expected that, due to the inherent variability of EMA data as well as the complexity of the real-world multivariate time-series (MTS) data, clustering algorithms could produce quite different results leading to different groups. Thus, a critical point to be addressed is evaluating the clustering results.

Beyond the well-applied distance-based criteria capturing the quality of clusters, mostly explored in the previous chapters, clustering interpretability is another important aspect of evaluation. Interpretability ensures that the patterns and common characteristics of the groups identified through clustering can be understood, explained and validated in the context of mental disorders. Thus, uncovering the meaningful patterns of each cluster within EMA data could provide insight into intra-individual psychopathological processes, their temporal patterns, and their interrelationships among theoretically similar subtypes of disorders.

In this chapter, to address the need for explanations on clustering results, the proposed methodology focuses on investigating interpretable deep-learning mechanisms capable of handling MTS data, particularly within the domain of psychopathology. Our approach utilizes advanced deep-learning models, specifically attention-based mechanisms [171], to understand the complexities inherent in MTS data without relying on prior data transformations. Therefore, this approach ensures that our interpretations rely on the actual data dynamics rather than other transformations, providing a clearer and more accurate comprehension of MTS in psychopathology. As depicted in Figure 6.1, our methodology employs a multi-aspect framework that integrates both temporal and feature-level attention. Therefore, it is designed to provide explanations by identifying the important time-points and variables that play primary roles in the domain of psychopathology [172, 173].

While the proposed framework for extracting temporal and feature-level attention is fully described, a significant part of this chapter examines ways to analyze, summarize, and interpret the attention weights as well as validate the patterns underlying the important segments of the data that differentiate between clusters. More specifically, this includes the significant and distinct characteristics in cluster-, feature- and individual-level and their evaluation. Thus, such clustering explanations could prove beneficial for generalizing the existing concepts, uncover-

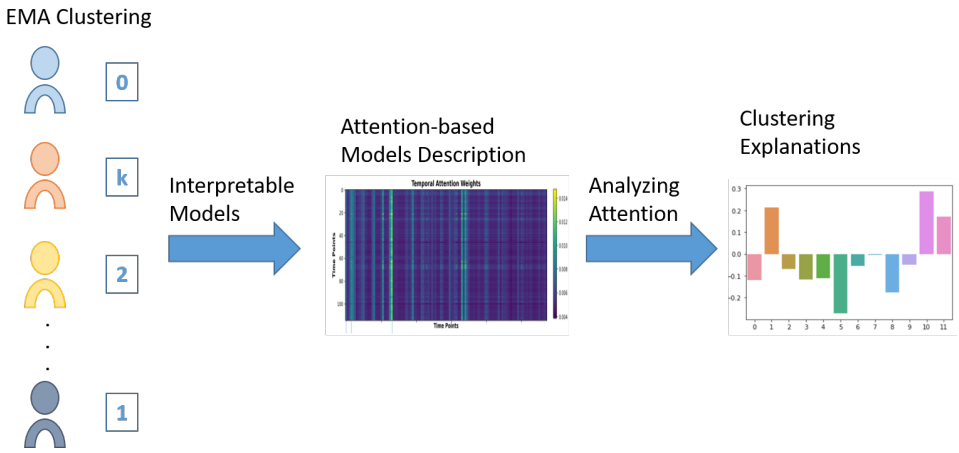


Figure 6.1: An overview of our methodological approach for explaining clustering results using attention-based interpretable models. These models are applied to the clustering results to analyze attention-based outputs, extracting meaningful insights and providing clear explanations of the clustering patterns.

ing new insights into psychopathology and network theory, and even enhancing our knowledge at an individual level. Furthermore, central to our framework is its independence from any specific clustering algorithms. This independence introduces a new theoretical perspective on evaluating the robustness of an examined clustering result, allowing for a more objective comparison and assessment of clustering algorithms. Overall, attention-derived interpretability, beyond contributing to a better understanding of the underlying data structures in psychopathology, could be theoretically used to benchmark clustering effectiveness.

6.2. Related Work

The field of time-series (TS) clustering research has attracted significant attention, with a focus on clustering explanations being introduced the recent years. First, related work on time-series clustering explanations is presented. This involves extracting meaningful representations for clusters' descriptions as well as classical explanation methods in the field of Explainable Artificial Intelligence (XAI).

6.2.1. Clusters' Descriptive Representation

In the context of understanding clustering results, the representation of clusters is crucial in unveiling meaningful insights into underlying constructs (connections in variables) and further facilitating decision-making. This is usually achieved by examining and visualizing all ele-

ments of a cluster. For instance, in temporal data, distinct trend lines can be observed by overlying the time-series data belonging to each cluster. Summary statistics, such as mean values and variance, can also be calculated to describe each cluster. However, the difficulty of distinguishing between them increases significantly when using many clusters or high-volume datasets with multiple variables.

In a more simplified way, a cluster can be represented by a center point or centroid, depending on the clustering methods applied. For example, k-medoids clustering uses a medoid or an actual data point within the cluster as the representative center, whereas k-means uses the average point or centroid. Although this approach performs well when clusters are compact or isotropic (spherical clusters), it falls short when dealing with more complex clusters [174]. The complexity increases even more when examining MTS data, where the centroid is also a temporal pattern in the dimensions of the whole dataset. For extracting MTS centroids, dynamic time warping barycenter averaging (DBA) was introduced by [148]. Thus, capturing all this information does not seem ideal for extracting any insights due to the inherent high dimensionality of MTS.

To address a dataset's high dimensionality, a common approach is to reduce the number of features in a way to better visualize the clusters on a two- or three-dimensional plot. This data transformation is typically achieved using Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) projections [175]. However, while these methods are effective for visualization, these methods transform data into a new feature space that does not preserve the interpretability of the original set of features. Moreover, it remains challenging how these could be directly applied to MTS datasets because different properties need to be considered, such as the time dependency of each point. In [176], different approaches are discussed regarding visualizing MTS by projecting them to lower dimensions while capturing time-related properties.

6.2.2. EXPLANATIONS ON TS CLUSTERING

Beyond statistically analyzing and exploring the data structures of each cluster, most research works focus on extracting the important factors (e.g., important variables) that influence cluster assignments, often by applying interpretable models. Despite the low number of publication outputs, the methodologies employed for explaining time-series clustering vary considerably.

Following the approach of transforming the data to different time-series representations, the work of [177] applies clustering on a range of interpretable extracted features, including intra-signal (or within feature) and inter-signal (across different features) characteristics, such as variance and correlation, respectively. After data preprocessing and dimensionality reduction techniques, a selection of features is retained for clustering, showing the high importance of inter-signal features as

always preserved in clustering. The number of features is then used as a measure of interpretability, meaning that fewer features facilitate clusters' comprehension.

Moreover, other work, such as [178, 179] was inspired by the general trend of training and explaining classification models to predict the derived cluster labels. On the one hand, according to [178], local interpretability methods, such as Local Interpretable Model-Agnostic Explanations (LIME) [180], SHapley Additive exPlanations (SHAP) [181] as well as Gradient-weighted Class Activation Mapping [182], are used with different classification models (e.g. XGBoost) and trained again on statistically-extracted temporal features. Examples of such features were auto-correlation and median difference. On the other hand, the work of [179] proposed a different approach to holistically perform clustering and provide explanations through training a decision tree. In practice, to achieve this, two different data sources were used, time-series and static or baseline data (such as demographic data), for clustering and interpretation, respectively. After applying clustering on TS, the cluster labels accompanied by the associated static data were used to train an interpretable decision tree model. Therefore, it aims to optimize both objectives, clustering and interpretation, at the same time.

According to the studies above, it becomes evident that existing methodologies for providing explanations rely on the applied transformations of the original data. This means that explanations also refer to the transformed feature space rather than the raw features raising some questions regarding the actual interpretability and transparency of the explanations. Therefore, beyond the limited work on the topic, there is a need for further exploration of the level of the original data.

6.3. REVIEW ON CHALLENGES OF EXPLAINING MTS DATA

Due to the inherent challenges of the topic and the limited available published research, the current chapter work was also inspired by the literature focusing on using interpretable classification models for explaining a target output, which in our case corresponds to the clustering-derived labels. Thus, an additional review of the challenges of applying interpretable classification models on MTS is necessary.

6.3.1. CLUSTERING EXPLANATIONS

Although it is acknowledged that the interpretability of clustering is of great importance in uncovering meaningful insights about data structures, limited work has been conducted in parallel with developing clustering algorithms. All the well-known clustering methods, such as k-means, were designed to group data, mainly considering various objective functions and dealing with different data types, but without taking into account any interpretability aspects of clustering. Therefore, when clustering results need to be explained, post-processing steps are com-

monly used, involving an additional classification model trying to predict and explain the clustering labels. For instance, in [183, 184], interpretable threshold decision trees having k leaf nodes are applied to explain the labels of k-means or k-medoids. Therefore, in most cases, the problem of clustering interpretability can be formulated as classification interpretability. This is a quite more studied issue, with a goal to explain the cluster labels as classification outputs.

Regarding classification explainability, beyond using inherently interpretable models (such as linear or decision tree models), post-hoc explainability methods are also commonly explored. These post-hoc methods fall into two main categories, model-specific and model-agnostic methods. Model-specific techniques correspond to particular groups of models, such as extracting feature importance from tree-based models and tree ensembles or utilizing layer-specific integrated gradients for deep-learning models [185]. On the other hand, model-agnostic explanation methods are applicable on top of various models regardless of their architecture. More specifically, these include the widely applied LIME focusing on generating local and instance-specific explanations, as developed by [180], and SHAP for feature-based explanations developed by [181].

However, when dealing with MTS clustering, challenges arise in two aspects of such formulation. First, the aforementioned XAI techniques of supervised learning models typically focus on images, text, and tabular data, limiting their application to time-series data. Second, it is not straightforward how to input time-series data into the classical classification models, without discarding their temporal nature. Therefore, adjustments should be made in both parts to better explain the dynamics of MTS data.

ADAPTING XAI METHODS TO MTS

As already discussed, the application of the aforementioned post-hoc XAI techniques is limited when time-series data is involved. Especially, in MTS, finding meaningful constructs in high-dimensional information is not trivial. To deal with such data, segmentation techniques are frequently applied to split time-series data into smaller, more manageable subsequences. By analyzing segments of the time series, rather than the entire series at once, it becomes easier to apply XAI methods.

To extend these methods for MTS, specific adaptations of LIME and SHAP have been developed. For instance, TS-MULE [186] is a LIME-based method that incorporates multiple advanced segmentation algorithms, such as matrix profile [187] and Symbolic Aggregate approXimation (SAX) [188]. This way, it is more likely that meaningful subsequences are uncovered, leading to motifs (reoccurring patterns) and local trends that are easily interpretable. Such methods can be also adapted to interpret forecasting output, instead of only dealing with the classical classification output. In the case of SHAP, multiple extensions

have been proposed and adjusted to a time-series setting. An example is an extension of KernelSHAP, adapted to explain time-series models, such as AR, ARIMA, VAR, VARMAX [189]. This method focuses on computing feature importance values for time series data. Another is the TimeSHAP method, aiming to explain more complex RNN-based models [190]. TimeSHAP provides explanations on multiple levels, by computing feature-, timestep-, and cell-level attributions.

ADAPTING MTS AS AN INPUT TO CLASSIFICATION MODELS

Another challenge of an MTS classification task arises from the complexity of time-series, which is not straightforwardly input to a classical machine learning model. The problem is the complex nature of time-series that makes it deviate from the conventional feature-vector representation. In the context of MTS, data is defined in a multi-dimensional feature space and characterized by special connections between the instances (time-points) as well as features. Thus, potential approaches focus on how to input the time-series data into the classical classification models.

The most straightforward way is to neglect any temporal dependencies by assuming instance and feature independence. This discards any time-oriented association and perceives the data directly as a vector input. Likewise, any classification model can be used on this dataset, but using the same output for all instances of the same time-series. Moreover, a well-applied approach addressing such data is to transform the complex-structured time-series data to a simple feature-vector representation. Such transformations can be achieved by using statistical-based or shapelet- and subsequence-based characteristics of the time-series [177]. After such transformation, a feature-vector representation is retained, which can be easily used in all existing classification models.

Alternatively, the necessary transformations can be achieved internally through models incorporating data representations and prediction. Recently, neural networks capable of handling multivariate time-series data have been increasingly used. More specifically, recurrent neural networks (RNN) models, such as long short-term memory (LSTM) and gated recurrent unit (GRU) represent the state-of-the-art group of models in tasks involving sequential decision-making. Besides these, attention mechanisms have also been introduced to sequential modeling with the ability to uncover and highlight the most important parts or periods of sequences [171, 191]. In other words, attention-based models offer interpretability for the results by using the learned attention weights. Utilizing attention weights is considered a form of inherent interpretability that is not commonly observed in NNs.

Given the attributes of attention-based models, the two main challenges of the current problem, handling MTS data and MTS interpretability, seem to be overcome. Thus, our methodology is specifically designed to enhance the explainability of MTS clustering by employing an attention-based mechanism, strategically integrating both temporal and

feature-level attention.

6.4. FRAMEWORK FOR CLUSTERING EXPLANATIONS

This section focuses on the proposed architecture that utilizes an interpretable attention-based framework and how this can eventually lead to clustering explanations. An overview of the proposed framework for providing explanations on EMA clustering is given in Figure 6.2. All components of the framework are described as follows.

6.4.1. INPUT: EMA DATA

The framework begins with the EMA data matrix X as input. In this chapter, the real-world NSMD dataset (previously described in Section 2.6) is examined. Due to the variability observed in missing observations of individuals, we implement a padding strategy to complete the examined EMA dataset so the dimension of X is $\{187, 12, T\}$. Thus, each individual's data is processed and filled when necessary, so that all have the same number of time-points, equal to the maximum number of time-points T observed across all individuals, for further analysis.

6.4.2. OUTPUT: CLUSTERING LABELS

In this setting, we adopt a framework that focuses on explaining clustering outcomes in EMA data without depending on a particular clustering method. That is, our framework relies on the fact that clustering is performed as a prior step, utilizing only the derived clustering labels to further provide explanations. An obvious advantage of this is the flexibility in the choice of clustering technique. Thus, as clustering is not an integrated component of the framework, it basically utilizes the clustering labels as the output to understand the reasoning behind the formation of each cluster.

6.4.3. INTERPRETABLE MODELS

Moving to the actual components of the proposed framework, in the case of k-clustering, it consists of k interpretable classification models. Each model is designed to predict a specific cluster while distinguishing it from all other clusters. For example, regarding the interpretable Model 1, where the goal is to predict Cluster0, the output labels are 1 for individuals of Cluster0 and 0 for Cluster1 and Cluster2. By predicting a single cluster through an interpretable model, we could have access to the description of the model highlighting the special characteristics, dynamics, and features that differentiate that cluster from the rest. Then, fitting k models could facilitate explaining all k derived clusters, which means exploring the important variables and time-points for each cluster.

Practically, although the input of each model is the whole EMA dataset, the difference of the k models lies in their respective outputs. This works

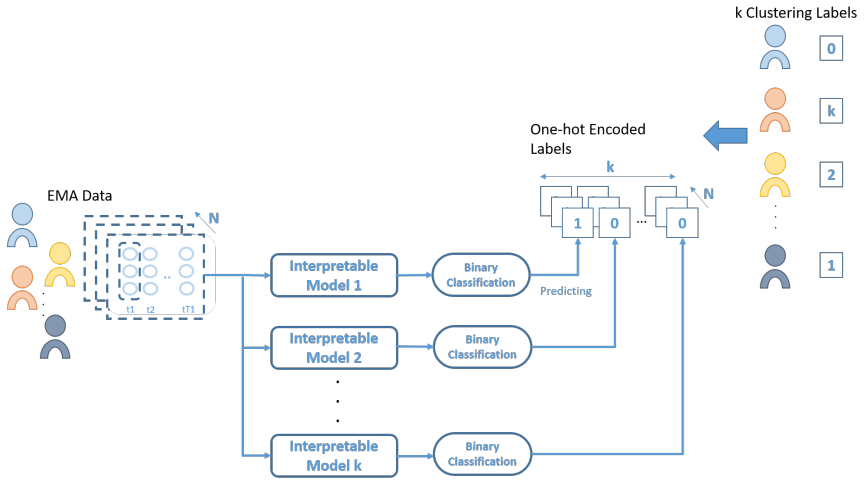


Figure 6.2: An overview of the proposed framework for providing explanations on EMA clustering. The data of N individuals is input into k interpretable models, each predicting an individual's cluster membership. To predict and explain the membership of individuals in the k derived clusters, the clustering-derived labels are binarized by one-hot encoding.

by one-hot encoding the clustering labels of the pre-applied clustering and using a different output vector for each model, as shown in Figure 6.2. This encoding method transforms the categorical cluster labels into a binary matrix, ultimately forming k binary classification models. It should be noted that in the case of a 2-clustering, the framework consists then of only one interpretable model. As the clustering output is already binary, there is no point in fitting two models.

6.4.4. Cluster-specific binary classification model

As already discussed, each binary classification model aims at predicting the individuals of a single cluster over the rest of the clusters. The main components of each model are shown in Figure 6.3. Each binary classification model is interpretable mainly relying on self-attention mechanisms [171]. By nature, the attention mechanism focuses on the temporal domain of the data, identifying the most important parts of the data in relation to the prediction task. By keeping only the relevant parts of the data and minimizing or filtering out the effect of irrelevant ones, the input space is effectively sparsified. This is implemented by setting different attention weights on each time-points depending on the contribution to the output. Therefore, such weights could be beneficial in recognizing the important parts of the data for classifying an individual in one cluster [191].

In the current framework, to address the complexity of EMA data, 2-

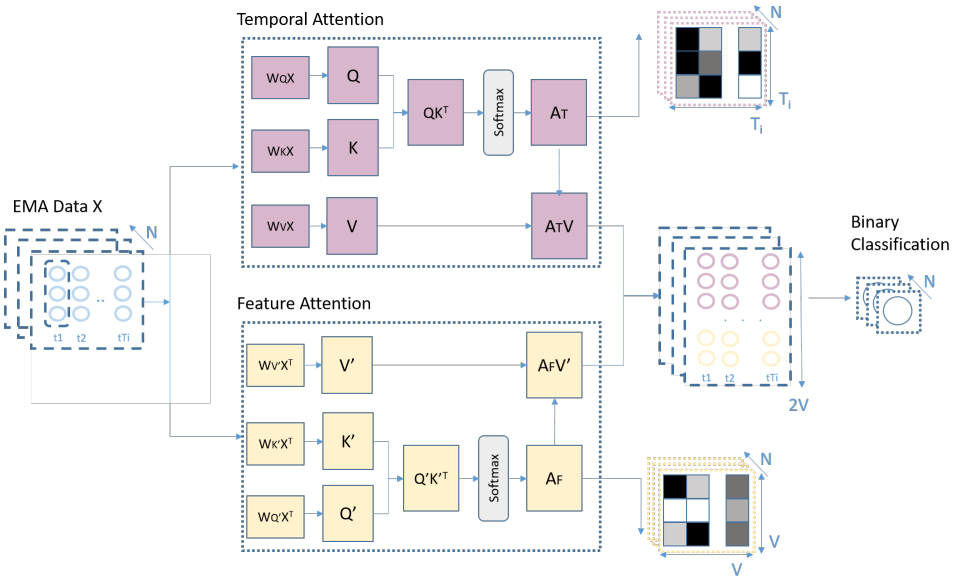


Figure 6.3: An overview of the main components of each interpretable model, consisting of the temporal and feature-level attention. The outputs from both attention levels are concatenated to perform binary classification, predicting cluster membership. Here, V refers to the value matrix of attention, distinct from the variable notation elsewhere in the dissertation.

level attention is used in parallel, each focusing on different aspects of the data [172, 173]. The first level of attention is dedicated to uncover the important parts in the temporal dimension, while the second one is to the important features. The value of analyzing data at the feature level is particularly evident when it comes to interpretation, as it is inherently more insightful to offer explanations based on specific features.

The description of each attention-based mechanism is as follows:

TEMPORAL ATTENTION

As shown in Figure 6.3, the calculations for Temporal Attention follow the procedure of a Scaled Dot-Product Attention self-attention block [171]. First, the 3 main components of the attention block, matrices query Q, key K, and value[1] V, are calculated by a linear transformation of the input X data. Then, the dot product of Q and K is the attention matrix, which after a softmax normalization produces the actual matrix of temporal attention A_T. For each individual i, the dimensions of A_T is $T_i \times T_i$. Subsequently, the attention matrix is multiplied by the matrix V (linearly

[1]The notation V here refers specifically to the standard attention mechanism terminology and it is different than V used to denote variables elsewhere in this dissertation.

transformed input X) to yield the context vector. In other words, the context vector is the weighted input based on the learned attention weights.

By learning the weights in A_T, the temporal attention component is designed to identify the most significant time-points within the EMA data. These weights are different for each individual and each time-point, resulting in a $T_i \times T_i$ matrix. This matrix captures the relative importance of each time point with respect to every other time-point, offering rich insights into temporal dynamics. However, its size and complexity pose challenges for direct interpretation and visualization. Despite the detailed and informative structure of the derived A_T, we employed a strategy of averaging over one dimension, while ensuring the other dimension is normalized. This approach averages the contributions across all time-points for each time-point, yielding only the average effect over time. This significantly simplifies the attention matrix, retaining only one dimension of size T_i or a time-series of T_i time-points.

FEATURE-LEVEL ATTENTION

In parallel to temporal attention, the feature-level attention mechanism assesses the importance of each feature within the EMA data. As observed in Figure 6.3, the block of feature-level attention is the same as of the temporal attention. The only difference is that the initial Q', K' and V' matrices derive from a linear transformation of the transposed EMA data, X^T. As before, this component results in the attention weight matrix A_F and the context vector. While the latter has the same size as the input X, the feature-level attention is in the dimension of $V \times V$. This is designed to map the inter-feature relationships and their contributions to the model's predictions. Through such weighting, the framework can distinguish which variables play an important role in influencing the outcome, providing an additional layer of explanation that complements the temporal insights.

Regarding interpretability, it is evident that being in the feature domain makes it more straightforward to provide a deeper understanding of the role of influence for each variable. Additionally, the low feature dimensionality allows us to directly represent, visualize, and analyze the feature-level attention weights. This offers clear insights into which features are the most relevant and how these features influence each other within the context of the prediction task.

6.4.5. CLUSTERING EXPLANATIONS THROUGH ATTENTION WEIGHTS ANALYSIS

To provide clustering explanations, the produced averaged temporal attention weights and feature-level attention weights are further analyzed. The weights analysis first facilitates the description of each model, aiming to differentiate each cluster, and then explaining the observed differences. In the following analysis, we adopt a multi-aspect approach to present our findings, covering a thorough exploration of the impact that

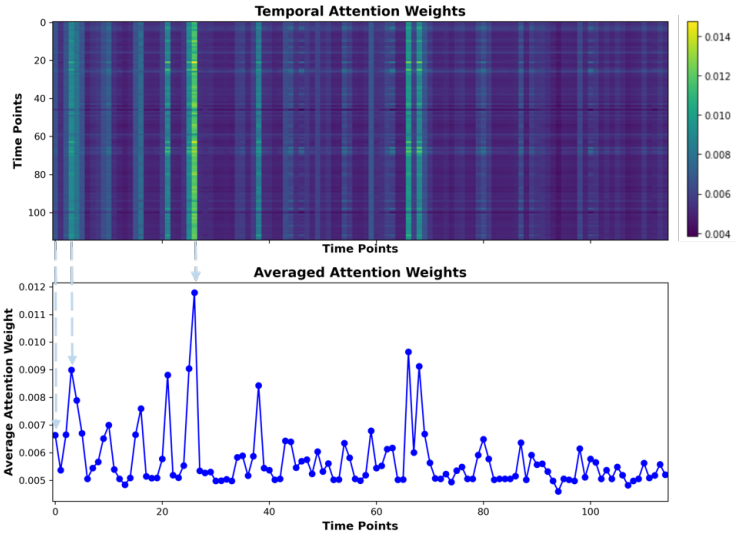


Figure 6.4: The averaging process of the full Temporal Attention matrix A_T to $A_{T_{av}}$.

attention mechanisms have at various but interconnected levels. These involve the cluster- and individual-level. From high- to low-level, each level zooms into different parts of the data, offering unique insights into the underlying relationships.

CLUSTER-LEVEL ANALYSIS

At the cluster level, the focus is on separately describing the individuals belonging to each cluster to get a clearer picture of the group-specific behaviors and attributes that distinguish one cluster from another. By examining the averages of attention weights across all individuals of each cluster, derived from the model describing each cluster, the aggregated behaviors and patterns of a cluster can be identified. At a high level, the average weights could uncover the special characteristics or strong effects that all these people have in common and possibly drive them to belong to the same cluster.

Although both temporal and feature-level attention weights are explored, getting access to the average effects across individuals based on the full temporal attention A_T, or even the averaged temporal attention $A_{T_{av}}$, is a bit challenging. As already discussed and shown in Figure 6.4, each individual is described by a time-series, $A_{T_{av}}$, showing their important time-points. However, it is not meaningful to average over different time-series, because the important time-points differ across individuals. To address this, the correlation between the attention weights $A_{T_{av}}$ and each feature's time-series is calculated, potentially identifying which features consistently align with the attention temporal trend. Eventually, at

this level of analysis, through both the temporal and feature-level attention, the importance of specific variables in leading to a particular cluster output could be uncovered.

Subsequently, at a cluster level, the patterns of inter-variable relationships or interactions dominant for each cluster are investigated. Based on the model's description through the attention weights, the relationship between data across individuals of one cluster can also be checked with respect to the acquired weights.

INDIVIDUAL-LEVEL ANALYSIS

At the individual level, the attention weights are examined separately for each individual, allowing for a personalized interpretation of the data. Without transforming or averaging the attention weights, the learned weights are analyzed along with the original feature space of each individual. Therefore, the focus is on understanding why or what was important to drive each individual to belong to a particular cluster. More specifically, regarding the time domain, the temporal attention weights facilitate uncovering what is happening underlying the time-points that are important for the prediction output. For instance, it is interesting to show which combinations of feature values get higher attention and which get lower.

6.5. ANALYSIS AND RESULTS

In our analysis, the real-world NSMD dataset is used, consisting of 187 individuals, 153 padded training time-points (70% of the total 224 time-points), and 12 distinct variables. After comparing different clustering methods based on different intrinsic evaluation measures in Chapter 4 (Section 4.5.3), a 3-clustering result derived from a GAK kernel k-means was chosen as the optimal clustering. To gain deeper insights into the identified clusters, we started by visually analyzing some of their characteristics. The following figures, Figure 6.5 and 6.6, illustrate the distribution of the average feature values across the clusters and the number of individuals in each cluster, respectively, providing a clearer picture of the clustering structure.

Then, the clustering labels derived from that particular method are further investigated to provide explanations regarding the formation of the clusters. Thus, the goal is to uncover what is different among clusters, meaning the important characteristics that drive each individual to belong to that specific cluster.

Taking as inputs the dataset X and the cluster labels, the proposed framework for describing and explaining clustering can be employed. According to the structure of the framework, for a 3-cluster grouping, a set of 3 interpretable models is trained on all individual EMA data, each aiming to predict one cluster over the rest. Therefore, the labels are one-hot encoded and each one-hot vector is the output of one model.

Figure 6.5: Distribution of the average feature values across individuals in all 3 clusters. Each point represents the average value of a specific feature for an individual, grouped by clusters.

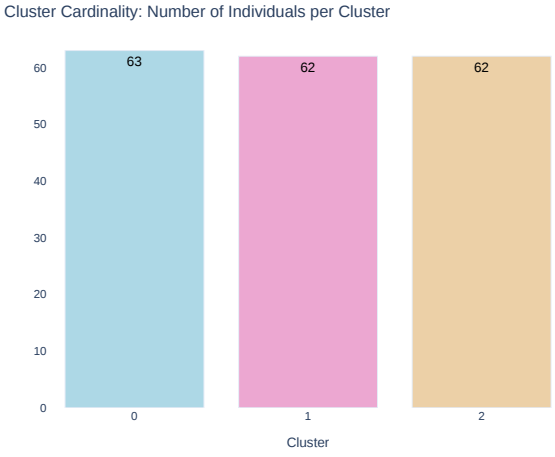


Figure 6.6: Cluster cardinality showing the number of individuals per cluster.

6.5.1. PERFORMANCE EVALUATION

Since the following analysis relies on the parameters derived from the 3 models, it is important to evaluate their effectiveness. The performance is then compared against the individual component of temporal attention and the baseline LSTM model. The assessment focuses on the ability of each approach to predict the cluster labels (provided by the predefined clustering) based on the EMA MTS data. After splitting the data into training (first 70% of individual time-points) and test (last 30%) sets, the accuracy, that is the number of correctly classified individuals (out of 187) is calculated. Each model was trained on the same training and test datasets, ensuring consistency in evaluation. The averaged results over the 3 models of each approach are presented in Table 6.1.

This comparison shows that the proposed framework performs at least as well as both the baseline LSTM model and the individual attention-based component. By effectively integrating the temporal and feature-level attention mechanisms, all individuals are identified in the correct cluster in the training set. On the test set, the framework correctly classified 105 out of 187 individuals, a slight improvement over the other models, demonstrating a modest but meaningful gain in generalization. Therefore, it is expected that the weights of a better-performing model could more accurately reflect the description of the underlying prediction task, which is the clustering. The rest of the analysis is conducted on the training set of all individual data.

6.5.2. CLUSTER-LEVEL EXPLANATIONS THROUGH TEMPORAL ATTENTION

To describe the characteristics of a cluster, the focus is on the model predicting that cluster label over the rest of the clusters. For example, for the first cluster, Cluster0, the analysis is conducted on the parameters of the first interpretable model, Model0. Regarding this model, the individuals of the two classes, which means belonging to one cluster over the rest of the clusters, are separately analyzed. Thus, by showing the average effects of the parameters of the individuals belonging to Cluster0, we could provide some description of Cluster0. A similar procedure holds when describing all different clusters. In other words, the parameters of Model1 are used for describing Cluster1 and the parameters of Model2 for Cluster2. Then, across all 3 models, the individuals belonging to the associated cluster, are separately analyzed.

As already discussed, to give an overview of the temporal effects on cluster level, a correlation analysis is employed between the temporal attention weights $A_{T_{av}}$ and the time-series of each feature. The average correlation scores across all individuals of each cluster are presented in Figure 6.7.

According to Figure 6.7, the important effects of features in distinguishing each cluster from the rest are identified. It can easily be seen that the patterns of the correlations of EMA features and temporal attention

Table 6.1: Comparison of models performance across 187 individuals, summarizing the training and test accuracy for three models.

Model	Training Accuracy	Test Accuracy
Baseline LSTM	144/187	100/187
Temporal-Attention	159/187	100/187
Proposed Framework	**187/187**	**105/187**

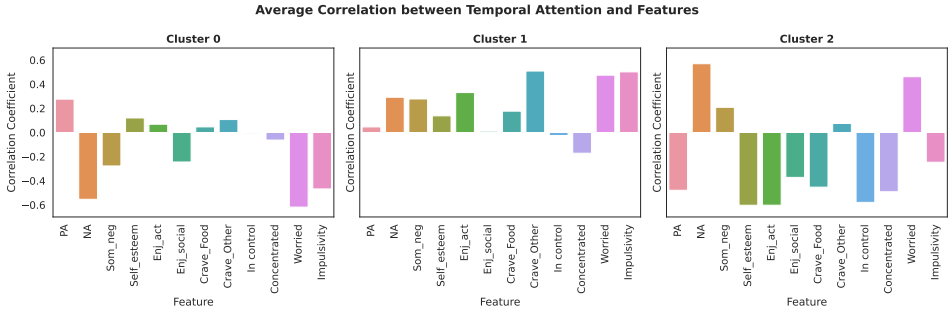


Figure 6.7: Cluster-level average correlation effects between the temporal attention weights and the EMA features.

weights are mostly distinct for each of the clusters. Furthermore, particular high (low) correlations stand out. In Cluster0, "Negative Affect" (given as NA), "Worried" and "Impulsivity" have strong negative correlations, meaning that the high values of these features get a lower attention weight. In Cluster1, almost all derived correlations are positive, with the highest being for "Craving_Other", "Worried" and "Impulsivity", whereas, in Cluster2, almost all correlations are in a stronger range, apart from the feature "Craving_Other". Although a cluster description has been uncovered by the analysis above, it still relies on the average (across all individuals of a cluster) effects. In other words, some individuals' effects may deviate from the average ones. Thus, explanations at the individual level should also be further explored.

Despite the first findings on clusters' composition, the actual role of attention has still been unclear. For instance, we need to understand what it means to get a higher or lower attention score. Therefore, we could analyze the average attention weights of individuals belonging to each cluster (represented by Class1 in each model) or those not belonging to that cluster (represented by Class0). As before, across all models, the average attention weight of all 187 individuals is shown in Figure 6.8 with respect to their average values of one feature. In each sub-figure, all 187 individuals are shown, represented by a point and colored according to their output class in each model. Regarding the coloring, these are different on the first and second row of the figure: while on the first row,

the real class label (0 or 1) is depicted, on the second row the actual cluster label.

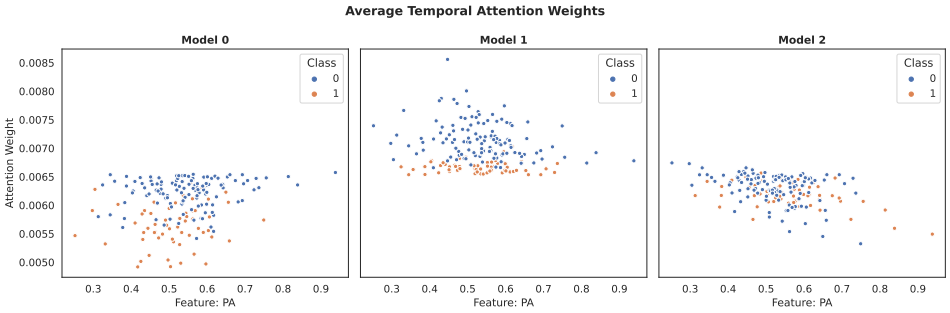
We notice that for Model0 and Model1, individuals belonging to Cluster0 and Cluster1, respectively, get lower (on average) attention scores than the rest. This could possibly reflect the fact that Class1 is always the minority class compared to Class0. Thus, each model gives more attention to the majority class. Nevertheless, the findings of Model3 are not similarly clear. It is noticeable that the attention weights of Cluster0 and Cluster2 are slightly mixed, but getting lower values than Cluster1. These unclear results should be further investigated as it may show that the third cluster may not be needed.

Additionally, no significant effect is observed on the feature level. We can see that individuals with average values ranging from 0.3 to 0.9 can belong to all possible clusters. This is also apparent when plotting across any other feature, where the same patterns are found.

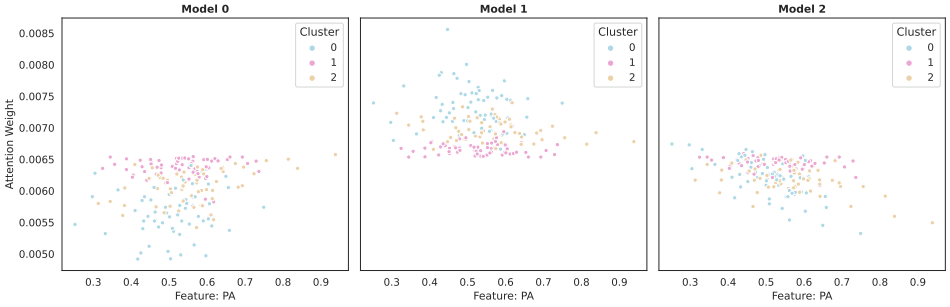
In response to the overlapping attention weights of Cluster0 and Cluster2, the similarities across clusters should be analyzed. Since our clustering relies on the GAK similarities, these are plotted for all individuals. In Figure 6.9, the similarities of all individuals (colored by their true cluster) are depicted in Cluster1 and Cluster2. Practically, for each individual, the average similarity to all the individuals of each cluster is given. Because of the limitations of a 2-dimensional plot, similarities to Cluster0 are not fully shown.

As expected, it is interesting to see that there is a similarity between Cluster0 and Cluster2. Although individuals of Cluster0 have very low similarity to Cluster1, lower than 0.05, most of them have a similarity between 0.10 and 0.30 to Cluster2. This range is not far from the within-cluster similarity which only reaches the level of 0.40. This comparable level of similarity between and within clusters suggests that some individuals might share characteristics across clusters, adding a layer of ambiguity in their cluster membership and highlighting shared features, especially between Cluster0 and Cluster2. Such findings raise questions regarding the quality and robustness of the chosen clustering. Therefore, it is indicated that the current clustering method may not fully capture the underlying heterogeneity of the dataset.

In the context of cluster-level analysis, beyond the initial feature-related effects, we could further elaborate on the analysis using feature interactions. For example, the interaction between two features, "Positive Affect" (PA) and "Negative Affect" (NA), along with the acquired attention weight is shown in Figure 6.10. For each cluster, each point represents a time-point of the individuals belonging to that cluster. According to this figure, different interaction patterns underlying each cluster can be identified. More specifically, for Cluster0, high NA values lead to low attention weights, whereas the opposite effect is seen for Cluster1 and Cluster2. Although for Cluster0 and Cluster1, no visible interactions were observed, a quite clear pattern is seen for Cluster2. The combination of low PA and high NA leads to high attention weights,



(a) Differentiation of attention weights across Class0 and Class1 of all 3 models.



(b) Differentiation of attention weights across the actual 3 clusters.

Figure 6.8: Relationship between "Positive Affect" and Temporal Attention weights.

whereas high PA and low NA lead to lower attention weights. By exploring all possible feature interactions, we have the opportunity to get more insights into the underlying structure of each cluster. Thus, the potential distinctions could point to differences in which variables are most relevant in characterizing each cluster.

6.5.3. CLUSTER-LEVEL EXPLANATIONS THROUGH FEATURE-LEVEL ATTENTION

Regarding the learned feature-level attention weights, the raw A_F weights are directly analyzed. Similarly to the previous section, to get some insights at a cluster-level, the average effects of all individuals within each cluster are aggregated. The average feature-level attention weights for each cluster are given in Figure 6.11. This figure of 12 × 12 heatmap presents the asymmetric inter-relations among features, specifically showing the relative contribution of Feature0 (x-axis) to Feature1 (y-axis).

The averaged attention weights can be similarly interpreted as the relative importance of various features in distinguishing one cluster from the rest. At first glance, it is noticeable that the same features emerge

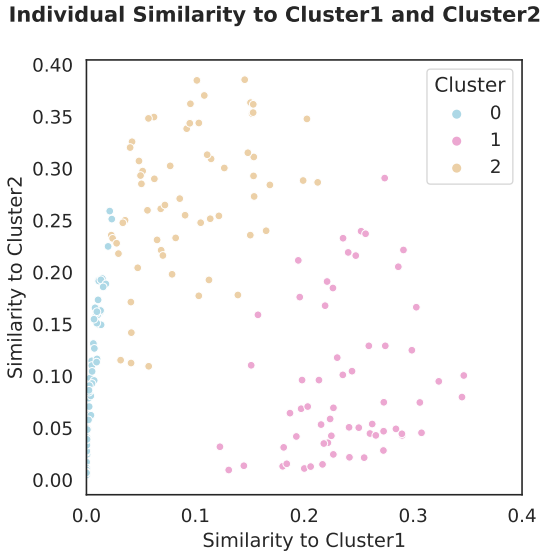


Figure 6.9: Similarities of all individuals to Cluster1 and Cluster2.

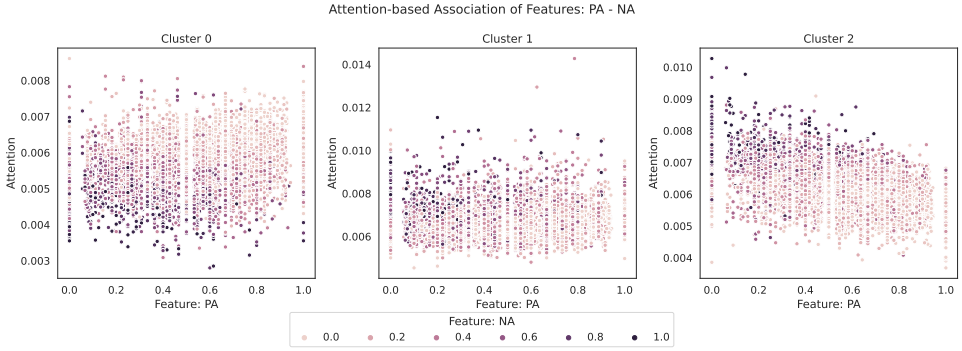


Figure 6.10: Features Interaction between "Positive Affect" (PA) and "Negative Affect" (NA) with respect to the Temporal Attention.

across all clusters, "Enjoying Social Activities" and "In Control", but to a different importance degree. For instance, this means that high values of "Enjoying Social Activities" are assigned to high attention weights in relation to any other variables. To illustrate this, an example of these interconnections is given in Figure 6.12. More specifically, the association of the average "Enjoying Social Activities" value with the average "Positive Affect" and "Crave Food" across all individuals within each cluster is depicted in the first and second rows, respectively. Each point represents an individual belonging to a particular cluster. As expected from

Figure 6.11, we can identify the high attention contribution of high values of "Enjoying Social Activities" to the other features along with other interesting patterns. For instance, in Cluster2, high attention weights are assigned to low values of "Crave Food" and high values of "Enjoying Social Activities". By analyzing the feature interplay across all combinations, from the perspective of feature-level attention, deeper insights into the underlying dynamics can be uncovered. Thus, a more detailed examination can highlight specific characteristics and patterns that are important for distinguishing each cluster.

6.5.4. INDIVIDUAL-LEVEL EXPLANATIONS

While cluster-level analysis offers valuable insights into the commonalities within clusters, undoubtedly it smooths over individual differences. To more thoroughly understand the important patterns of cluster formation, it is essential to examine the data at the individual level as well.

Beyond unfolding the individual effects of the cluster-level analysis conducted before, it is important to investigate the underlying interactions of features in response to the attention weights learned by the model dedicated to the cluster each individual belongs. The example of the first individual, Individual0, belonging to Cluster2, is used for the rest of the analysis. A detailed summary plot including all feature interactions and their attention weights is presented in Figure 6.13. Particularly, for each feature on the y-axis, all time-points of Individual0 are plotted, while ordered according to their attention weights, and colored by the corresponding feature value. Through this, we can have a detailed exploration of the combinations of feature values that lead to higher attention. According to the model, higher attention weights show the important characteristics for distinguishing individuals between clusters. Although it was indicated before that on average low attention was assigned to individuals classified as Class1, which corresponds to Cluster0 and Cluster1, this pattern was not that dominant for Cluster2. Thus, further investigation of all individuals is necessary for a deeper understanding of the distinct feature dynamics, especially in the case of Cluster2.

To more clearly understand the role of attention on each individual, the differences in the weights learned by all models should be additionally studied. Specifically, for each individual, feature interactions could be compared against the attention weights derived from each model. This could facilitate uncovering how much each feature interplay influences each model's decisions for each individual. An example of an individual belonging to Cluster2 is shown in Figure 6.14. By examining how the same feature interactions are colored based on the 3 models, some distinctions are apparent. For example, it is observed that the combination of high values in both "Positive Affect" and "Enjoying Social" leads to higher weights for Model0, whereas to lower ones for Model2. Thus, these patterns possibly reflect the impact that the specific time-points have on predicting Cluster0 and Cluster2, respectively. Also, after the

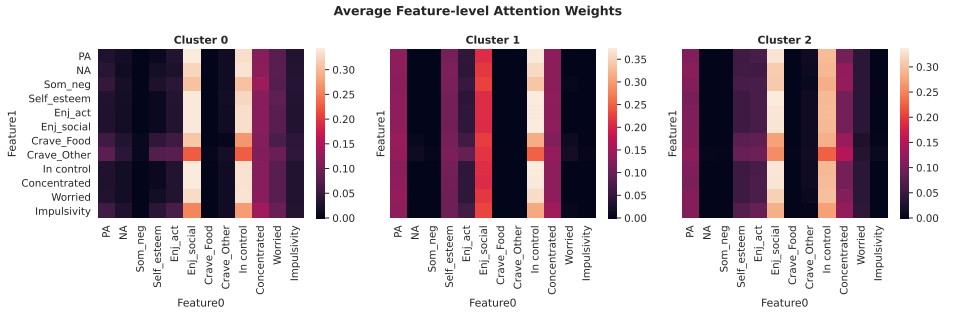


Figure 6.11: Cluster-level average feature-level attention weights.

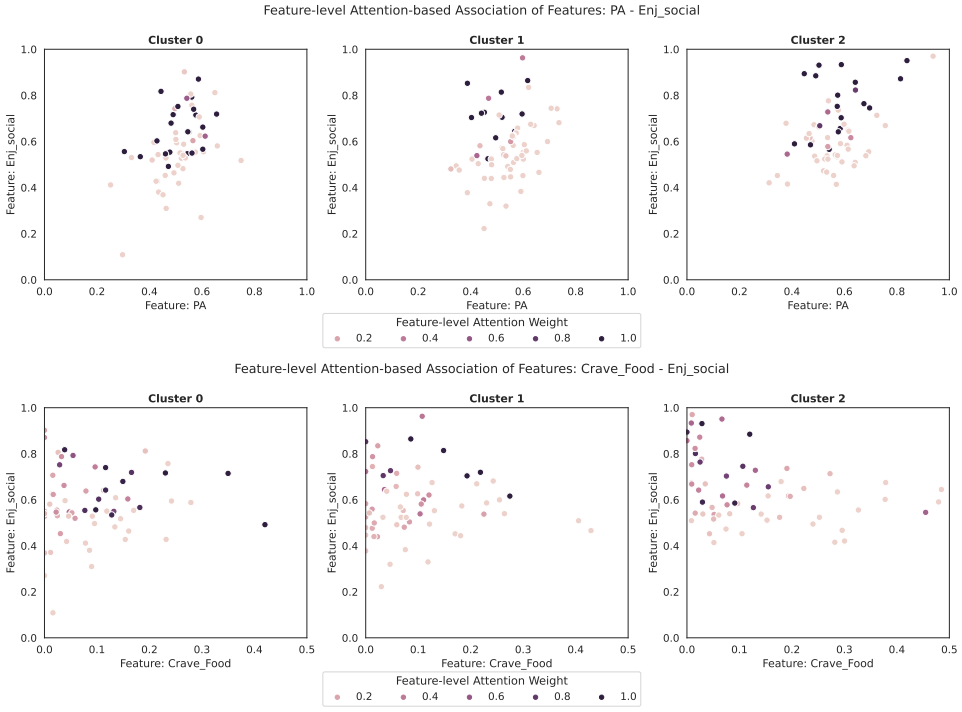


Figure 6.12: Feature-level Attention: Unfolding the inter-connection of "Enjoying Social Activities" to two other features: "Positive Affect" and "Crave Food".

previous indication that lower attention weights are linked to a particular Cluster, the lower attention of Model2 could highlight the most relevant information for predicting that individual as Cluster2. Therefore, such a comparison offers valuable insights into how different models prioritize and interpret the same set of features of an individual.

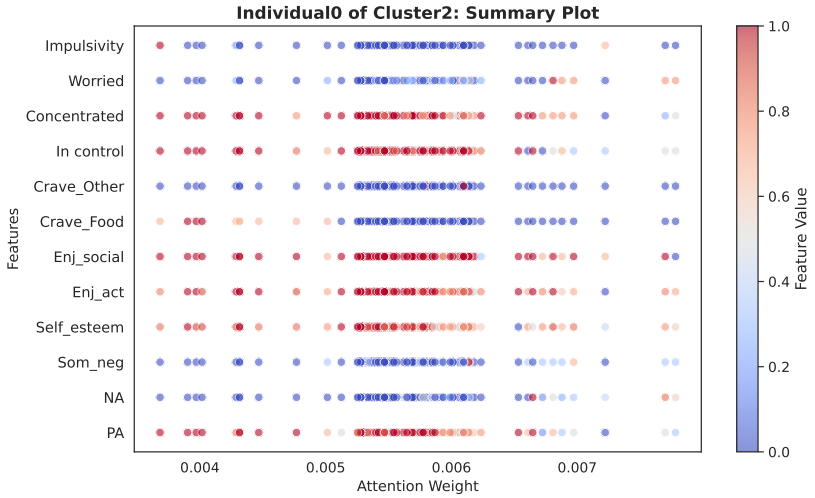


Figure 6.13: The summary plot of feature interactions for an individual of Cluster2.

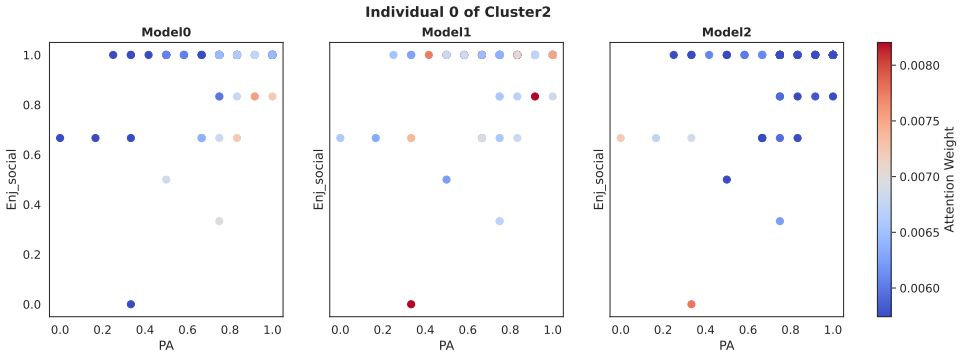


Figure 6.14: The attention weights of the interaction between "Positive Affect" and "Enjoying Social" derived from all 3 models.

6.6. DISCUSSION

In this chapter, the understudied problem of providing explanations on clustering results in the context of MTS data is explored. A novel interpretable framework is proposed and examined using a real-world EMA dataset. To address the complexity of EMA data, our framework offers interpretability by integrating 2 levels of attention mechanisms, in the temporal and feature-level dimensions. Through its multi-aspect attention design (Section 6.4.4) and analysis (Section 6.4.5), this framework eventually facilitates a deeper understanding of clustering, providing interpretations of the important underlying patterns. Next, we focus on the role of the multi-aspect framework's design and analysis and the impact

of prior applied clustering on interpretation and its validation.

6.6.1. THE ROLE OF THE MULTI-ASPECT ATTENTION

Although attention mechanisms have typically been employed in the temporal dimension, the current approach of integrating multi-aspect attention mechanisms, focusing on both temporal and feature levels, has shown that it enhances the performance of the overall accuracy of the downstream model. Based on Table 6.1, this integrated approach improved the performance of all baseline methodologies, by predicting more individuals in the correct clusters in training and test sets. While the increase in test accuracy is modest, it suggests that the learned model weights contribute to better distinguishing individuals across clusters.

Moving to interpretability, the multi-aspect attention mechanism offers detailed information on the data patterns each model considers important. While temporal attention highlights the significance of particular time-points, as shown in Figure 6.4, such information remains meaningful on the individual level since each individual exhibits behaviors at different time-points. To enrich the learned information and our understanding on cluster-level, feature-level attention can additionally identify the important interconnection among features. Thus, incorporating both temporal and feature-level attention enhances the interpretability of complex data, such as EMA, where the understanding of dynamic patterns is crucial.

6.6.2. THE ROLE OF THE MULTI-LEVEL ANALYSIS

The analysis of all learned attention weights is conducted at 2 levels, cluster- and individual-level. The cluster-specific insights provide a deep investigation of the commonalities as well as distinctions across clusters. After aggregating the effects of all individuals in each cluster, we derived the most influential time-points as well as the underlying features (Figure 6.7) and feature interactions (6.11) based on a model's decision-making. Although some first cluster descriptions were derived, it should be noted that the average effects smooth over the real individual differences. Therefore, to better understand each cluster formation, examining the data at the individual level is essential. This individual-level analysis facilitates uncovering important feature interactions and patterns unique to each individual (Figure 6.13), while also revealing how each individual is reflected on the weights learned by models focusing on different clusters (Figure 6.14).

6.6.3. THE IMPACT OF THE ALGORITHM-AGNOSTIC META-CLUSTERING FRAMEWORK

A key advantage of the framework is its algorithm-agnostic nature regarding the generation of clustering labels. By design, as clustering is

not integrated into the framework, it shows great flexibility allowing the use of every possible clustering technique as a preliminary step. Thus, it could be used as an effective tool for evaluating the result given by any clustering method. This capability not only enhances the framework's utility but also broadens its applicability, enabling the comparison and evaluation of clustering algorithms based on the quality and relevance of the insights they produce.

However, it is reasonable that the quality of the clustering results utilized in the framework plays an important role in the accuracy of the explanations. In other words, when clustering labels are not robust, the proposed framework would always provide some explanations on the cluster- and individual-level importance effects, but without these being meaningful or accurate. This scenario may also apply to the chosen clustering of our EMA dataset. Given that the true clusters are unknown and the between-cluster similarities are found close, the examined clustering may not represent the optimal solution. This can be caused by the clustering algorithm utilized or the complexity of the data. Nevertheless, even when using that clustering, the framework was capable of uncovering similarities in the influential data patterns of different clusters, suggesting that a 2-clustering result may be more possible. Therefore, regarding the evaluation of this framework, the clustering results of other clustering algorithms should be investigated. This necessity also comes since simulation studies on explanations are not typically performed because in principles explanations tend to be subjective. This issue of subjectivity is particularly observed in human-centric fields, like EMA studies, where data inherently includes significant subjective variation. Consequently, evaluating the quality of explanations through objective criteria becomes quite challenging. Therefore, it is more appropriate for explanations to be assessed by domain experts on a case-by-case basis.

Furthermore, explanations could be enriched by using various types of ground-truth information. This might include baseline data collected before or during the study, which offers valuable contextual insights into the examined individuals. For instance, demographic details or health-related information, such as high depressive symptomatology, could be used. Thus, such information provides crucial background that can help interpret patterns and variations within the data.

6.7. CONCLUSION

This chapter presents an interpretable framework for explaining and evaluating an MTS clustering result. By analyzing the attention-derived important time-points and feature interactions at both cluster and individual levels. This dual-level analysis not only uncovers the patterns and interactions that define each cluster, but also highlights unique individual contributions, offering a comprehensive understanding of the clustering process.

Such insights are particularly valuable in complex fields such as psychopathology, where a better understanding of clustered individuals could be beneficial for personalized interventions and mechanistic understanding. Therefore, this framework bridges the gap between data-driven clustering methods and their practical interpretability, ensuring that the results can be utilized effectively to address real-world challenges.

Having thoroughly examined the validity of clustering results in Chapters 4, 5 and 6, the next step is to take advantage of this information augmenting the personalized models. Specifically, Chapter 5 revealed that incorporating data from clustering-derived similar individual profiles could improve the personalized predictive performance. Building on these results, it is hypothesized that integrating clustering-based results in a more advanced way could further enhance the personalized approaches. This leads us to the exploration of transfer learning in the next chapter, Chapter 7. By employing transfer learning methods, the aim is to harness the rich information derived from clustering to strengthen nomothetic approaches while focusing on personalized information. This way, the idea is to take advantage of the insights gained from similar individuals while prioritizing the target individual each time, capturing both individual and group-level patterns for better predictive results.

7

TRANSFER LEARNING APPROACH FOR EMA MODELING

Considering the success of utilizing clustering-derived groups of similar individuals to build group-based models that enhance the performance of both personalized and using-all-data models, as investigated in Chapter 5, the next step is to take advantage of similar individuals in a more sophisticated modeling approach. In particular, transfer learning approaches can be applied to improve predictions for a specific individual (target domain) by incorporating data from other individuals (source domain). Among the existing transfer learning approaches, in this chapter, boosting-based methodologies, focusing on enhancing Transfer Adaptive Boosting (TrAdaBoost), are further explored. More specifically, the optimal selection of similar source domains and the development of effective weighting strategies, ensuring that the knowledge from relevant sources is utilized in a way beneficial for the target, is emphasized.

Parts of this chapter have been published in

- M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. "Enhanced Boosting-based Transfer Learning for Modeling Ecological Momentary Assessment Data". In: ML4ITS2023 - 3rd Workshop on Machine Learning for Irregular Time Series: Advances in Generative Models, Global Models and Self-Supervised Learning. ECML. 2024

7.1. INTRODUCTION

Starting from Chapter 3, one of the primary objectives in modeling EMA data is to develop accurate personalized models that could provide reliable descriptions about each individual. However, one of the main challenges in building personalized models is the limited number of data points available for each individual. Small datasets often lead to overfitted models without being capable of generalizing, or even to situations where models cannot be trained at all. Thus, information collected from other individuals in the same EMA study can be beneficial for modeling [45]. To address this, Chapter 3 utilized nomothetic models that pool data from multiple individuals, whereas Chapter 5 focused on clustering-derived group models. Although such models improved personalized performance by incorporating more data and capturing general patterns reflective across individuals, it is hypothesized that they might ignore important individual differences. Thus, there is a clear need for methodologies that can balance these approaches by integrating the benefits of both personalized and group-based approaches. Such methodologies can be derived from the concept of Transfer Learning.

Transfer learning is a Machine Learning (ML) paradigm where knowledge gained from one domain (the source domain) is additionally used to improve the performance in another domain (the target domain) [193, 194]. In the context of EMA data, transfer learning can be applied by incorporating data from multiple individuals (source domain) to enhance personalized predictions for a specific individual (target domain) [195]. While individuals may originate from the same data collection, they are not necessarily considered part of a statistically homogeneous population. Variations in data availability, patterns, and temporal dynamics across individuals almost always introduce challenges that resemble transfer learning scenarios [196]. Transfer learning can be particularly beneficial in cases where the target individual has limited data but shares similarities with individuals in the source domain. By transferring knowledge from a larger, similar dataset, it is possible to enhance EMA modeling for the target individual effectively. Therefore, the main objective of the chapter is to investigate whether transferring knowledge from other individuals could improve individual predictive performance using a real-world EMA dataset.

Among various existing transfer learning approaches, boosting-based methodologies are explored in the current chapter. In particular, a methodology adapted from transfer learning using Adaptive Boosting (AdaBoost) is investigated, taking advantage of both boosting and sample reweighting strategies [197]. Boosting algorithms aim to build an ensemble of predictive models, iteratively adjusting the weights of all data points (instances) depending on their misclassification rate and whether they belong to the target or source domain. Misclassification is determined by evaluating the model's predictions at each iteration: if the predicted label for a data point does not match the actual label, it is considered misclassified. For the misclassified instances in the

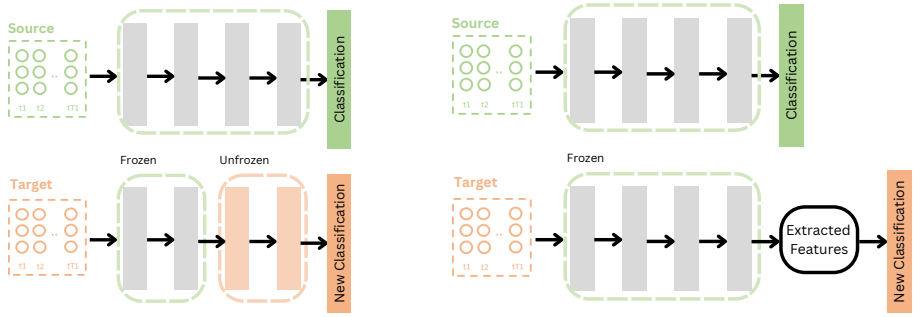
source domain, weight updating is based on the similarity to the target individual, whereas instances in the target domain should be more influential. Thus, the impact of different reweighting strategies as well as the number of similar source individuals are thoroughly investigated in a way to ultimately improve individual performance. Regarding the source data, different approaches are explored for selecting the optimal set of individuals, including similarity-based and clustering-derived results.

7.2. RELATED WORK

Transfer learning is an advanced ML paradigm that aims to improve the performance of a target domain or task by incorporating knowledge from a related source domain or task. Comprehensive reviews of transfer learning strategies and their applications can be found in [194, 198, 199]. Specifically, various advanced transfer-learning strategies have been proposed, focusing on sharing model-related information. The learnable information includes model parameters, feature-based transformations, and instance-based weights. Figure 7.1 provides an overview of the first two, illustrating their use in transfer learning for adapting and enhancing model performance for target tasks. Building mostly on large deep learning models, parameter-sharing strategies involve learning from models trained on source data and fine-tuning (adapting) the target dataset by freezing some deep learning layers and allowing others to learn during training [200]. Feature-based transfer learning utilizes learned transformations or frozen layers from the pre-trained model, using these as inputs for training on the target data to benefit from robust representations [201].

These techniques have seen a wide application in deep learning, especially in fields like computer vision (e.g., [202]) and natural language processing (e.g., [203]), where data across different datasets or domains are widely available. However, since in our case all data come from the same data collection, and the dataset is not that large, instance-based transfer learning is further explored [193].

Instance-based transfer learning is the most straightforward way of sharing additional information by providing models with more input data [204]. Importantly, this is possible within the concept of domain adaptation, where target and source data share the same feature space, coming from the same domain or EMA data collection [205]. While data pooling remains a valid initial strategy, it does not differentiate the influence between target and source data. This is not always optimal particularly when source data outnumbers the target data leading to a higher influence from the source [196, 206]. Such an issue is still under-explored since most existing instance-based transfer methods fail to adequately balance the contribution of target and source. Therefore, it is essential to study and apply balancing approaches where appropriate instances' reweighting could guide the optimal selection of instances in source individuals. Specifically, methods inspired by boosting, such as Trans-



(a) Sharing models' parameters. (b) Sharing feature transformations.

Figure 7.1: Learned knowledge for transfer learning.

fer AdaBoost (TrAdaBoost), can provide a more effective solution [197]. TrAdaBoost extends the concept of boosting, combining multiple weak learners (small and low-performed models) into a stronger one, by dynamically reweighting source and target instances across iterations. It reduces the weight of source instances that negatively impact the target task while increasing the emphasis on target instances and beneficial source data. This iterative process ensures that the resulting model not only utilizes valuable knowledge from the source domain to improve predictive performance on the target domain but also promotes distribution invariance across different domains [207]. By aligning the distributions within a shared feature space, the model ensures robustness and generalizability, even when there are differences in the underlying data distributions between source and target.

Nevertheless, according to previous research [73, 208], several limitations of TrAdaBoost have been identified that need to be taken into account and handled accordingly.

- Its learning process is highly prone to negative inference, which occurs when the knowledge transferred from the source domain obscures, rather than helps, the learning process in the target domain. This issue arises when the source domain data is not sufficiently similar to the target. In such cases, instead of enhancing the model's predictive capability, the transferred information may introduce noise and biases that adversely affect the learning process.

- The utilized reweighting strategy causes the weights of source instances to decrease progressively, eventually converging to zero. Based on the reweighting effects, misclassified instances in the source get lower and lower weights, minimizing their influence. Even the correctly classified source data, whose weights are expected to remain stable, get a gradual reduction because of the relative increase in the target. Therefore, the impact of the source is minimal.

- In the case of imbalanced datasets, where the distribution of class labels is not uniform, the prediction of the minority instances can become challenging. This happens because the model tends to be biased toward the majority class, as it is more represented during training. When minority instances are scarce in the target dataset, the model may fail to generalize without being able to predict underrepresented instances, potentially leading to poor performance in rare but important cases.

- TrAdaBoost can be computationally demanding, requiring multiple runs of training weak learners and reweighting instances. The effect is more challenging in the case of large datasets or real-time applications.

7.3. METHODOLOGY

This section starts by introducing how TrAdaBoost can be effectively applied to EMA data. Having also identified the issues inherent in TrAdaBoost, we illustrate these challenges and present the associated enhancements of this approach.

7.3.1. TRADABOOST ON EMA DATA

During an EMA study, data from multiple individuals are typically collected, all represented by the same set of variables. This is a key characteristic that makes EMA a promising application of transfer learning. In this setting, starting from one individual as the target domain, the goal of TrAdaBoost is to accurately predict the 1-lag future data points collected from that particular individual. In addition to the target data, the model gets input data from other available individuals, referred to as source domain data. Incorporating data from other individuals has the potential to enhance the model's ability to generalize and improve predictions for individuals with insufficient data, in terms of both size and quality. However, as already discussed, selecting the appropriate source data is crucial, as it can significantly affect individual performance. Therefore, it is crucial to explore different options for determining the optimal number of sources and assessing their relevance to the target. Regarding relevance, different methods for identifying the most similar individuals of the source are investigated. Given the nature of time-series data, temporally-oriented Dynamic Time Warping (DTW) distance or Global Alignment Kernel (GAK) similarity can be applied [85]. According to these measures, various sets of source data can be explored by keeping the most similar one each time.

Alternatively, more advanced approaches using clustering approaches can be utilized to discover the most similar individuals based on the derived clusters. By grouping individuals into clusters, the number of sources used can differ among individuals, depending on the number of

individuals in each cluster. For example, if clusters are not balanced, some individuals are trained using a higher number of source data than others. Such information can be used in the subsequent modeling steps.

7.3.2. MODELING PROCESS

During the TrAdaBoost modeling process, a number of boosting iterations take place, where several important steps are involved. These mainly include training a weak learner (or classifier for a classification task), calculating the training error, reweighting all the instances in the target and source domain, and normalizing the weights. The whole TrAdaBoost modeling process is depicted in Figure 7.2. Finally, after all iterations, the predictions of all weak learners need to be combined using an aggregation strategy. More specifically, all the proposed enhancements are described below.

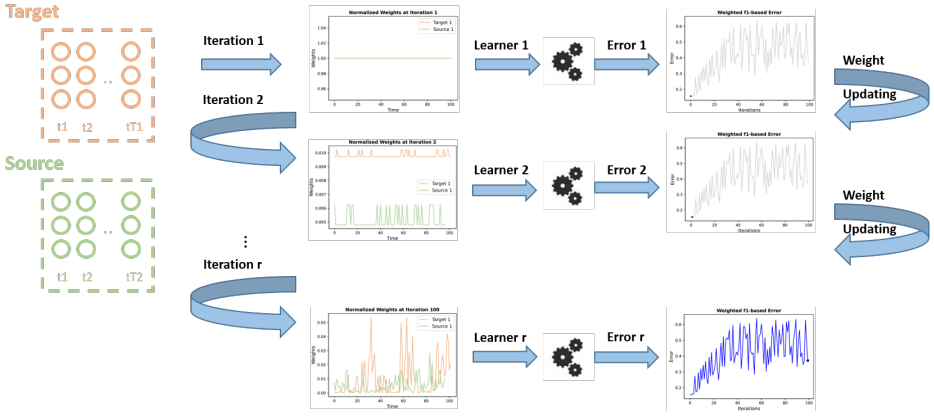


Figure 7.2: The iterative TrAdaBoost modeling process over r iterations for one individual as target and one individual as source. Each iteration involves: normalizing the weights, training a weak learner, calculating the training error, and reweighting all the instances in the target and source domain.

CHANGING THE ERROR METRIC: WEIGHTED F1 SCORE

According to all boosting algorithms, the training error at each iteration plays a significant role in the learning process, as it is involved in all weight updating and predictions' aggregation strategies. All boosting algorithms typically use the weighted average of the absolute error during training between the true label y and the predicted label y'. However, the absolute error may not be optimal in scenarios where the data is imbalanced. To address this and take into account the error for both minority and majority classes, a modified training error metric is proposed based on the F1 score. More specifically, using the weighted average of

an error based on the F1 score (defined as $1 - F_1$) provides a more representative measure reflecting the misclassification rate of both majority and minority classes.

As observed in Figure 7.2, the errors typically vary a lot across iterations in an attempt to correct predictions for the most challenging data points. It should be noted that an overall decrease is not necessarily expected because the goal is that different weak learners focus on different parts of the data.

TARGET AND SOURCE REWEIGHTING STRATEGY

Subsequently, the derived training error is used to update the weights for each instance in the target domain ($target_i$). According to the updating Equation 7.1, the error, $error_r$, is mainly utilized on the changing/learning rate, represented by the parameter β_r for each iteration (or round) r.

$$w_{r+1}^{target_i} = w_r^{target_i} \cdot \beta_r^{-|y_{target_i} - y'_{target_i,r}|}, \text{ where } \beta_r = \frac{error_r}{1 - error_r} \tag{7.1}$$

To retain the originally designed updating effects, which aims to increase the misclassified instances as expressed by $\beta_r^{-|y-y'|}$), the basis of the exponential β_r is constrained between 0 and 1. Subsequently, the weighted error $error_r$ should be lower than 0.5. Because of the exponential expression, this threshold is crucial for the validity of the whole process, so any $error_r$ greater than 0.5 is forced to be 0.5. This adjustment indicates that the original procedure may not have been entirely fair, as iterations with $0.5 < error_r < 1$ were effectively disregarded. To make use of all calculated errors and simplify the process, our approach adapts the learning rate to be based on the weighted F1 score $1 - F_1^r$. The update of weights (from w_{r+1} to w_r) for a target instance ($target_i$) is shown in Equation 7.2.

$$w_{r+1}^{target_i} = w_r^{target_i} \cdot (1 - F_1^r)^{-|y_{target_i} - y'_{target_i,r}|} \tag{7.2}$$

$$w_{r+1}^{source_i} = w_r^{source_i} \cdot \beta_0^{|y_{source_i} - y'_{source_i,r}|} \tag{7.3}$$

Moving to the core of the boosting concept, the original weight adjustment ensures that the model focuses on the harder instances in the next training iterations. Initially, all instances have equal weights. As training progresses, instances that are misclassified receive higher weights, increasing their importance in the training process (already seen in Figure 7.2). Although it may take several iterations until the harder target instances are correctly classified, these are supposed to be useful since they come from the target domain or individual of interest. However, this concept is not always applicable to the source. The misclassified source instances generally indicate that they may not be valuable for the target

and thus should get decreased weights expressed by $\beta_0^{|y-y'|}$ (with a fixed β_0 rate according to Equation 7.3) [197]. Although this is sometimes plausible, some source instances could have been misclassified because they are hard or challenging but valuable instances. In such cases, the algorithm may need several iterations for these being correctly classified. Thus, in our approach, we explore updating strategies that initially increase and then decrease the weights of the misclassified source data. The number of previous consecutive iterations (steps) where an instance can be wrongly classified but still has its weight increased, is a chosen hyperparameter *step*. If this specified number of steps is exceeded and it is not correctly classified yet, its weights start decreasing, as it should according to the original source updating strategy. The proposed source updating equations are presented in Equation 7.4.

$$w_{r+1}^{source_i} = w_r^{source_i} \cdot \beta_0^{-s \cdot |y_{source_i} - y'_{source_i,r}|}, \text{ if } \prod_r^{r-step} |y_{source_i} - y'_{source_i,r}| = 0$$
$$w_{r+1}^{source_i} = w_r^{source_i} \cdot \beta_0^{s \cdot |y_{source_i} - y'_{source_i,r}|}, \text{ if } \prod_r^{t-step} |y_{source_i} - y'_{source_i,r}| = 1 \tag{7.4}$$

As introduced in Equation 7.4, the source weights are updated not only based on the classification errors but also considering the relevance and similarity (s) to the target domain. If a misclassified source is highly similar to the target, there is a high probability that this is relevant despite the misclassification. Therefore, data from different sources should be updated differently based on their similarity level to the target. For example, in case of a misclassified but similar to the target source instance, the weight decrease should be less severe compared to less similar data. Through this, the model does not completely disregard potentially valuable data from the source domain, maintaining a balance between similarity and the impact of errors.

WEIGHTS NORMALIZATION

In Transfer AdaBoost, normalizing the weights of both target and source instances is an important step of each iteration to prevent excessive increases in target instance weights. However, significant issues arise when normalizing all weights together. When the weight updating strategy causes a proportion of weights to be increased or decreased and then all are normalized, the actual relative differences change. This leads to three main issues that can impact the learning process and the balance between target and source data.

The first issue is that the normalization process can cause changes to be either stronger or less impactful than intended, depending on the other weight adjustments. For example, when the weights of misclassified instances in the source domain decrease and are compared to the

potentially increased weights in the target, the decreased weights decrease even more when normalized, as depicted in Figure 7.3a. When weights have already decreased a lot, any attempts of change are not impactful or visible. Similarly, as shown in Figure 7.3b, the weights of the correctly classified source instances that should remain stable, when compared to the overall increase of other instances, also eventually decrease after normalization. Consequently, normalization drives the source's contribution to diminish due to the weights convergence to zero.

Furthermore, normalization can change the effect of the desired update. According to all equations above, during the same phases of prediction (i.e., consecutive iterations leading to the same prediction for an instance), the effect of change in weights should be the same, either weight increase or decrease. However, according to Figure 7.3c, there are obvious changes during the same phases.

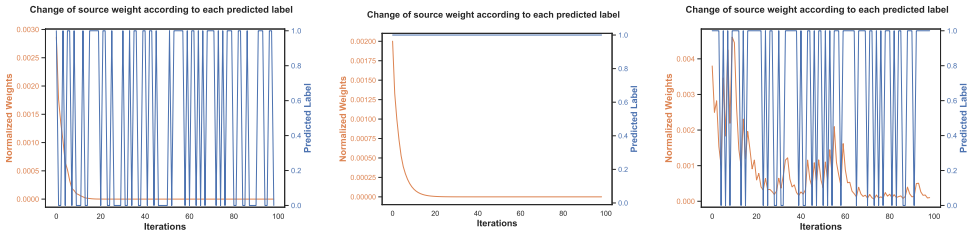
To address the first issue, it is important to normalize the weights of target and each source separately. This facilitates maintaining the unique characteristics and importance of instances within each domain. However, in this case, depending on the number of instances in each source, data in the source can overpower the data in the target. To prevent this, a 50% threshold is set for the normalization of source weights.

In a way to handle the issue regarding the change of expected effects in the target domain, a slight increase in the originally steady weights is necessary. This adjustment ensures that the contribution of the correctly classified target instances is not diminished by the normalization process, maintaining to some extent its level of importance for the following iterations. This is described in Equation 7.5, where the zero difference of $|y_{target_i} - y'_{target_i,r}|$ is adjusted by a parameter ϵ. This parameter can be optimized based on the examined data, and for this setting, $\epsilon = 0.2$ is selected. By all these adjustment strategies, the model can more effectively take advantage of target and source, but also balance the influence of the correctly classified instances, potentially leading to an improved performance.

$$w_{r+1}^{target_i} = w_r^{target_i} \cdot (1 - F_1^r)^{-\epsilon}, \text{ when } y_{target_i} - y'_{target_i,r} = 0 \tag{7.5}$$

PREDICTIONS AGGREGATION

After the specified number of iterations, where all weak learners have been sequentially trained, their predictions need to be combined in a way to produce a strong final prediction. Using AdaBoost, the predictions are aggregated through weighted voting, where each model's contribution is proportional to its accuracy. However, in TrAdaBoost, more complicated formulas are used for calculating each model's contribution and eventually, the aggregation of the last half iterations [197]. To simplify this, in our approach, a weighted average approach is used, where each model's prediction $y'_{i,r}$ of sample i at iteration r is weighted according to



(a) Weight convergence to zero.

(b) Weights decrease although correctly classified.

(c) Weight's effect change within the same prediction phase (consecutive same predictions).

Figure 7.3: Examples of the normalization issues. The orange line represents the weights of a source/target instance over 100 iterations, while the blue line shows the predicted labels $y'_{i,r}$ (0 or 1) that cause the weight changes.

its F_1^r performance. This is described in Equation 7.6. This way, while the learners of all iterations participate in determining the final prediction y'_i, good-performing models have more influence on the final decision.

$$y'_i = \frac{\sum_r F_1^r \cdot y'_{i,r}}{\sum_r F_1^r} \qquad (7.6) \qquad AUC = \frac{\sum_r F_1^r \cdot AUC_r}{\sum_r F_1^r} \qquad (7.7)$$

Similarly, for calculating the aggregated Area Under the Receiver Operating Characteristic Curve (AUC) metric, several approaches can be considered. A solution is given by using the F1-based weighted average of probabilities, which are then used for calculating the final AUC. Alternatively, another solution is to calculate the probabilities and subsequently compute the AUC metric for each learner, denoted as AUC_r. According to Equation 7.7, a weighted average of these AUC scores can be obtained using F1-based weights for each learner, resulting in an overall AUC score, AUC. In this analysis, the latter method will be employed to ensure that each learner's contribution is appropriately reflected in the final metric.

7.4. EXPERIMENTAL SETUP

7.4.1. EXAMINED EMA DATASET

The examined EMA dataset is the real-world NSMD dataset, described in Section 2.6. Before modeling, EMA data preprocessing is crucial in preparing both target and source data. The initial step in this process involves splitting the data into training and test subsets. Each individual's data is divided using always a 70:30 split ratio. Because these are time-series data, the split is performed sequentially and not randomly,

meaning that the first 70% is used for training and the last 30% for testing. For a fair comparison, all models are evaluated on the test sets of each individual in the target. The next preprocessing step is data normalization. Each individual dataset is normalized separately on training and test sets, allowing for transforming and aligning the feature space across target and source.

7.4.2. OUTPUT TASK

In this setting, the output task is 1-lag binary classification. This means that for all variables at time-point $t-1$, we aim to predict all corresponding variables at time-point t. Due to an observed sparsity in the original $1-7$ scale, the variables at time-point t are dichotomized, transforming them into binary outcomes regarding their relation to the average of each individual variable. This approach focuses on predicting whether each variable changes towards a positive or negative state from one time-point to the next, without accessing the exact rating. Thus, this setting facilitates a more straightforward analysis, predicting the variables' point-to-point transition.

7.4.3. EXPERIMENTAL SETTING

VALIDATION METRIC

Since the task is multivariate binary classification, the evaluation of the proposed Transfer AdaBoost approach is conducted using appropriate classification metrics, averaged across all variables. Given the sparsity in each output variable, it is important to note that our individual datasets are quite imbalanced. To address this, two metrics that take into account such output characteristics are the F1 score and AUC. Through these, the focus is equally split on the correct classification of both majority and minority classes. While using AUC, where different classification thresholds are considered in calculations, the performance reflects both imbalanced classes.

BASE LEARNER

In each iteration, when focusing on a specific target individual, a weak classifier is trained using the training data of the target and the utilized source data. This learner is typically a simple algorithm, such as a decision stump, where the ensemble of all these could lead to a strong classifier. In the current experiments, a decision tree with its depth set to 2 is used as the base learner. Given that the total number of iterations is set to 100, 100 base learners are trained to target one individual. This process will be repeated for all available individuals and all 12 variables separately, resulting in the training of 2244 models for each examined case that are discussed below.

BASELINE COMPARISONS

To assess the performance of the proposed improvements in the Transfer AdaBoost approach, we need to conduct some baseline comparisons against several approaches:

- Personalized classification using the examined base learner (Decision tree with depth equal to 2): Only the training data of the target is used to train the model that is later examined in a transfer learning scenario.
- Personalized classification using AdaBoost and Explainable Boosting Machines (EBMs, as described in Section 2.4): In a personalized setting, a model is trained only on the target domain data without incorporating any source domain data.
- Original TrAdaBoost classification: The original implementation of TrAdaBoost is trained using data from 2 and 10 similar sources.

7.5. EXPERIMENTAL RESULTS

In this section, a thorough analysis of the different experimental choices made in the proposed Transfer AdaBoost approach is presented. Specifically, in Experiment A, we examine the impact of varying the number of individuals in the source domain as well as the method of updating the source weights on the overall performance. The overall performance is evaluated in terms of F1 score, averaged over all 12 variables. After identifying the optimal experimental setting for TrAdaBoost, in Experiment B, we compare this with some baseline modeling approaches.

7.5.1. EXPERIMENT A

To explore the optimal number of sources, we conducted experiments using different numbers of sources, where always the most similar, in terms of GAK, are used. For instance, we compared the performance of using the single most similar source domain for each target against combining multiple source domains, with a maximum of the 10 most similar. At the same time, different source reweighting strategies are examined. These include the proposed strategy of Section 7.3.2, involving both source weights increase and decrease, and the original concept of weight decreasing. For the proposed approach, the impact of different numbers of *step* is also examined. All the examined strategies have been adapted to the changes proposed in Section 7.3.2. Figure 7.4 presents the boxplots of the F1 performance results across all 187 individuals.

The extracted results reveal distinct trends regarding the optimal number of sources based on the source reweighting strategy employed. Starting with a single source domain, all reweighting strategies yielded a similar performance, with the original weight-decreasing strategy providing the highest average F1 score at 0.51. This was followed by the

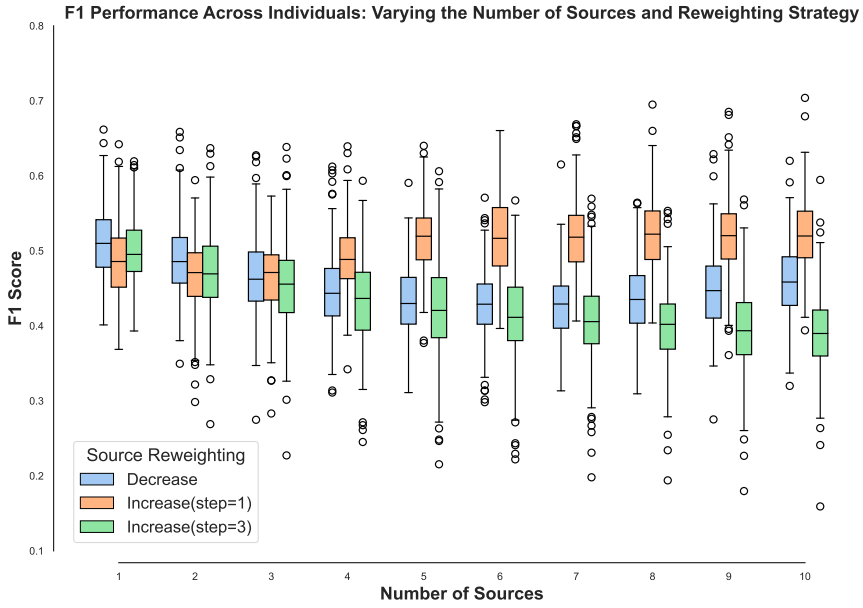


Figure 7.4: Comparison of experimental settings.

increasing strategy with $step = 3$ and $step = 1$. The same trend continued, with decreasing scores, until incorporating four or more sources. After this point, the performance of the increasing strategy with $step = 1$ starts rising significantly, with the average value converging to approximately 0.53. This shows that additional sources can provide valuable information and improve performance depending on how we deal with the weights of the source data. On the contrary, for some strategies, more source data is being handled like noise.

It should also be noted that the relative individual performance also depends on the similarity degree of sources to each target. Although the sources are particularly selected for each target, always picking the most similar ones in each case, the level of similarity can still vary. Some sources are more closely related to the target than others. Especially, when the number of sources increases, it becomes less likely that all sources will be highly similar to the target. This is also indicated by the fact that a milder increasing ($step = 1$) approach starts becoming effective as more sources are added, while for a small number of sources, weight increasing with $step = 3$ shows a better performance. Future experiments could explore a more refined approach by setting a threshold on the similarity degree between the target and the selected sources. This strategy would ensure that only the most relevant sources are utilized, leading to a more focused and effective transfer of knowledge.

7.5.2. EXPERIMENT B

To further investigate the impact of the proposed methodology for TrAdaBoost, a thorough comparison was conducted against several baseline methods. These methods include both personalized models without any transfer learning components (Decision Tree, AdaBoost and EBMs) and original TrAdaBoost (using 1 and 10 sources). The aim is to evaluate the effectiveness of our proposed approach in two scenarios. Beyond the best scenario identified through Experiment A, we examine a scenario where the sources are selected based on a clustering result, referred to as TrAda_cl. Specifically, as derived from Chapter 4 (Section 4.5.3), we apply kernel k-means clustering (using the GAK similarity and $k = 3$) to group the EMA data and then use the individuals of the same cluster as sources. The produced F1 and AUC scores are demonstrated in Figures 7.5 and 7.6, respectively.

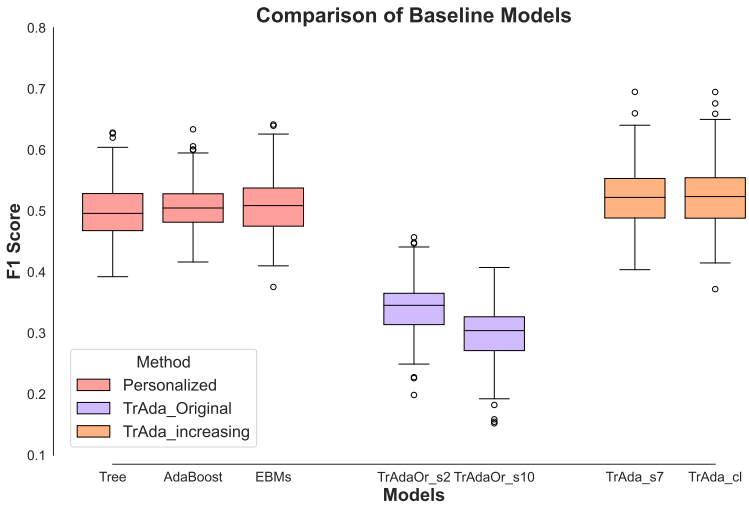


Figure 7.5: Comparison of the proposed TrAdaBoost enhancements (TrAda) based on F1 scores against baseline approaches, including personalized methods (Tree, AdaBoost, EBMs) and the original implementation of TrAdaBoost using 2 (TrAdaOr_s2) and 10 (TrAdaOr_s10) sources. For the proposed TrAdaBoost enhancements, the two distributions (highlighted in orange) represent the incorporation of 7 sources (TrAda_s7) and sources derived from clustering (TrAda_cl).

The proposed enhancements in TrAdaBoost provide the highest F1 scores, demonstrating the effectiveness of our refined approach, reaching an F1 score of 0.53. The clustering-based source selection also shows promising results, especially if we consider all the statistical properties of the individuals' distribution. This suggests that using structurally similar sources can further enhance the performance. Although the average

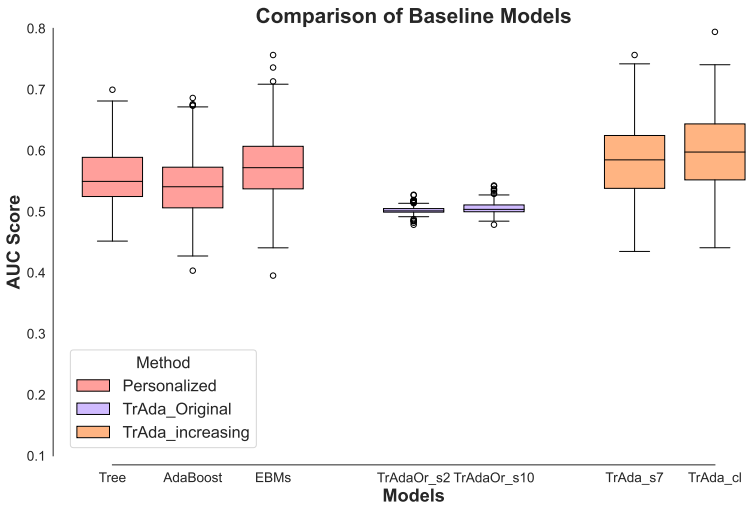


Figure 7.6: Comparison of the proposed TrAdaBoost enhancements (TrAda) based on AUC scores against baseline approaches, including personalized methods (Tree, AdaBoost, EBMs) and the original implementation of TrAdaBoost using 2 (TrAdaOr_s2) and 10 (TrAdaOr_s10) sources. For the proposed TrAdaBoost enhancements, the two distributions (highlighted in orange) represent the incorporation of 7 sources (TrAda_s7) and sources derived from clustering (TrAda_cl).

improvement compared to personalized models is not significant, further statistical analysis revealed that 132 out of 187 individuals achieved an improved F1 score (average increase of 5.1%) compared to trees, while 116 individuals (average increase of 3.5%) compared to both AdaBoost and EBMs. This indicates that additional information improves the majority of individuals, but not all, compared to the simple models. Therefore, a more extensive exploration should be performed on an individual- and feature- level to determine which cases have the potential to be improved. For instance, it is expected that transfer learning can improve predictive performance in scenarios with limited target data. Moreover, a similar trend is observed when comparing the baseline models against the AUC scores in Figure 7.6, with the enhanced TrAdaBoost models yielding the highest performance. More specifically, we can distinguish a slightly elevated average score for both EBMs and TrAda_cl reaching an average of 0.57 and 0.6, respectively. The potential of incorporating clustering is further highlighted by the AUC scores over personalized models, with clustering-based TrAdaBoost showing an improvement of 7.5% compared to tree models, 10.7% compared to AdaBoost, and 4.8% compared to EBMs. This demonstrates its significant contribution to enhancing the modeling process.

It is also interesting to notice that the original TrAdaBoost, when us-

ing either a single or multiple sources, exhibits very poor performance on our dataset. This was expected based on all the issues discussed in Section 7.2. Despite utilizing the same set of similar sources, the low scores could be attributed to the original modeling processes regarding weight updating and, most importantly, prediction aggregation, where only the last half of the iterations are taken into account. Such poor performance highlights the need for improved methodologies in transfer learning. Our proposed approach addresses these limitations by incorporating a more sophisticated weighting strategy and thoroughly evaluating the relevance of source data. These enhancements help ensure that the model is not only more adaptable but also achieves improved predictive accuracy and reliability across diverse individuals.

7.6. CONCLUSION

In a way to balance the personalized and nomothetic modeling approaches, this chapter presents an enhanced version of Transfer AdaBoost aiming to improve the predictive performance of individual EMA-based models. After identifying and discussing the issues of the original implementation of TrAdaBoost, our proposed methodological enhancements address various aspects of the modeling processes. More specifically, our approach focuses on the optimal selection of similar source domains as well as their careful exploitation, through advanced source reweighting strategies. After a set of experiments investigating the impact of all these choices, the results highlight that the initial source weight increase is a necessary step, emphasizing the difficult source instances before it is determined whether they may not significantly contribute to the target. In that case, incorporating more source domains also positively impacts the overall performance. Compared to the baseline methods, the proposed approach proved to outperform the original TrAdaBoost, but the average performance gain over personalized models was not significant. Interestingly, when examining the average percentage of individual differences, the improvement in F1 score reaches 5.1%. This suggests that, whereas our proposed methodology provides a more advanced approach to transfer learning, a thorough investigation is necessary to identify the specific cases where it may not be very effective. For instance, when there is a significant distribution shift between the training and test data for each target individual, the transfer learning model may struggle to generalize effectively. Additionally, when a target individual's characteristics do not sufficiently match those of the selected source data, the knowledge transfer may be less effective. In such cases, incorporating a threshold cutoff to determine the minimum similarity between target and source individuals could serve as an alternative strategy. This would help ensure that only sufficiently relevant source data is used, potentially enhancing the model's performance and reliability.

8

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation addressed significant challenges in the field of psychopathology, specifically within the context of modeling Ecological Momentary Assessment (EMA) data. The primary objective was to develop more advanced predictive models than the baseline linear network models, improving their accuracy and robustness. Such advanced models could better reflect individual EMA patterns and ultimately facilitate our understanding of mental disorders. This research systematically explored and evaluated various predictive modeling approaches, focusing on both individual and group levels. A significant part of this work was dedicated to various methodologies for identifying homogeneous groups of individuals as well as effectively utilizing these groupings to improve model predictive accuracy and personalization. By taking advantage of individual variability and commonalities within data, this work could provide additional enhancement towards refining individual or personalized models.

This chapter provides a summary of the contributions of this dissertation, discussing how each of the initial research questions was addressed. Finally, the chapter concludes by proposing potential directions for future work. These suggestions aim to explore more advanced integration of group-based approaches and expand the understanding of individual dynamics and patterns shared within groups in psychopathology.

8.1. ADDRESSING RESEARCH QUESTIONS

At the beginning of this dissertation, in Section 1.7, five different research questions were defined. All these are separately discussed as follows:

RQ1

Are non-linear individual models capable of outperforming the linear network models?

Research Question 1 (Chapter 3) explored the effectiveness of linear and non-linear personalized modeling approaches for analyzing EMA data. Although, traditionally, this research area has been dominated by network models [26, 37], which are linear, this chapter went a step further towards introducing the use of non-linear ML models. Non-linear models, such as decision tree-based models, have been widely used models known for their ability to extract complex, non-obvious patterns that linear models might ignore. Discovering such patterns is particularly crucial in the field of psychopathology, where the inherent complexity of mental disorders frequently presents non-linear dynamics and involves non-linear interactions.

Among various non-linear models that were examined, a recent implementation of Explainable Boosting Machines (EBMs) emerged as a prominent example [105, 106]. EBMs represent a sophisticated type of boosting-based model that integrates flexible, non-linear feature functions along with pairwise feature interactions. Through such functions, EBMs can effectively model complex relationships within the data while maintaining interpretability. The interpretability with state-of-the-art accuracy scores reported in other published works (e.g., [107]) made them particularly valuable in psychopathology. Motivated by these strengths, we further extended the investigation into their application across different output tasks and chapters, specifically in Chapters 5 and 7.

After applying advanced non-linear techniques, experimental results have demonstrated potential for improving modeling accuracy across different classification scenarios, predicting the occurrence of future EMA-relevant events within psychopathology. Specifically, these were tested using both real-world and synthetic datasets. In real-world datasets, the distribution of AUC scores for non-linear models showed less variability compared to linear ones, showing greater consistency among individuals' results. Nevertheless, it is important to note that model performance significantly depends on the size and quality of the data characteristics specific to each individual, leading to varied results across individuals, datasets, and the models examined. Moreover, the generation and utilization of synthetic data, which was crucial in addressing the challenge of limited data availability, played an important role in this research. These synthetic datasets, designed to mimic real-world conditions, were further explored in Chapter 4.

The shift from traditional linear models to advanced non-linear methods marks a significant advancement in EMA modeling. This transition is driven by the need for more accurate and robust individual predictions. Consequently, these improved models could provide more reliable descriptions of different individuals, capturing their complex underlying

dynamics.

RQ2

Could nomothetic modeling approaches, by integrating more data, exceed the predictive performance of individual models?

Two different research directions were explored to investigate the effectiveness of nomothetic or group-based modeling approaches and answer this research question. The first approach involved training models on aggregated data (referred to as using-all-data models), which includes data from all available individuals within the same data collection (population data). This offers the opportunity to capture patterns of behaviors and processes that are shared between individuals. Models developed from such aggregated data capture broader patterns across individuals, making them useful in situations where personalization is not possible due to a lack of individual data. Such models can then be effectively applied to new individuals who were not included in the original dataset. However, the generalizability of these models depends on the similarity between the individuals in the training data and those in the target population. Extensive experiments, conducted in Chapter 3, demonstrated successful performance on both synthetic and real-world datasets, with high AUC scores achieved when predicting 1-lag events. Particularly, in the challenging ThinkSlim2 dataset [123], it is interesting to notice that the AUC performance shows an average improvement of 14% compared to the relatively poor performance of personalized models.

Building on this success, the second approach focused on further refining the process of selecting individuals to form cohesive sub-population groups, and, hence optimizing the input of group-based models. Through clustering, the aim is to produce accurate and meaningful groupings, ensuring that the derived cluster-based models are tailored to the distinct characteristics of grouped individuals. In Chapter 5, after having explored two model-based clustering methods, their effectiveness was evaluated in a downstream forecasting task. The results highlighted that clustering was found to enhance the overall forecasting performance, compared to the personalized, aggregated as well as randomized cluster models. Thus, the results confirmed that the superiority of clustering performance is not a random effect arising from the use of a mixture of models. Specifically, clustering enhanced the overall performance, achieving a maximum of 7.19% MSE improvement over personalized models and 7.99% over the using-all-data models.

The insights gained from both approaches demonstrate that incorporating additional information from other individuals can significantly enhance models' predictive accuracy and relevance in complex scenarios. While personalized models offer several advantages, particularly in cases where sufficient individual data is available, the results of this research indicated that nomothetic methods by integrating population-

level and sub-population group (or cluster) information can enhance the relevance and generalization of predictive models. This approach is especially beneficial in the context of psychopathology, where individual data is often limited, and understanding shared behavioral patterns can lead to better interventions and treatment strategies.

RQ3

How could nomothetic modeling approaches effectively integrate group-based information while maintaining the focus on individual data?

Having identified the strengths and weaknesses of different modeling strategies, including idiographic as well as nomothetic and cluster-based approaches, the focus shifted towards achieving a balance between models that are either too broad or overfit to a particular individual. To address this question, we proposed two methodologies integrating both group- and individual-focused strategies. First, we proposed the Knowledge Distillation (KD) approach in Chapter 3 [74]. This approach strengthens the using-all-data nomothetic model, by sequentially integrating personalized models. According to the evaluation results on both synthetic and real-world data, the performance of the KD method improved significantly, showing a maximum AUC improvement of 17% compared to personalized models and achieving results comparable to using-all-data models.

Second, another transfer learning approach was investigated in Chapter 7, taking advantage of both boosting and sample-weighting strategies tailored to different individuals (target and source data) to enhance the 1-lag predictive performance of all psychopathology-related variables. Inspired by the TrAdaBoost methodology [197], this chapter focused on adapting it to the context of EMA and enhancing most of the parts in the training process, including training a weak classifier, calculating the training error, reweighting all the instances in the target and source domain, and normalizing the weights. Since all these methodological enhancements play a significant role in the learning process, different choices were examined to optimize its effectiveness. An important enhancement was to change the role of the misclassified instances in the source domain. Instead of directly considering these as useless, coming from a different distribution to the target, some source instances could have been misclassified because they are hard or challenging but useful instances. Additionally, this assumes that individuals in the source should be carefully selected based on their temporal similarity to the target [148]. After exploring different updating strategies, we showed the necessity of initially increasing for a small number of iterations and then decreasing the weights of the misclassified source data. Moreover, great emphasis was placed on the optimal number of similar individuals in the source that could enhance the performance. To examine this, vari-

ous experiments were conducted on different numbers but also sources derived from clustering. The clustering-based source selection showed promising results, suggesting that using structurally similar sources can further enhance individual performance, with a maximum percentage of improvement of 5.1% in F1 score and 10.7% in AUC score.

By balancing group-level and personalized approaches, both the Knowledge Distillation method and transfer learning strategies demonstrated improvements in predictive accuracy across psychopathology-related tasks. The utilization of clustering-based additional individuals further emphasized the importance of selecting similar individuals to enhance model performance. Thus, combining nomothetic and idiographic information can offer a more accurate framework for the personalized modeling of psychopathology.

RQ4

What individual characteristics extracted from time-series can be used to effectively group individuals into homogeneous clusters?

Inspired by the success of nomothetic approaches, it was obvious that additional information from other individuals could complement and enhance the personalized model. However, large heterogeneity in the whole population necessitated the refinement of the individuals' subset by only utilizing meaningful clusters or groups of similar individuals in the modeling process. Therefore, a significant part of this work covered the exploration of two different categories of clustering methods adopted for EMA data. The first category of time-series clustering was investigated in Chapter 4. Considering that all well-known clustering algorithms, like k-means or hierarchical clustering, can be used for time-series, challenges relate to the selection of all clustering-related predefined parameters (e.g., number of clusters), the appropriate distance metric, and how to evaluate the validity of the clustering solution. Due to the unsupervised nature of the problem, to evaluate their validity we conducted a large-scale EMA simulation study. The simulations cover different scenarios that mimic real-world cases, involving multiple individuals, noisy features and/or irregular time-series data. Through thorough experimentation, we showed that all methods achieved a good performance when applied to datasets with few or no noisy features. However, for datasets containing a high percentage of noisy features, as exhibited in real-world EMA data, employing more sophisticated data representations, such as kernel transformations, has great potential to better capture its unique characteristics and underlying patterns. A kernel transformation, such as the Global Alignment Kernel (GAK), allows for mapping the data into a higher-dimensional space, where complex, nonlinear relationships become more distinguishable, making it a promising first step when cluster-analyzing EMA data. Even in such cases, because of their structure complexity, the evaluation scores are not realistically

expected to be as good as in simulation studies. Thus, we should evaluate the scores in comparison to the produced values of other methods.

Beyond relying on time-series data, clustering can be investigated using alternative sources of information. According to model-based clustering approaches, each individual can be described by different characteristics or parameters extracted from their personalized model. In Chapter 5, we examined static characteristics, such as coefficients or feature importance, as well as more dynamically optimized information, like predictive performance. Through a series of experiments, all these approaches were assessed along with other important clustering-related choices, such as the number of clusters and the base model used. As a result, the superiority of clusters relying on dynamically optimized performance is confirmed by both the Silhouette coefficients and overall forecasting performance produced by group models.

Regarding the majority of real-world EMA datasets, it is important to acknowledge that there is no definitive answer about the true or optimal number and composition of groups. Each method, according to its objective function and parameters, aims to separate data in the most appropriate way. As a result, different methods may yield varying group separations, reflecting a distinct perspective on how the data can be best organized.

RQ5

How can we evaluate the time-series clustering results derived from different unsupervised clustering algorithms?

Having acknowledged that clustering can effectively enhance the modeling process, it becomes crucial to extract meaningful groups of individuals. However, given that clustering is an unsupervised task, evaluating its results poses significant challenges. To address this, besides only checking clusters quality and stability, we proposed two additional approaches. First, in Chapter 5, we evaluated the effectiveness of the clustering results by examining their performance in downstream predictive tasks. This approach allowed us to determine whether the clusters were not only well-formed but also practically useful for improving predictive accuracy. Specifically, we used the clusters to build group models and tested their ability to predict future outcomes, providing a comprehensive assessment of their utility. After comparing the performance of experiments, involving different k values and different clustering methods, we observed that cluster-based models consistently achieved lower loss scores, indicating more meaningful and effective clustering results. This approach demonstrates significant potential, as it not only outperformed personalized and using-all-data models but also provided straightforward insights into the practical applicability and predictive power of the clusters in real-world scenarios.

Following the predictive performance evaluation, we shifted our focus

to the explainability of the clusters. Clustering explainability ensures that the group patterns identified through clustering are comprehensible, explainable, and valid within the context of mental disorders. However, due to the inherent challenges of handling EMA data and limited available published work on clustering explainability, we got inspiration from literature focusing on interpretable classification models for explaining an output, which in our setting are the distinct cluster labels. A key advantage of this method is its versatility. It can be broadly applied to any set of cluster groupings used as output labels, regardless of the clustering method employed, allowing for the comparison and evaluation of clustering algorithms based on the quality and relevance of the insights they produce.

Specifically, in Chapter 6, we proposed an advanced interpretative deep-learning model, utilizing attention-based mechanisms, that could internally handle the complexities inherent in MTS data. Such deep-learning models can rely on the actual data dynamics rather than other data transformations to provide a clearer and more accurate comprehension of examined MTS. Among different experiments regarding its design, integrating both temporal and feature-level attention provided the best classification performance, potentially leading to a more accurate description of the model and, consequently, improved cluster interpretation. The power of employing two-level attention gave us the opportunity to identify the important time-points and variables that play primary roles in distinguishing between clusters.

Following this exploration, we focused on analyzing, summarizing, and interpreting the attention weights as well as evaluating the patterns underlying the important segments of the data that differentiate across clusters. Nevertheless, it is reasonable that the interpretability of the derived patterns is heavily influenced by the quality of the clustering results utilized in the framework. As observed, a sub-optimal clustering can be reflected when similarities in the influential data patterns of different clusters are extracted, suggesting a smaller number of clusters.

Besides this, explanations should always be assessed by domain experts on a case-by-case basis. Alternatively, explanations need to be investigated using some different types of ground truth or baseline information. This could include baseline data (such as demographic details or health-related information) collected before or during the study.

8.1.1. SUMMARY OF CONCLUSIONS

This dissertation explored various approaches aiming at building more robust and reliable machine learning models, focusing on improving both predictive accuracy and interpretability. The research was guided by key questions aimed at addressing the trade-off between personalized and generalizable models, the utilization of clustering techniques, and the role of transfer learning in enhancing predictions.

First, the investigation of non-linear individual models highlighted the

need for more flexible mechanisms capable of capturing the complex interactions underlying psychopathology. Experiments demonstrated that non-linear models outperformed the linear baseline models for predicting the short-term future of psychopathology-related variables, suggesting their potential to uncover more reliable data patterns reflective of psychopathological processes. However, the broad variation in performance across individuals showed that the EMA data quality and quantity of each individual played an important role. Despite the observed individual heterogeneity, it has been established that there are commonalities shared between different individuals. Therefore, nomothetic approaches, proposing aggregating data from all individuals, showed enhanced performance compared to personalized models.

Next, clustering techniques were employed to group individuals based on shared patterns. These clustering-based models not only outperformed both personalized and aggregated models but also offered practical insights into the underlying structure of the data. Clusters were found to capture meaningful patterns that improved the overall predictive accuracy, particularly when the number of clusters and clustering methods were carefully evaluated and optimized for the data.

The exploration of both idiographic and nomothetic models revealed that combining both approaches could provide a way to balance individual-specific predictions with generalizable insights. Methods based on Knowledge Distillation and transfer learning, adapted for multivariate time-series data, showed consistent improvement in the 1-lag predictive performance of all psychopathology-related variables. To summarize, all explored methodologies converged on a common principle: integrating group-level data with personalized insights enhances model robustness and predictive accuracy.

8.2. FUTURE DIRECTIONS

With the goal of gaining a deeper understanding of psychopathology and individual EMA patterns, this dissertation focuses on two main methodological shifts, setting the foundations for exploring complex non-linear models as well as more advanced group-based modeling strategies. The research findings demonstrate that these approaches significantly contributed to achieving the set objectives, while also highlighting potential new directions for future investigation. Specifically, this section presents a set of some potential directions:

- **Clustering based on similarities in EMA subsequences (shorter segments) of individuals' time-series.**

  In Chapter 4, a thorough exploration was conducted on clustering methods based on raw time-series data, mainly relying on global similarities across the entire temporal multivariate sequence. However, one of the key characteristics of EMA data is its dynamic nature with individuals exhibiting multiple patterns that evolve over

time. Therefore, relying on global similarities may overlook important, short-term patterns that are crucial for understanding individual variability and underlying psychopathological processes. Subsequences refer to shorter segments of an individual's time-series data that capture specific patterns over limited intervals of time. These segments can be extracted from the entire time-series sequence to identify recurring or meaningful local behaviors that may not be apparent when analyzing the entire sequence.

As a future direction, it would be valuable to shift the focus toward clustering based on similarities in subsequences [209]. Subsequence-based clustering focuses on capturing specific temporal windows during which individuals exhibit common patterns or trends, called time-series subsequences, offering deeper insights into individual trajectories and shared behaviors. This approach enables us to identify patterns that are otherwise hidden when analyzing the entire time-series. Moreover, by breaking down each individual's time-series data, we could uncover a set of fine-grained patterns, called motifs, which are frequently occurring and distinct patterns [210]. These motifs represent important segments that distinguish meaningful patterns from noise. Identifying motifs within subsequences provides a clearer understanding of the underlying dynamics in psychopathology.

Furthermore, based on subsequences, another direction would involve incremental or dynamic clustering, where individuals are allowed to change clusters over time as their patterns change [211]. Incremental clustering would allow models to adapt dynamically, repeatedly assigning individuals to clusters as more subsequences are revealed. Such clustering could be considered as a form of fuzzy clustering, as an individual may belong to multiple clusters at different time-points of the whole period. By reflecting real-world scenarios in mental disorders, such as comorbidities, this method would provide a more realistic and accurate grouping of individuals over time.

- **Additional data from sensors and digital phenotyping data.**

  Another promising direction is the utilization of additional data to EMA, which can be passively collected from smartphone's sensors during an EMA study [212]. These data include sensors' information from individual's geolocation (Global Positioning System, GPS), physical activity, sleep patterns, or other biometric information, as well as digital phenotyping data, time spent online, usage of specific applications, and response times to messages, etc [36]. Since such data are measured automatically, they can be collected in a much higher frequency. Therefore, they represent a rich additional dataset of objective measurements that can potentially provide valuable additional insights into individual aspects that may not be fully captured through self-reported EMA data.

By, subsequently, integrating these data into modeling approaches, we can build more accurate models that offer a richer understanding of individual variability and psychopathology [213]. However, further modeling challenges arise particularly due to the varying sampling frequencies among different sources of passive data, as well as compared to self-reported EMA data, which requires an advanced way for integration.

- **Causal ML for facilitating interventions and treatment.**

  To align with the final goal of the NSMD project, which is to facilitate individual intervention and treatment in a healthcare setting, another promising future direction is the application of causal ML techniques on EMA data [36]. Unlike traditional ML, which is mostly used in this dissertation, simply learning associations between different variables, causal ML aims to uncover the underlying causal mechanisms driving mental health outcomes [214]. For example, after having identified the most important variables of a model, causal ML can examine how these can influence changes in other variables. Such causal discoveries could be utilized in how a system would respond to an intervention. Thus, causality would ultimately allow for more targeted interventions. Moreover, causal ML can be used to predict the potential outcome in response to different treatments. Specifically, it could predict the risk of relapse under different treatment plans, guiding medical practitioners in selecting the most effective, personalized treatment strategy for each individual [215, 216].

8.3. CONCLUDING REFLECTION

In this dissertation, several advanced machine-learning modeling approaches were explored to address the complexities of EMA data in the field of psychopathology. Starting from individual non-linear models, it became evident that integrating data from other individuals could enhance the predictive power, reliability and robustness of the predictive models, particularly in targeting 1-lag future values. Specifically, by balancing idiographic and nomothetic methods, using transfer learning and clustering approaches, this work demonstrated how shared patterns and insights could improve individual predictions and, subsequently our understanding of individual and generalized group characteristics in mental disorders.

As we move forward, refining these methodologies along with incorporating emerging data sources, such as passive and sensor data, holds the potential to further improve both the accuracy and interpretability of predictive models. These advancements have real-world implications. By continuing to explore innovative methodologies, we can contribute to more personalized, scalable, and effective solutions in mental health

care. Such advancements contribute to more personalized and effective treatments and interventions for mental health disorders, but they can also help clinicians better understand individual variability and the underlying patterns of mental disorders. Ultimately, this work sets the stage for further advancements in data-driven psychopathology modeling, providing a foundation for future research to build upon.

BIBLIOGRAPHY

[1] M. W. DeVries. *The experience of psychopathology: Investigating mental disorders in their natural settings*. Cambridge University Press, 1992.

[2] E. I. Fried, C. D. van Borkulo, A. n. O. Cramer, L. Boschloo, R. A. Schoevers, and D. Borsboom. “Mental disorders as networks of problems: A review of recent insights”. In: *Social psychiatry and psychiatric epidemiology* 52 (2017), pp. 1–10.

[3] D. Borsboom, A. O. Cramer, V. D. Schmittmann, S. Epskamp, and L. J. Waldorp. “The small world of psychopathology”. In: *PloS One* 6.11 (2011), e27407.

[4] W. W. Eaton, S. S. Martins, G. Nestadt, O. J. Bienvenu, D. Clarke, and P. Alexandre. “The burden of mental disorders”. In: *Epidemiologic Reviews* 30.1 (2008), pp. 1–14.

[5] R. F. Krueger. “The structure of common mental disorders”. In: *Archives of General Psychiatry* 56.10 (1999), pp. 921–926.

[6] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove. “The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013”. In: *International Journal of Epidemiology* 43.2 (2014), pp. 476–493.

[7] D. Borsboom, A. O. Cramer, and A. Kalis. “Brain disorders? Not really: Why network structures block reductionism in psychopathology research”. In: *Behavioral and Brain Sciences* 42 (2019), e2.

[8] D. Borsboom. “A network theory of mental disorders”. In: *World Psychiatry* 16.1 (2017), pp. 5–13.

[9] K. S. Kendler, P. Zachar, and C. Craver. “What kinds of things are psychiatric disorders?” In: *Psychological medicine* 41.6 (2011), pp. 1143–1150.

[10] W. W. M. H. S. Consortium *et al.* “Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys”. In: *Jama* 291.21 (2004), pp. 2581–2590.

[11] R. V. Bijl, R. de Graaf, E. Hiripi, R. C. Kessler, R. Kohn, D. R. Offord, T. B. Ustun, B. Vicente, W. A. Vollebergh, E. E. Walters, *et al.* “The prevalence of treated and untreated mental disorders in five countries”. In: *Health Affairs* 22.3 (2003), pp. 122–133.

[12] N. Carragher, R. F. Krueger, N. R. Eaton, and T. Slade. “Disorders without borders: current and future directions in the metastructure of mental disorders”. In: *Social Psychiatry and Psychiatric Epidemiology* 50 (2015), pp. 339–350.

[13] A. Fagiolini, A. Goracci, *et al.* “The effects of undertreated chronic medical illnesses in patients with severe mental disorders.” In: *Journal of Clinical Psychiatry* 70.suppl 3 (2009), pp. 22–29.

[14] R. Kohn, S. Saxena, I. Levav, and B. Saraceno. “The treatment gap in mental health care”. In: *Bulletin of the World Health Organization* 82.11 (2004), pp. 858–866.

[15] L. H. Andrade, J. Alonso, Z. Mneimneh, J. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. De Girolamo, R. De Graaf, *et al.* “Barriers to mental health treatment: results from the WHO World Mental Health surveys”. In: *Psychological Medicine* 44.6 (2014), pp. 1303–1317.

[16] M. Solmi, S. Cortese, G. Vita, M. De Prisco, J. Radua, E. Dragioti, O. Köhler-Forsberg, N. M. Madsen, C. Rohde, L. Eudave, *et al.* “An umbrella review of candidate predictors of response, remission, recovery, and relapse across mental disorders”. In: *Molecular Psychiatry* 28.9 (2023), pp. 3671–3687.

[17] L. F. Bringmann and M. I. Eronen. “Don’t blame the model: Reconsidering the network approach to psychopathology.” In: *Psychological Review* 125.4 (2018), p. 606.

[18] M. E. Aristodemou, R. A. Kievit, A. L. Murray, M. Eisner, D. Ribeaud, and E. I. Fried. “Common cause versus dynamic mutualism: An empirical comparison of two theories of psychopathology in two large longitudinal cohorts”. In: *Clinical Psychological Science* 12.3 (2024), pp. 380–402.

[19] A. Bystritsky, A. Nierenberg, J. Feusner, and M. Rabinovich. “Computational non-linear dynamical psychiatry: A new methodological paradigm for diagnosis and course of illness”. In: *Journal of Psychiatric Research* 46.4 (2012), pp. 428–435.

[20] R. R. Sokal. “Classification: Purposes, Principles, Progress, Prospects: Clustering and other new techniques have changed classificatory principles and practice in many sciences.” In: *Science* 185.4157 (1974), pp. 1115–1123.

[21] D. American Psychiatric Association, D. American Psychiatric Association, *et al. Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Vol. 5. 5. American Psychiatric Association Washington, DC, 2013.

[22] D. A. Regier, E. A. Kuhl, and D. J. Kupfer. “The DSM-5: Classification and criteria changes”. In: *World Psychiatry* 12.2 (2013), pp. 92–98.

[23] A. Caspi and T. E. Moffitt. “All for one and one for all: Mental disorders in one dimension”. In: *American Journal of Psychiatry* 175.9 (2018), pp. 831–844.

[24] A. O. Cramer, L. J. Waldorp, H. L. Van Der Maas, and D. Borsboom. “Comorbidity: A network perspective”. In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 137–150.

[25] J. J. Newson, V. Pastukh, and T. C. Thiagarajan. “Poor separation of clinical symptom profiles by DSM-5 disorder criteria”. In: *Frontiers in Psychiatry* 12 (2021), p. 775762.

[26] D. Borsboom and A. O. Cramer. “Network analysis: An integrative approach to the structure of psychopathology”. In: *Annual Review of Clinical Psychology* 9 (2013), pp. 91–121.

[27] P. J. Jones, A. Heeren, and R. J. McNally. “Commentary: A network theory of mental disorders”. In: *Frontiers in Psychology* 8 (2017), p. 1305.

[28] D. Borsboom. “Psychometric perspectives on diagnostic systems”. In: *Journal of Clinical Psychology* 64.9 (2008), pp. 1089–1108.

[29] K. Börner, S. Sanyal, A. Vespignani, *et al.* “Network science”. In: *Annual Review of Information Science and Technology* 41.1 (2007), pp. 537–607.

[30] S. Milgram. “The small world problem”. In: *Psychology Today* 2.1 (1967), pp. 60–67.

[31] D. J. Watts and S. H. Strogatz. “Collective dynamics of ‘small-world’networks”. In: *Nature* 393.6684 (1998), pp. 440–442.

[32] L. F. Bringmann, T. Elmer, S. Epskamp, R. W. Krause, D. Schoch, M. Wichers, J. T. Wigman, and E. Snippe. “What do centrality measures measure in psychological networks?” In: *Journal of Abnormal Psychology* 128.8 (2019), p. 892.

[33] D. J. Robinaugh, R. H. Hoekstra, E. R. Toner, and D. Borsboom. “The network approach to psychopathology: A Review of the literature 2008–2018 and an agenda for future research”. In: *Psychological Medicine* 50.3 (2020), p. 353.

[34] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters. “Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication”. In: *Archives of General Psychiatry* 62.6 (2005), pp. 617–627.

[35] J. Wigman, J. Van Os, D. Borsboom, K. Wardenaar, S. Epskamp, A. Klippel, W. Viechtbauer, M. Wichers, *et al.* “Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach”. In: *Psychological Medicine* 45.11 (2015), pp. 2375–2387.

[36] A. Roefs, E. I. Fried, M. Kindt, C. Martijn, B. Elzinga, A. W. Evers, R. W. Wiers, D. Borsboom, and A. Jansen. “A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology”. In: *Behaviour Research and Therapy* 153 (2022), p. 104096.

[37] D. Borsboom, M. K. Deserno, M. Rhemtulla, S. Epskamp, E. I. Fried, R. J. McNally, D. J. Robinaugh, M. Perugini, J. Dalege, G. Costantini, *et al.* “Network analysis of multivariate data in psychological science”. In: *Nature Reviews Methods Primers* 1.1 (2021), p. 58.

[38] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, *et al.* “A survey of statistical network models”. In: *Foundations and Trends® in Machine Learning* 2.2 (2010), pp. 129–233.

[39] D. J. Robinaugh, N. J. LeBlanc, H. A. Vuletich, and R. J. McNally. “Network analysis of persistent complex bereavement disorder in conjugally bereaved adults.” In: *Journal of Abnormal Psychology* 123.3 (2014), p. 510.

[40] S. Guloksuz, L. Pries, and J. Van Os. “Application of network methods for understanding mental disorders: pitfalls and promise”. In: *Psychological Medicine* 47.16 (2017), pp. 2743–2752.

[41] S. Epskamp and E. I. Fried. “A tutorial on regularized partial correlation networks.” In: *Psychological Methods* 23.4 (2018), p. 617.

[42] L. von Klipstein, D. Borsboom, and A. Arntz. “The exploratory value of cross-sectional partial correlation networks: Predicting relationships between change trajectories in borderline personality disorder”. In: *PloS One* 16.7 (2021), e0254496.

[43] E. I. Fried, S. Epskamp, R. M. Nesse, F. Tuerlinckx, and D. Borsboom. “What are ’good’ depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis”. In: *Journal of Affective Disorders* 189 (2016), pp. 314–320.

[44] C. G. DeYoung and R. F. Krueger. “Understanding psychopathology: Cybernetics and psychology on the boundary between order and chaos”. In: *Psychological Inquiry* 29.3 (2018), pp. 165–174.

[45] E. I. Fried and A. O. Cramer. “Moving forward: Challenges and directions for psychopathological network theory and methodology”. In: *Perspectives on Psychological Science* 12.6 (2017), pp. 999–1020.

[46] S. Epskamp, C. D. van Borkulo, D. C. van der Veen, M. N. Servaas, A.-M. Isvoranu, H. Riese, and A. O. Cramer. “Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections”. In: *Clinical Psychological Science* 6.3 (2018), pp. 416–427.

[47] C. R. Brewin, B. Andrews, and I. H. Gotlib. “Psychopathology and early experience: A reappraisal of retrospective reports.” In: *Psychological Bulletin* 113.1 (1993), p. 82.

[48] T. E. Moffitt, A. Caspi, A. Taylor, J. Kokaua, B. J. Milne, G. Polanczyk, and R. Poulton. “How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment”. In: *Psychological Medicine* 40.6 (2010), pp. 899–909.

[49] M. Mestdagh and E. Dejonckheere. “Ambulatory assessment in psychopathology research: Current achievements and future ambitions”. In: *Current Opinion in Psychology* 41 (2021), pp. 1–8.

[50] J. Ruwaard, L. Kooistra, and M. Thong. “Ecological Momentary Assessment in Mental Health Research: A Practical Introduction, With Examples in R (–build 2018-11-26)”. In: *Amsterdam: APH Mental Health* (2018).

[51] S. Shiffman, A. A. Stone, and M. R. Hufford. “Ecological momentary assessment”. In: *Annual Review of Clinical Psychology* 4 (2008), pp. 1–32.

[52] R. Larson, M. Csikszentmihalyi, *et al.* “The experience sampling method”. In: *New Directions for Methodology of Social and Behavioral Science* 15.15 (1983), pp. 41–56.

[53] T. J. Trull and U. W. Ebner-Priemer. “Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section.” In: (2009).

[54] N. Bolger and J.-P. Laurenceau. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press, 2013.

[55] R. C. Moore, C. A. Depp, J. L. Wetherell, and E. J. Lenze. “Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults”. In: *Journal of Psychiatric Research* 75 (2016), pp. 116–123.

[56] M. Wichers, C. Simons, I. Kramer, J. A. Hartmann, C. Lothmann, I. Myin-Germeys, A. Van Bemmel, F. Peeters, P. Delespaul, and J. Van Os. “Momentary assessment technology as a tool to help patients with depression help themselves”. In: *Acta Psychiatrica Scandinavica* 124.4 (2011), pp. 262–272.

[57] B. Wild, M. Eichler, H.-C. Friederich, M. Hartmann, S. Zipfel, and W. Herzog. “A graphical vector autoregressive modelling approach to the analysis of electronic diary data”. In: *BMC Medical Research Methodology* 10 (2010), pp. 1–13.

[58] F. Schultze-Lutter, S. J. Schmidt, and A. Theodoridou. “Psychopathology—a precision tool in need of re-sharpening”. In: *Frontiers in Psychiatry* 9 (2018), p. 446.

[59] P. Dolce, D. Marocco, M. N. Maldonato, and R. Sperandeo. “Toward a machine learning predictive-oriented approach to complement explanatory modeling. an application for evaluating psychopathological traits based on affective neurosciences and phenomenology”. In: *Frontiers in Psychology* 11 (2020), p. 446.

[60] D. Stamate, A. Katrinecz, D. Stahl, S. J. Verhagen, P. A. Delespaul, J. van Os, and S. Guloksuz. “Identifying psychosis spectrum disorder from experience sampling data using machine learning approaches”. In: *Schizophrenia research* 209 (2019), pp. 156–163.

[61] A. B. Shatte, D. M. Hutchinson, and S. J. Teague. “Machine learning in mental health: A scoping review of methods and applications”. In: *Psychological Medicine* 49.9 (2019), pp. 1426–1448.

[62] T. Yarkoni and J. Westfall. “Choosing prediction over explanation in psychology: Lessons from machine learning”. In: *Perspectives on Psychological Science* 12.6 (2017), pp. 1100–1122.

[63] K. Husen, E. Rafaeli, J. Rubel, E. Bar-Kalifa, and W. Lutz. “Daily affect dynamics predict early response in CBT: Feasibility and predictive validity of EMA for outpatient psychotherapy”. In: *Journal of Affective Disorders* 206 (2016), pp. 305–314.

[64] R. Wang, W. Wang, M. S. Aung, D. Ben-Zeev, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, E. A. Scherer, *et al.* “Predicting symptom trajectories of schizophrenia using mobile sensing”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–24.

[65] S. L. Connolly and L. B. Alloy. “Rumination interacts with life stress to predict depressive symptoms: An ecological momentary assessment study”. In: *Behaviour Research and Therapy* 97 (2017), pp. 86–95.

[66] J. Torous, M. E. Larsen, C. Depp, T. D. Cosco, I. Barnett, M. K. Nock, and J. Firth. “Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps”. In: *Current Psychiatry Reports* 20.7 (2018), pp. 1–6.

[67] A. G. Wright and W. C. Woods. “Personalized models of psychopathology”. In: *Annual Review of Clinical Psychology* 16 (2020), pp. 49–74.

[68] N. R. Eaton, L. F. Bringmann, T. Elmer, E. I. Fried, M. K. Forbes, A. L. Greene, R. F. Krueger, R. Kotov, P. D. McGorry, C. Mei, *et al.* “A review of approaches and models in psychopathology conceptualization research”. In: *Nature Reviews Psychology* 2.10 (2023), pp. 622–636.

[69] C. R. van Genugten, J. Schuurmans, F. Lamers, H. Riese, B. W. Penninx, R. A. Schoevers, H. M. Riper, and J. H. Smit. “Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders”. In: *Journal of Clinical Medicine* 9.2 (2020), p. 322.

[70] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. “Learning from class-imbalanced data: Review of methods and applications”. In: *Expert Systems with Applications* 73 (2017), pp. 220–239.

[71] J. Wenzel, N. Dreschke, E. Hanssen, M. Rosen, A. Ilankovic, J. Kambeitz, A.-K. Fett, and L. Kambeitz-Ilankovic. “Ecological momentary assessment (EMA) combined with unsupervised machine learning shows sensitivity to identify individuals in potential need for psychiatric assessment”. In: *European Archives of Psychiatry and Clinical Neuroscience* 274.7 (2024), pp. 1639–1649.

[72] R. P. Masini, M. C. Medeiros, and E. F. Mendes. “Machine learning advances for time series forecasting”. In: *Journal of Economic Surveys* 37.1 (2023), pp. 76–111.

[73] S. Al-Stouhi and C. K. Reddy. “Transfer learning for class imbalance problems with inadequate data”. In: *Knowledge and Information Systems* 48 (2016), pp. 201–228.

[74] G. Hinton, O. Vinyals, and J. Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[75] D. O. Hoare, D. S. Matteson, and M. T. Wells. “K-ARMA Models for Clustering Time Series Data”. In: *arXiv preprint arXiv:2207.00039* (2022).

[76] I. Myin-Germeys, Z. Kasanova, T. Vaessen, H. Vachon, O. Kirtley, W. Viechtbauer, and U. Reininghaus. “Experience sampling methodology in mental health research: new insights and technical developments”. In: *World Psychiatry* 17.2 (2018), pp. 123–132.

[77] T. Kuhlmann, M. Dantlgraber, and U.-D. Reips. “Investigating measurement equivalence of visual analogue scales and Likert-type scales in Internet-based personality questionnaires”. In: *Behavior Research Methods* 49 (2017), pp. 2173–2181.

[78] M. W. Heymans and J. W. Twisk. “Handling missing data in clinical research”. In: *Journal of Clinical Epidemiology* 151 (2022), pp. 185–188.

[79] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. “A survey on missing data in machine learning”. In: *Journal of Big Data* 8 (2021), pp. 1–37.

[80] S. Fielding, P. M. Fayers, A. McDonald, G. McPherson, M. K. Campbell, and R. S. Group. “Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data”. In: *Health and Quality of Life Outcomes* 6 (2008), pp. 1–9.

[81] K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen. “Time series cluster kernel for learning similarities between multivariate time series with missing data”. In: *Pattern Recognition* 76 (2018), pp. 569–581.

[82] V. I. Paulsen and M. Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152. Cambridge university press, 2016.

[83] J. M. Haslbeck, A. Jover-Martínez, A. J. Roefs, E. I. Fried, L. H. Lemmens, E. Groot, and P. A. Edelsbrunner. “Comparing Likert and Visual Analogue Scales in Ecological Momentary Assessment”. In: (2024).

[84] L. Li and B. A. Prakash. “Time series clustering: Complex is simpler!” In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011, pp. 185–192.

[85] M. Cuturi. “Fast global alignment kernels”. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011, pp. 929–936.

[86] L. F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, and F. Tuerlinckx. “A network approach to psychopathology: new insights into clinical longitudinal data”. In: *PloS One* 8.4 (2013), e60188.

[87] S. Epskamp, J. Kruis, and M. Marsman. “Estimating psychopathological networks: Be careful what you wish for”. In: *PloS One* 12.6 (2017), e0179891.

[88] J. C. Biesanz. “Autoregressive longitudinal models.” In: (2012), pp. 459–471.

[89] L. F. Bringmann. “Person-specific networks in psychopathology: Past, present, and future”. In: *Current Opinion in Psychology* 41 (2021), pp. 59–64.

[90] L. van der Krieke, A. C. Emerencia, E. H. Bos, J. G. Rosmalen, H. Riese, M. Aiello, S. Sytema, P. de Jonge, *et al.* “Ecological momentary assessments and automated time series analysis to promote tailored health care: A proof-of-principle study”. In: *JMIR Research Protocols* 4.3 (2015), e4000.

[91] M. Eichler. “A graphical approach for evaluating effective connectivity in neural systems”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457 (2005), pp. 953–967.

[92] G. Shukur and P. Mantalos. “A simple investigation of the Granger-causality test in integrated-cointegrated VAR systems”. In: *Journal of Applied Statistics* 27.8 (2000), pp. 1021–1031.

[93] A. Shojaie and E. B. Fox. “Granger causality: A review and recent advances”. In: *Annual Review of Statistics and Its Application* 9.1 (2022), pp. 289–319.

[94] F. Dablander and M. Hinne. “Node centrality measures are a poor substitute for causal inference”. In: *Scientific Reports* 9.1 (2019), p. 6846.

[95] K. Bulteel, F. Tuerlinckx, A. Brose, and E. Ceulemans. “Using raw VAR regression coefficients to build networks can be misleading”. In: *Multivariate Behavioral Research* 51.2-3 (2016), pp. 330–344.

[96] S. Epskamp, D. Borsboom, and E. I. Fried. “Estimating psychological networks and their accuracy: A tutorial paper”. In: *Behavior Research Methods* 50 (2018), pp. 195–212.

[97] S. Epskamp, L. J. Waldorp, R. Mõttus, and D. Borsboom. “The Gaussian graphical model in cross-sectional and time-series data”. In: *Multivariate Behavioral Research* 53.4 (2018), pp. 453–480.

[98] J. Haslbeck and L. J. Waldorp. “mgm: Estimating time-varying mixed graphical models in high-dimensional data”. In: *arXiv preprint arXiv:1510.06871* (2015).

[99] J. M. Haslbeck, L. F. Bringmann, and L. J. Waldorp. “A tutorial on estimating time-varying vector autoregressive models”. In: *Multivariate behavioral research* (2020), pp. 1–30.

[100] T. Krone, C. J. Albers, P. Kuppens, and M. E. Timmerman. “A multivariate statistical model for emotion dynamics.” In: *Emotion* 18.5 (2018), p. 739.

[101] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu. “Causal interpretability for machine learning-problems, methods and evaluation”. In: *ACM SIGKDD Explorations Newsletter* 22.1 (2020), pp. 18–33.

[102] T. J. Hastie. “Generalized additive models”. In: *Statistical Models in S*. Routledge, 2017, pp. 249–307.

[103] R. Tibshirani and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.

[104] Y. Lou, R. Caruana, and J. Gehrke. “Intelligible models for classification and regression”. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, pp. 150–158.

[105] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. “Accurate intelligible models with pairwise interactions”. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013, pp. 623–631.

[106] H. Nori, S. Jenkins, P. Koch, and R. Caruana. “Interpretml: A unified framework for machine learning interpretability”. In: *arXiv preprint arXiv:1909.09223* (2019).

[107] H. Nori, R. Caruana, Z. Bu, J. H. Shen, and J. Kulkarni. “Accuracy, interpretability, and differential privacy via explainable boosting”. In: International Conference on Machine Learning. PMLR. 2021, pp. 8227–8237.

[108] P. D. Soyster, L. Ashlock, and A. J. Fisher. “Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility.” In: *Psychology of Addictive Behaviors* (2021).

[109] G. Spanakis, G. Weiss, B. Boh, and A. Roefs. “Network analysis of ecological momentary assessment data for monitoring and understanding eating behavior”. In: Smart Health. Springer International Publishing, 2016, pp. 43–54.

[110] A. J. Martínez, L. Lemmens, E. I. Fried, and A. Roefs. “Developing a Transdiagnostic Ecological Momentary Assessment Protocol for Psychopathology.” In: (2023).

[111] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Using explainable boosting machine to compare idiographic and nomothetic approaches for ecological momentary assessment data”. In: International Symposium on Intelligent Data Analysis. Springer. 2022, pp. 199–211.

[112] A. G. Wright and J. Zimmermann. “Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement.” In: *Psychological Assessment* 31.12 (2019), p. 1467.

[113] L. F. Bringmann, C. Albers, C. Bockting, D. Borsboom, E. Ceulemans, A. Cramer, S. Epskamp, M. I. Eronen, E. Hamaker, P. Kuppens, *et al.* “Psychopathological networks: Theory, methods and practice”. In: *Behaviour Research and Therapy* 149 (2022), p. 104011.

[114] D. Cicchetti and F. A. Rogosch. “Equifinality and multifinality in developmental psychopathology”. In: *Development and Psychopathology* 8.4 (1996), pp. 597–600.

[115] I. H. Sarker. “Machine learning: Algorithms, real-world applications and research directions”. In: *SN Computer Science* 2.3 (2021), p. 160.

[116] A. M. Beltz, A. G. Wright, B. N. Sprague, and P. C. Molenaar. “Bridging the nomothetic and idiographic approaches to the analysis of clinical data”. In: *Assessment* 23.4 (2016), pp. 447–458.

[117] J. V. Tu. “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes”. In: *Journal of Clinical Epidemiology* 49.11 (1996), pp. 1225–1231.

[118] J. Gou, B. Yu, S. J. Maybank, and D. Tao. “Knowledge distillation: A survey”. In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819.

[119] L. Wang and K.-J. Yoon. “Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.6 (2021), pp. 3048–3068.

[120] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. “Model compression”. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006, pp. 535–541.

[121] S. Fukui, J. Yu, and M. Hashimoto. “Distilling Knowledge for Non-Neural Networks”. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE. 2019, pp. 1411–1416.

[122] B. Boh, L. H. Lemmens, A. Jansen, C. Nederkoorn, V. Kerkhofs, G. Spanakis, G. Weiss, and A. Roefs. “An Ecological Momentary Intervention for weight loss and healthy eating via smartphone and Internet: study protocol for a randomised controlled trial”. In: *Trials* 17.1 (2016), pp. 1–12.

[123] G. Spanakis, G. Weiss, B. Boh, L. Lemmens, and A. Roefs. “Machine learning techniques in eating behavior e-coaching”. In: *Personal and Ubiquitous Computing* 21.4 (2017), pp. 645–659.

[124] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Clustering individuals based on multivariate EMA time-series data”. In: The Annual Meeting of the Psychometric Society. Springer. 2022, pp. 161–171.

[125] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Evaluating multivariate time-series clustering using simulated ecological momentary assessment data”. In: *Machine Learning with Applications* 14 (2023), p. 100512.

[126] P. A. Jaskowiak, I. G. Costa, and R. J. Campello. “The area under the ROC curve as a measure of clustering quality”. In: *Data Mining and Knowledge Discovery* 36.3 (2022), pp. 1219–1245.

[127] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. “Time-series clustering–A decade review”. In: *Information Systems* 53 (2015), pp. 16–38.

[128] A. Javed, B. S. Lee, and D. M. Rizzo. “A benchmark study on time series clustering”. In: *Machine Learning with Applications* 1 (2020).

[129] E. E. Özkoç. “Clustering of time-series data”. In: *Data Mining–Methods, Applications and Systems* (2020), pp. 1–19.

[130] M. Ali, A. Alqahtani, M. W. Jones, and X. Xie. “Clustering and classification for time series data in visual analytics: A survey”. In: *IEEE Access* 7 (2019), pp. 181314–181338.

[131] A. Alqahtani, M. Ali, X. Xie, and M. W. Jones. “Deep time-series clustering: A review”. In: *Electronics* 10.23 (2021), p. 3001.

[132] N. Tavakoli, S. Siami-Namini, M. Adl Khanghah, F. Mirza Soltani, and A. Siami Namin. “An autoencoder-based deep learning approach for clustering time series data”. In: *SN Applied Sciences* 2 (2020), pp. 1–25.

[133] B. Hammer, A. Micheli, N. Neubauer, A. Sperduti, M. Strickert, *et al.* “Self organizing maps for time series”. In: Proceedings of Workshop on Self-Organizing Maps. Vol. 2005. 2005, pp. 115–122.

[134] J. Paparrizos, C. Liu, A. J. Elmore, and M. J. Franklin. “Querying Time-Series Data: A Comprehensive Comparison of Distance Measures.” In: *IEEE Data Engineering Bulletin* 46.3 (2023), pp. 69–88.

[135] C. Genolini, R. Ecochard, M. Benghezal, T. Driss, S. Andrieu, and F. Subtil. “kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes”. In: *Plos One* 11.6 (2016), e0150738.

[136] W. Choi, J. Cho, S. Lee, and Y. Jung. “Fast constrained dynamic time warping for similarity measure of time series data”. In: *IEEE Access* 8 (2020), pp. 222841–222858.

[137] M. Cuturi and M. Blondel. “Soft-dtw: A differentiable loss function for time-series”. In: International Conference on Machine Learning. PMLR. 2017, pp. 894–903.

[138] J. Zhao and L. Itti. “shapeDTW: Shape dynamic time warping”. In: *Pattern Recognition* 74 (2018), pp. 171–184.

[139] M. Vlachos, G. Kollios, and D. Gunopulos. “Discovering similar multidimensional trajectories”. In: Proceedings 18th International Conference on Data Engineering. IEEE. 2002, pp. 673–684.

[140] J. Paparrizos and L. Gravano. “k-shape: Efficient and accurate clustering of time series”. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015, pp. 1855–1870.

[141] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh. “Generalizing DTW to the multi-dimensional case requires an adaptive approach”. In: *Data Mining and Knowledge Discovery* 31.1 (2017), pp. 1–31.

[142] M. Badiane, M. O’Reilly, and P. Cunningham. “Kernel Methods for Time Series Classification and Regression.” In: AICS. 2018, pp. 54–65.

[143] K. Hebbrecht, M. Stuivenga, T. Birkenhäger, M. Morrens, E. Fried, B. Sabbe, and E. Giltay. “Understanding personalized dynamics to inform precision medicine: A dynamic time warp analysis of 255 depressed inpatients”. In: *BMC Medicine* 18.1 (2020), pp. 1–15.

[144] A. A. Wagan, S. Talpur, and S. Narejo. “Clustering uncertain overlapping symptoms of multiple diseases in clinical diagnosis”. In: *PeerJ Computer Science* 10 (2024), e2315.

[145] S. Nagesh, A. Moreno, S. M. Carpenter, J. Yap, S. Chatterjee, S. L. Lizotte, N. Wan, S. Kumar, C. Lam, D. W. Wetter, *et al.* “Transformers for prompt-level EMA non-response prediction”. In: *arXiv preprint arXiv:2111.01193* (2021).

[146] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui. “A kernel for time series based on global alignments”. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07. Vol. 2. IEEE. 2007, pp. II–413.

[147] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, *et al.* “Tslearn, a machine learning toolkit for time series data”. In: *Journal of Machine Learning Research* 21.118 (2020), pp. 1–6.

[148] F. Petitjean, A. Ketterlin, and P. Gançarski. “A global averaging method for dynamic time warping, with applications to clustering”. In: *Pattern Recognition* 44.3 (2011), pp. 678–693.

[149] G. Tzortzis and A. Likas. “The global kernel k-means clustering algorithm”. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE. 2008, pp. 1977–1984.

[150] H. Izakian, W. Pedrycz, and I. Jamal. “Fuzzy clustering of time series data using dynamic time warping distance”. In: *Engineering Applications of Artificial Intelligence* 39 (2015), pp. 235–244.

[151] Y. Ding and X. Fu. “Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm”. In: *Neurocomputing* 188 (2016), pp. 233–238.

[152] D. Xu and Y. Tian. “A comprehensive survey of clustering algorithms”. In: *Annals of Data Science* 2 (2015), pp. 165–193.

[153] U. Von Luxburg *et al.* “Clustering stability: An overview”. In: *Foundations and Trends® in Machine Learning* 2.3 (2010), pp. 235–274.

[154] A. Dachraoui, A. Bondu, and A. Cornuéjols. “Early classification of time series as a non myopic sequential decision making problem”. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15. Springer. 2015, pp. 433–447.

[155] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Model-based Clustering of Individuals’ Ecological Momentary Assessment Time-series Data for Improving Forecasting Performance”. In: BNAIC/ BeNeLearn 2023: Joint International Scientific Conferences on AI and Machine Learning. 2023.

[156] D. C. McLean, J. Nakamura, and M. Csikszentmihalyi. “Explaining system missing: Missing data and experience sampling method”. In: *Social Psychological and Personality Science* 8.4 (2017), pp. 434–441.

[157] T. W. Liao. “Clustering of time series data—a survey”. In: *Pattern Recognition* 38.11 (2005), pp. 1857–1874.

[158] Á. López-Oriona, P. Montero-Manso, and J. A. Vilar. “Clustering of Time Series Based on Forecasting Performance of Global Models”. In: International Workshop on Advanced Analytics and Learning on Temporal Data. Springer. 2022, pp. 18–33.

[159] Q. Ma, J. Zheng, S. Li, and G. W. Cottrell. “Learning representations for time series clustering”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[160] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier. “End-to-end deep representation learning for time series clustering: A comparative study”. In: *Data Mining and Knowledge Discovery* 36.1 (2022), pp. 29–81.

[161] M. Corduas and D. Piccolo. “Time series clustering and classification by the autoregressive metric”. In: *Computational Statistics & Data Analysis* 52.4 (2008), pp. 1860–1872.

[162] K. Kalpakis, D. Gada, and V. Puttagunta. “Distance measures for effective clustering of ARIMA time-series”. In: Proceedings 2001 IEEE International Conference on Data Mining. IEEE. 2001, pp. 273–280.

[163] S. Ghassempour, F. Girosi, and A. Maeder. “Clustering multivariate time series using hidden Markov models”. In: *International Journal of Environmental Research and Public Health* 11.3 (2014), pp. 2741–2763.

[164] M. Bicego. “K-random forests: A K-means style algorithm for random forest clustering”. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE. 2019, pp. 1–8.

[165] T. Shi and S. Horvath. “Unsupervised learning with random forest predictors”. In: *Journal of Computational and Graphical Statistics* 15.1 (2006), pp. 118–138.

[166] M. Längkvist, L. Karlsson, and A. Loutfi. “A review of unsupervised feature learning and deep learning for time-series modeling”. In: *Pattern Recognition Letters* 42 (2014), pp. 11–24.

[167] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. “A recurrent latent variable model for sequential data”. In: *Advances in Neural Information Processing Systems* 28 (2015).

[168] Q. Ma, S. Li, W. Zhuang, J. Wang, and D. Zeng. “Self-supervised time series clustering with model-based dynamics”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.9 (2020), pp. 3942–3955.

[169] A. Amelio and C. Pizzuti. “Is normalized mutual information a fair measure for comparing community detection methods?” In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015, pp. 1584–1585.

[170] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention”. In: Explainable Artificial Intelligence. Ed. by L. Longo, S. Lapuschkin, and C. Seifert. Cham: Springer Nature Switzerland, 2024, pp. 75–99. isbn: 978-3-031-63797-1.

[171] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[172] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič. “Feature importance estimation with self-attention networks”. In: *arXiv preprint arXiv:2002.04464* (2020).

[173] T.-Y. Hsieh, S. Wang, Y. Sun, and V. Honavar. “Explainable multivariate time series classification: A deep neural network which learns to attend to important variables as well as time intervals”. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021, pp. 607–615.

[174] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little. “What to do when K-means clustering fails: A simple yet principled alternative algorithm”. In: *PloS One* 11.9 (2016), e0162259.

[175] J. Pareek and J. Jacob. “Data compression and visualization using PCA and T-SNE”. In: Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019. Springer. 2021, pp. 327–337.

[176] M. Nguyen, S. Purushotham, H. To, and C. Shahabi. “m-tsne: A framework for visualizing high-dimensional multivariate time series”. In: *arXiv preprint arXiv:1708.07942* (2017).

[177] A. Bonifati, F. D. Buono, F. Guerra, and D. Tiano. “Time2Feat: learning interpretable representations for multivariate time series clustering”. In: *Proceedings of the VLDB Endowment* 16.2 (2022), pp. 193–201.

[178] O. Ozyegen, N. Prayogo, M. Cevik, and A. Basar. “Interpretable Time Series Clustering Using Local Explanations”. In: *arXiv preprint arXiv:2208.01152* (2022).

[179] H. Hwang and S. E. Whang. “XClusters: explainability-first clustering”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. 7. 2023, pp. 7962–7970.

[180] M. T. Ribeiro, S. Singh, and C. Guestrin. “Model-agnostic interpretability of machine learning”. In: *arXiv preprint arXiv:1606.05386* (2016).

[181] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: Advances in Neural Information Processing Systems. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[182] K. Vinogradova, A. Dibrov, and G. Myers. “Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract)”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 10. 2020, pp. 13943–13944.

[183] M. Moshkovitz, S. Dasgupta, C. Rashtchian, and N. Frost. “Explainable k-means and k-medians clustering”. In: International Conference on Machine Learning. PMLR. 2020, pp. 7055–7065.

[184] S. Bandyapadhyay, F. V. Fomin, P. A. Golovach, W. Lochet, N. Purohit, and K. Simonov. “How to find a good explanation for clustering?” In: *Artificial Intelligence* (2023), p. 103948.

[185] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic attribution for deep networks”. In: International Conference on Machine Learning. PMLR. 2017, pp. 3319–3328.

[186] U. Schlegel, D. L. Vo, D. A. Keim, and D. Seebacher. “Ts-mule: Local interpretable model-agnostic explanations for time series forecast models”. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer. 2021, pp. 5–14.

[187] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, Z. Zimmerman, D. F. Silva, A. Mueen, and E. Keogh. “Time series joins, motifs, discords and shapelets: A unifying view that exploits the matrix profile”. In: *Data Mining and Knowledge Discovery* 32 (2018), pp. 83–123.

[188] Y. Sun, J. Li, J. Liu, B. Sun, and C. Chow. “An improvement of symbolic aggregate approximation distance measure for time series”. In: *Neurocomputing* 138 (2014), pp. 189–198.

[189] M. Villani, J. Lockhart, and D. Magazzeni. “Feature Importance for Time Series Data: Improving KernelSHAP”. In: *arXiv preprint arXiv:2210.02176* (2022).

[190] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro. “Timeshap: Explaining recurrent models through sequence perturbations”. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021, pp. 2565–2573.

[191] E.-Y. Hsu, C.-L. Liu, and V. S. Tseng. “Multivariate time series early classification with interpretability using deep learning and attention mechanism”. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer. 2019, pp. 541–553.

[192] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Enhanced Boosting-based Transfer Learning for Modeling Ecological Momentary Assessment Data”. In: ML4ITS2023 - 3rd Workshop on Machine Learning for Irregular Time Series: Advances in Generative Models, Global Models and Self-Supervised Learning. ECML. 2024.

[193] K. Weiss, T. M. Khoshgoftaar, and D. Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3 (2016), pp. 1–40.

[194] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[195] Y. Yao and G. Doretto. “Boosting for transfer learning with multiple sources”. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE. 2010, pp. 1855–1862.

[196] A. Storkey. “When training and test sets are different: characterizing learning transfer”. In: *Dataset Shift in Machine Learning*. The MIT Press, 2008.

[197] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. “Boosting for transfer learning”. In: Proceedings of the 24th International Conference on Machine Learning. 2007, pp. 193–200.

[198] M. Iman, H. R. Arabnia, and K. Rasheed. “A review of deep transfer learning and recent advancements”. In: *Technologies* 11.2 (2023), p. 40.

[199] S. Niu, Y. Liu, J. Wang, and H. Song. “A decade survey of transfer learning (2010–2020)”. In: *IEEE Transactions on Artificial Intelligence* 1.2 (2020), pp. 151–166.

[200] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. “A survey on deep transfer learning”. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer. 2018, pp. 270–279.

[201] Y. Bengio. “Deep learning of representations for unsupervised and transfer learning”. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning. JMLR Workshop and Conference Proceedings. 2012, pp. 17–36.

[202] W. Ge and Y. Yu. “Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 1086–1095.

[203] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. “Parameter-efficient transfer learning for NLP”. In: International Conference on Machine Learning. PMLR. 2019, pp. 2790–2799.

[204] A. Asgarian, P. Sobhani, J. C. Zhang, M. Mihailescu, A. Sibilia, A. B. Ashraf, and B. Taati. “A hybrid instance-based transfer learning method”. In: *arXiv preprint arXiv:1812.01063* (2018).

[205] W. M. Kouw and M. Loog. “An introduction to domain adaptation and transfer learning”. In: *arXiv preprint arXiv:1812.11806* (2018).

[206] S. Sun, H. Shi, and Y. Wu. “A survey of multi-source domain adaptation”. In: *Information Fusion* 24 (2015), pp. 84–92.

[207] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. “Invariant models for causal transfer learning”. In: *Journal of Machine Learning Research* 19.36 (2018), pp. 1–34.

[208] S. Al-Stouhi and C. K. Reddy. “Adaptive boosting for transfer learning using dynamic updates”. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11. Springer. 2011, pp. 60–75.

[209] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh. “A review of subsequence time series clustering”. In: *The Scientific World Journal* 2014.1 (2014), p. 312521.

[210] S. Torkamani and V. Lohweg. “Survey on time series motif discovery”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7.2 (2017), e1199.

[211] V. Kuznetsov and M. Mohri. “Time series prediction and online learning”. In: Conference on Learning Theory. PMLR. 2016, pp. 1190–1213.

[212] R. Hays, M. Keshavan, H. Wisniewski, and J. Torous. “Deriving symptom networks from digital phenotyping data in serious mental illness”. In: *BJPsych Open* 6.6 (2020), e135.

[213] N. Martinez-Martin, T. R. Insel, P. Dagum, H. T. Greely, and M. K. Cho. “Data mining for health: staking out the ethical territory of digital phenotyping”. In: *NPJ Digital Medicine* 1.1 (2018), p. 68.

[214] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar. “Causal machine learning for predicting treatment outcomes”. In: *Nature Medicine* 30.4 (2024), pp. 958–968.

[215] B. Schölkopf. “Causality for machine learning”. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804.

[216] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O’Neil, and S. A. Tsaftaris. “Causal machine learning for healthcare and precision medicine”. In: *Royal Society Open Science* 9.8 (2022), p. 220638.

[217] M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Exploiting Individual Graph Structures to Enhance Ecological Momentary Assessment (EMA) Forecasting”. In: 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW). IEEE. 2024, pp. 158–166.

SUMMARY

The overall goal of this dissertation is to develop and evaluate advanced analysis methods tailored to the network approach for studying psychopathology, which offers a novel perspective on understanding mental disorders. The network approach to psychopathology conceptualizes mental disorders as a complex system of interacting psychopathology-related variables, such as emotions, behaviors and experiences, that directly influence one another. To study these interactions, this dissertation leverages time-intensive, repeated, intra-individual measurements using Ecological Momentary Assessment (EMA), a method that has grown significantly in psychology and health research over the past decade. EMA enables the real-time and contextual monitoring of various psychopathology-related variables, such as emotions, behaviors and experiences within participants' natural environment. This method allows the collection of temporal data regarding all these variables, providing great insights into their temporal dynamics.

The rich structure and granularity of the EMA data can be particularly valuable for building robust predictive models capable of forecasting the course of mental disorders, examining treatment responses, or developing tailored psychological interventions. Accurate predictive models are crucial not only for facilitating early interventions, potentially mitigating the severity of mental health episodes, but also for providing reliable representations of the interactions between key variables. These representations enrich our understanding of the complex mechanisms underlying mental disorders, offering a pathway to more personalized and effective therapeutic strategies.

As introduced in Chapter 2, a commonly used method for studying the network approach to psychopathology is through the Vector Autoregressive (VAR) model, which applies a linear framework to capture temporal dependencies and interactions between various psychopathology-related variables. While VAR has been widely adopted due to its interpretability, its linear nature may fail to fully capture the inherent complexity of mental disorders. Psychopathology-related variables probably exhibit non-linear interactions, reflecting the dynamic and complex nature of mental health processes. Relying solely on linear models risks oversimplifying these relationships and missing critical aspects of the underlying patterns.

Motivated by the constraints of linear models, this dissertation explores the use of advanced machine learning techniques, with a particular emphasis on non-linear models. Such advanced models offer the potential to recognize more complicated patterns among EMA data, enhancing

the accuracy of predictions related to the occurrence and intensity of different variables. While the initial focus is on developing personalized non-linear models that tailor predictions to individual patterns, this dissertation also investigates the integration of data from other individuals of the same EMA study to further enhance predictive performance. By combining these group-level insights with personalized approaches, the models achieve a balance between capturing individual specificity and extracting generalizable patterns, thus advancing predictive modeling in psychopathology.

Given the significant individual heterogeneity, where individuals exhibit unique EMA patterns, Chapter 3 starts with exploring the performance of non-linear models based on tree ensembles. In this chapter, we propose applying the Explainable Boosting Machines (EBMs), which is a non-linear interpretable model. The experimental evaluation demonstrated strong consistency in the results of the personalized (or idiographic) approach, with non-linear models significantly improving performance on both synthetic and real-world datasets.

One of the main challenges in building personalized models is the limited amount of data points available for each individual. These small datasets often result in models that cannot be trained at all, or in some cases, overfitted models that lack generalizability. To address this, data collected from other individuals in the same EMA study can prove beneficial for modeling. In Chapter 3, we also examine nomothetic prediction models, which integrate data from all individuals of the same EMA study. After a series of experiments, the proposed knowledge distillation method emerged as the most beneficial method for improving the performance of personalized models. This was achieved by transferring knowledge from a larger, more general model to a smaller, more specialized one, achieving up to a 17% improvement in AUC performance in one real-world dataset.

While nomothetic approaches with broad data integration offer valuable insights into general population trends, the considerable individual variability within large datasets can sometimes hide the unique individual patterns. By grouping individuals with shared characteristics, we can reduce noise in the data and focus on more specific patterns within subgroups of the population. To better refine the input data, strategically selecting meaningful groups of individuals could enhance our understanding of the underlying processes at both the individual and the group level. An effective way to achieve this is through clustering. Consequently, Chapters 4 and 5 focus on evaluating the performance of various clustering approaches for grouping individuals based on the similarity of their raw time-series data patterns and model-based information, respectively. Since clustering is an unsupervised task, where the true underlying groups are typically unknown, evaluating the results can be challenging. Therefore, Chapter 4 investigates various clustering methods and clustering-related parameters by analyzing data from simulations. The simulations are designed to mimic real-world EMA datasets,

involving multiple individuals, noisy features and/or irregular time-series data. The results, based on comparisons across various evaluation measures, suggest that employing alternative data representations, such as Global Alignment Kernel (GAK) transformations, has great potential to better capture the unique characteristics and underlying patterns of EMA data.

In addition to traditional intrinsic evaluation measures, Chapters 5 and 6 take different approaches to assess the efficacy of clustering methods and the practical usefulness of the created groups. First, Chapter 5 evaluates clustering within the context of a downstream predictive/forecasting task, providing a more comprehensive understanding of their practical utility. The results demonstrated that the superiority of clustering performance is not a random effect arising from using a mixture of models, but is instead driven by the quality of the clusters themselves. Achieving lower MSE scores with clustering indicates that using data from similar individuals helps capture more relevant patterns and make accurate predictions. Notably, the proposed performance-optimized clustering approach (POC) achieved a maximum improvement of 7.99% in MSE scores over the personalized models when using Random Forest. Second, Chapter 6 shifts the focus to explainability as an additional criterion for evaluating clustering. By generating explanations through an interpretable framework as well as analyzing the attention-derived important time-points and feature interactions at both cluster- and individual-level, valuable insights into the patterns characterizing each cluster are identified. These explanations can be particularly useful for clinicians and researchers, as they provide a clearer understanding of the underlying mechanisms driving group-specific behaviors, ultimately facilitating the development of more effective interventions.

After thoroughly validating the clustering results in Chapters 4, 5 and 6, the next step is to explore how information from similar individuals, even without being derived from clustering, can complement and enhance personalized models. Specifically, we propose employing transfer learning approaches to improve predictions for a specific individual (target domain) by incorporating data from one or several other individuals (source domain). In Chapter 7, we focus on methodologically refining all the modeling aspects of the Transfer Adaptive Boosting (TrAdaBoost) process. After a set of experiments investigating the impact of the optimal selection of similar source domains and the target and source reweighting strategies, the results highlight the presence of difficult and useful source instances, but reveal that not all source data significantly contribute to target prediction. Furthermore, the incorporation of similar source domains, optimally individuals of the same cluster as identified through clustering, also positively impacts overall performance, reaching a maximum improvement of 10.7% in AUC score compared to personalized AdaBoost.

Finally, Chapter 8 offers a detailed summary of the research conducted in this dissertation, addressing each of the research questions

posed in Chapter 1 and outlining potential directions for future research. The findings of this dissertation demonstrate the potential of advanced non-linear methods in capturing the complex dynamics of psychopathology. Moreover, personalized models can be enhanced by incorporating the data of more individuals through nomothetic and clustering-based approaches, providing complementary insights, particularly when data is sparse and limited. To balance these approaches, adapting transfer learning approaches further enhances the predictive performance of group-based models, highlighting the value of utilizing shared knowledge across similar individuals. Overall, this dissertation introduces methodological advancements for studying mental disorders, paving the way for deeper insights into the mechanisms underlying psychopathology.

SAMENVATTING

Het doel van dit proefschrift is het ontwikkelen en evalueren van geavanceerde analysemethoden die passen bij de netwerkmethode voor het bestuderen van psychopathologie. Deze methode biedt een nieuw perspectief op het begrijpen van psychische stoornissen. De netwerkmethode conceptualiseert psychische stoornissen als een complex dynamisch systeem van interacterende variabelen, , zoals emoties, gedrag en ervaringen. Om dit dynamische systeem te bestuderen, wordt in dit proefschrift gebruik gemaakt van, herhaalde intra-individuele metingen via Ecological Momentary Assessment (EMA), een methode die de afgelopen tien jaar aanzienlijk is gegroeid binnen de psychologie en gezondheidswetenschappen. EMA houdt in dat proefpersonen via hun smartphone meerdere keren per dag korte vragenlijstjes invullen, wat realtime en contextuele monitoring van verschillende psychopathologie-gerelateerde variabelen, binnen de natuurlijke omgeving van deelnemers mogelijk maakt. Deze data bieden inzichten in het dynamisch verloop van relevante variabelen binnen een proefpersoon.

De rijke structuur van EMA-data kan bijzonder waardevol zijn voor het ontwikkelen van robuuste voorspellende modellen die in staat zijn het verloop van psychische stoornissen te voorspellen, behandelreacties te analyseren of op maat gemaakte psychologische interventies te ontwikkelen. Nauwkeurige voorspellende modellen zijn niet alleen cruciaal voor het faciliteren van vroege interventies—die mogelijk de ernst van psychische episoden kunnen verminderen—maar ook voor het bieden van betrouwbare representaties van de interacties tussen relevante variabelen. Deze representaties verrijken ons begrip van psychische stoornissen en bieden een pad naar meer gepersonaliseerde en effectieve therapeutische strategieën.

Zoals geïntroduceerd in Hoofdstuk 2, is het Vector Autoregressive (VAR) model een veelgebruikte methode die past in de netwerkmethode van psychopathologie. In een VAR model worden de relaties tussen de variabelen lineair gemodelleerd om temporele afhankelijkheden en interacties tussen verschillende psychopathologie-gerelateerde variabelen te beschrijven. Hoewel VAR in dit veld veelvuldig wordt gebruikt, heeft het lineaire karakter van VAR beperkingen voor de beschrijving van de complexiteit van psychische stoornissen. Psychopathologie-gerelateerde variabelen vertonen waarschijnlijk niet-lineaire interacties, wat de dynamische en complexe aard van psychische processen weerspiegelt. Een exclusieve focus op lineaire modellen kan deze relaties oversimplificeren en cruciale aspecten van onderliggende patronen missen.

Gezien de beperkingen van lineaire modellen, onderzoeken we in

dit proefschrift het gebruik van geavanceerde machine learning technieken, met een specifieke nadruk op niet-lineaire modellen. Dergelijke geavanceerde modellen bieden de mogelijkheid om complexere patronen binnen EMA-data te ontdekken, waardoor de nauwkeurigheid van voorspellingen met betrekking tot het voorkomen en de intensiteit van verschillende variabelen wordt verbeterd. In het eerste deel van dit proefschrift ligt de focus ligt op het ontwikkelen van gepersonaliseerde niet-lineaire modellen die voorspellingen baseren op individuele datasets. Daarnaast toetsen we in dit proefschrift ook de integratie van data van andere individuen uit dezelfde EMA-studie om de voorspellende prestaties verder te verbeteren. Door deze groepsinzichten te combineren met gepersonaliseerde benaderingen, bereiken de modellen een balans tussen individuele specificiteit en het extraheren van generaliseerbare patronen, waardoor voorspellend modelleren in de psychopathologie wordt verbeterd.

Gezien de significante interindividuele heterogeniteit, waarbij individuen unieke EMA-patronen vertonen, begint Hoofdstuk 3 met het onderzoeken van de prestaties van niet-lineaire modellen op basis van boomgebaseerde ensemblemodellen. In dit hoofdstuk toetsen we het gebruik van Explainable Boosting Machines (EBM), een niet-lineair en interpreteerbaar model. De evaluatie toonde aan dat de gepersonaliseerde (of idiografische) methode sterke consistentie in resultaten vertoonde , waarbij niet-lineaire modellen de prestaties significant verbeterden op zowel synthetische datasets, alsook echte datasets.

Een van de grootste uitdagingen bij het bouwen van gepersonaliseerde modellen is de beperkte hoeveelheid beschikbare gegevens per individu. Deze kleine datasets leiden vaak tot modellen die niet getraind kunnen worden of, in sommige gevallen, te veel op de gegeven dataset lijken en daardoor hun generaliseerbaarheid verliezen. Om dit probleem aan te pakken, kan data die is verzameld van andere individuen binnen dezelfde EMA-studie nuttig zijn voor het modelleren. In Hoofdstuk 3 onderzoeken we ook deze zogenoemde nomothetische voorspellingsmodellen, die data van alle individuen uit dezelfde EMA-studie integreren. Na een reeks experimenten bleek de voorgestelde Knowledge Distillation method de meest effectieve aanpak voor het verbeteren van de prestaties van gepersonaliseerde modellen. Dit werd bereikt door kennis over te dragen van een groter, algemener model naar een kleiner, meer gespecialiseerd model, wat resulteerde in een verbetering tot 17% in AUC-prestaties met een echte dataset.

Hoewel nomothetische benaderingen met data-integratie waardevolle inzichten bieden in algemene populatietrends, kan de aanzienlijke individuele variabiliteit binnen grote datasets soms de unieke individuele patronen verbergen. Door individuen met gedeelde kenmerken te groeperen, kunnen we ruis in de data verminderen en ons richten op specifiekere patronen binnen subgroepen van de populatie. Een effectieve manier om dit te bereiken is via clustering. Daarom richten Hoofdstukken 4 en 5 zich op het evalueren van de prestaties van verschil-

lende clusteringmethodes voor het groeperen van individuen op basis van de gelijkenis van hun ruwe tijdreeksgegevens en modelgebaseerde informatie. Omdat clustering een niet-gesuperviseerde taak is, waarbij de ware onderliggende groepen meestal onbekend zijn, kan de evaluatie van de resultaten een uitdaging zijn. Daarom onderzoekt Hoofdstuk 4 verschillende clusteringmethoden en clusteringgerelateerde parameters door data uit simulaties te analyseren. De simulaties zijn ontworpen om echte EMA-datasets na te bootsen, met meerdere individuen, ruisgevoelige kenmerken en/of onregelmatige tijdreeksdata. De resultaten suggereren dat het gebruik van alternatieve datarepresentaties, zoals Global Alignment Kernel (GAK)-transformaties , veelbelovend is voor het beter nauwkeuriger vastleggen van de unieke kenmerken en onderliggende patronen van EMA-data.

Naast traditionele evaluatiemethoden worden in Hoofdstukken 5 en 6 verschillende methodes gebruikt om de effectiviteit van clusteringmethoden en de praktische bruikbaarheid van de gecreëerde groepen te beoordelen. Hoofdstuk 5 evalueert clustering binnen de context van een voorspellende taak en toont aan dat verbeteringen in clustering niet willekeurig zijn, maar worden beïnvloed door de kwaliteit van de clusters. Het prestatie-geoptimaliseerde clusteringmodel (POC) verbeterde de Mean Squared Error (MSE)-score met maximaal 7.99% ten opzichte van gepersonaliseerde modellen met Random Forest. Hoofdstuk 6 richt zich vervolgens op verklaarbaarheid als aanvullend evaluatiecriterium, waarbij interpreteerbare verklaringen en aandacht-gebaseerde (attention-based) inzichten worden gegenereerd om de patronen binnen clusters te begrijpen.

Na een grondige validatie van de clusteringresultaten in Hoofdstukken 4, 5 en 6, is de volgende stap om te onderzoeken hoe informatie van vergelijkbare individuen, zelfs wanneer deze niet rechtstreeks voortkomt uit clustering, gepersonaliseerde modellen kan aanvullen en verbeteren. Specifiek toetsen we de toepassing van transfer learning-methodes om de voorspellingen voor een specifiek individu (doeldomein) te verbeteren door gegevens van één of meerdere andere individuen (brondomein) te integreren. In Hoofdstuk 7 richten we ons op de methodologische verfijning van alle modelleeraspecten van het Transfer Adaptive Boosting (TrAdaBoost)-proces. Na een reeks experimenten waarin de invloed van de optimale selectie van vergelijkbare brondomeinen en de herweging van doel- en brondomeinen werd onderzocht, geven de resultaten aan dat, hoewel er zowel moeilijke als nuttige bronvoorbeelden in de dataset aanwezig zijn, niet alle brondata significant bijdragen aan de voorspellingen van het doeldomein.. Bovendien heeft de integratie van vergelijkbare brondomeinen—bij voorkeur individuen uit dezelfde cluster—ook een positieve invloed op de algehele prestaties, met een maximale verbetering van 10.7% in AUC-score ten opzichte van gepersonaliseerde AdaBoost.

Tot slot biedt Hoofdstuk 8 een uitgebreide samenvatting van het onderzoek in dit proefschrift, waarin de gestelde onderzoeksvragen uit Hoofd-

stuk 1 worden beantwoord en mogelijke richtingen voor toekomstig onderzoek worden besproken. De bevindingen in dit proefschrift onderstrepen het potentieel van geavanceerde niet-lineaire modellen om de complexe dynamiek van psychopathologie beter te begrijpen. Bovendien blijkt dat gepersonaliseerde modellen aanzienlijk kunnen worden verbeterd door gegevens van andere individuen te integreren via nomothetische en clustering-gebaseerde benaderingen. Dit biedt waardevolle aanvullende inzichten, vooral wanneer de beschikbare gegevens per individu schaars of beperkt zijn. Door transfer learning toe te passen, wordt een balans gevonden tussen individuele specificiteit en het benutten van gedeelde kennis, wat leidt tot robuustere en nauwkeurigere voorspellende modellen. Samenvattend introduceert dit proefschrift methodologische innovaties voor het bestuderen van psychische stoornissen, waarmee een basis wordt gelegd voor diepgaandere inzichten in de onderliggende mechanismen van psychopathologie en de ontwikkeling van effectievere, op maat gemaakte interventies.

IMPACT PARAGRAPH

The Regulations for obtaining a doctoral degree at Maastricht University require the addition of an impact paragraph to the thesis (Article 12, Paragraph 8, entry into force on 1 February 2023). This paragraph consists of a reflection on the scientific impact of the results of the research described in the thesis, as well as, if applicable, the social impact anticipated or already achieved. Scientific impact is the short-term and long-term contribution of the results of scientific research to shifting insight and stimulating science, method, theory and application within and between disciplines. Social impact is the short-term and long-term contribution of the results of scientific research to changes in or development of social sectors and to social challenges. This impact paragraph addresses the four questions provided in the regulations.

Research

What is the main purpose of the research described in the thesis, and what are the main results and conclusions?

The primary goal of this dissertation is to develop reliable and robust approaches, applying advanced data-driven models to Ecological Momentary Assessment (EMA) data, to accurately capture individual psychopathology. Starting from personalized (idiographic) and group-based (nomothetic) approaches, this research work places significant emphasis on balancing these perspectives through more sophisticated group-based modeling strategies.

Building on the challenges and assumptions of the widely employed linear models, this research aims to bridge the gap between traditional linear models and more advanced non-linear models within the network approach to psychopathology. Specifically, it demonstrates how adopting non-linear models, such as Explainable Boosting Machines (EBMs), can be effectively integrated into this framework to uncover complex and hidden relationships between variables. Non-linear models also provide a more realistic understanding of mental health processes, enabling their application in more advanced and sophisticated modeling approaches to further enhance predictive accuracy and clinical utility.

A key focus throughout this dissertation is on improving the predictive performance of personalized models, not only to achieve higher accuracy, but also to ensure that models can provide reliable and consistent predictions across various scenarios or predictive tasks. These tasks include forecasting different psychopathology-related variables, capturing

dynamic interactions, and explaining the underlying psychopathological mechanisms at an individual and group level. By addressing these objectives, the dissertation aims to enhance our understanding of mental health dynamics, support the development of personalized interventions, and contribute to more generalizable methodologies that can be applied across diverse populations and contexts.

Specifically, this dissertation (Chapter 3) starts with exploring the application of non-linear EBM models within both idiographic and nomothetic approaches. Experimental results demonstrate that non-linear models, including EBMs, outperform linear models in terms of predictive accuracy. Moreover, nomothetic modeling approaches based on EBMs show improved performance, highlighting their potential to more accurately predict future outcomes.

To further improve nomothetic approaches, this dissertation (Chapters 4 and 5) introduces clustering techniques to group individuals with similar patterns, reducing noise and variability in EMA datasets. Clustering-derived group models demonstrate superior performance compared to personalized and traditional nomothetic approaches that utilize all data. The results show that such a strong performance is not merely a random outcome of combining multiple models but rather is driven by the quality and relevance of the clusters themselves. The discovery of such meaningful subgroups within the population is further explored (Chapter 6), which introduces methods for explaining clustering results. Specifically, deep learning attention models were utilized, aiming to provide clearer insights into the structures and key factors that differentiate the clusters. This method is model-agnostic, making it applicable across various clustering algorithms. Furthermore, it serves as a general clustering evaluation measure, enabling the assessment of any clustering result by uncovering the patterns and features that define each cluster.

Inspired by the success of incorporating data from similar individuals into group models, the subsequent goal is to balance idiographic and nomothetic approaches by placing greater emphasis on personalized data. Specifically, this dissertation (Chapter 7) explores the application and refinement of transfer learning methods, such as Transfer AdaBoost, for EMA data, making a significant contribution to the field of predictive modeling in psychology. While the average performance of models incorporating group data is comparable to that of personalized models, further analysis of individual-level changes reveals notable improvements for several individuals. This highlights the value of leveraging relevant data from additional individuals to enhance personalized predictions. This approach is not only impactful within the context of EMA data but also generalizable to other domains where data is sparse or limited at the individual level, helping to enhance the predictive accuracy across various scientific disciplines, such as healthcare, behavioral and social science.

Overall, in this dissertation, interdisciplinary collaboration between psychology, data science, and machine learning played a pivotal role.

By integrating expertise from these domains, the models developed were better aligned with the complexity of mental disorders and practical applications, such as personalized interventions and treatments. A critical component of this collaboration was ensuring interpretability, where domain experts and data scientists worked closely to clarify the goals of the analysis, what insights were most valuable, and how the findings could be applied in practice. This reciprocal communication allowed the models to balance the theoretical needs of psychological research with the technical possibilities and limitations of advanced data analysis, ultimately ensuring meaningful and actionable results.

Relevance

What is the (potential) contribution of the results of this research to science, and if applicable to societal sectors and societal challenges?

This research contributes not only to the scientific understanding of mental health modeling but also holds the potential to address key societal challenges in mental healthcare. With mental health issues on the rise globally, considering the challenges in proper diagnosis and treatment that lead to an increasing burden on healthcare systems, the research of this dissertation focuses on modeling and predicting individual psychopathology through EMA data, offering promising pathways for improved healthcare solutions.

By improving predictive models of psychopathology, this research primarily enhances our ability to better understand the complex hidden relationships between psychopathology-related variables, such as emotions, behaviors, and experiences. Specifically, using advanced nonlinear models, this research provides realistic and flexible representations of the underlying interactions, allowing for more precise identification of these dynamics. Moreover, clustering techniques provide valuable insights into individual mental health profiles, revealing how people both differ and share common patterns. This knowledge can refine our understanding of diverse mental health trajectories and help identify meaningful groupings within the population.

These enhanced models and enriched knowledge representation have the potential for more precise and tailored treatments. A deeper understanding of individual and group profiles enables clinicians and mental health professionals to develop interventions that are better suited to the specific needs by relying on the characteristics of individuals or clusters. This personalization can lead to more effective and timely interventions, ultimately improving patient outcomes and transforming how mental health care can be provided.

In addition, the proposed nomothetic models reduce the need for extensive individual data collection, by leveraging aggregated data from multiple individuals, which has important social implications. While these

models initially require sufficient data across a population to be effective, they mitigate the need for collecting large datasets from every individual. Through shared information, they enable the generalization of findings across populations, making the models more accessible and practical for broader applications. This approach is particularly valuable in resource-limited settings where data collection can be burdensome or costly. Ultimately, it contributes to bridging the gap in access to mental health services in regions where healthcare might be constrained due to financial, geographical, or infrastructural challenges.

Target Audience

To whom are the research findings interesting and/or relevant? And why?

This dissertation is designed to impact research in the fields of psychopathology and mental health, as well as data science, jointly advancing personalized medicine.

First, targeting researchers in psychopathology, particularly those working on methodological and modeling advancements, this dissertation aims to bridge the gap between traditional linear models and more advanced machine learning models. Particularly, Chapter 2 starts with the limitations and assumptions of the current linear methods used to model psychopathology, highlighting the complicated structure of EMA data and the complexity of mental disorders. An important part of this chapter is the connection between network linear approaches and more advanced non-linear models. It is important to highlight that adopting non-linear models can uncover the hidden relationships between variables, similar to linear models, but in a more realistic and complex way.

Additionally, clinical psychologists and scientists in mental healthcare can benefit from the findings, as the improved models offer practical insights for a better understanding of the individual dynamics, and interactions between variables, as well as the shared characteristics or profiles and variability within each group of individuals. This knowledge can help clinical psychologists design more personalized and effective interventions and treatment strategies, making them more responsive to individual needs. Particularly, within the context of the New Science of Mental Disorders (NSMD[1]) project, these models provide a valuable framework for studying mental health disorders, enabling more accurate predictions that could ultimately help to tailor interventions to the needs of diverse patient populations. Moreover, Chapter 6 specifically focuses on explaining the clustering techniques in a way that is accessible to domain experts. By making the clustering results interpretable, the research ensures that clinical psychologists can understand and use these findings to inform treatment decisions or design tailored interventions.

[1] **https://www.nsmd.eu/**

This emphasis on interpretability bridges the gap between data science methodologies and clinical practice, enabling their collaboration.

Furthermore, this work is highly relevant to data scientists and researchers working with EMA and time-series data, as it introduces novel approaches to handle challenges inherent to complex and dynamic data sets. Particularly, Chapters 3-7 present advanced methodologies for modeling and analyzing time-series data, exploring key challenges such as time-series similarity, time-series clustering evaluations and explanations. These methods, while developed in the context of studying psychopathology, have broad applicability across other fields that rely on time-series data, such as personalized healthcare monitoring, climate science, biology and finance. Therefore, by providing adaptable approaches, this research extends its impact beyond psychology, offering valuable tools for a wide range of disciplines dealing with sequential data.

Activity

In what way can these target groups be involved and informed about the research findings so that the knowledge gained can be used in the future?

The majority of the content chapters of this dissertation have already been described in various publications. In particular, Chapters 3 to 7 present studies that have been presented at five international conferences and published in their associated peer-reviewed conference proceedings. Chapter 4 was also partially published in the Elsevier journal Machine Learning with Applications. Further related works were also presented at the International Convention of Psychological Science (ICPS) in 2023 and at the IEEE 40th International Conference on Data Engineering Workshops (ICDEW) in 2024. Moreover, an interactive web application has been developed alongside this research to visually explore the EMA data and model outcomes, allowing users to interact with and better understand the data and predictions derived from the models presented in this dissertation. The web application can be accessed at: **`https://clustering-pilot-data.streamlit.app/`**.

CURRICULUM VITæ

Mandani Ntekouli

Mado Ntekouli was born in Athens, Greece, on December 1, 1993. She graduated in 2016 from the University of Patras (Greece) with a Diploma in Electrical and Computer Engineering, completing a thesis titled “An Implementation of a Decision-making Model for the Current Health Status of Patients Diagnosed with a Chronic Disorder - Application to Epilepsy”. The research thesis was conducted under the supervision of Professor Dr. Dimitrios Lymperopoulos. After her studies, she worked as a research assistant at the Wire Communication Laboratory of the same university, gaining valuable experience in decision support systems.

In 2017, Mado moved to the Netherlands to pursue her MSc in Biomedical Engineering at Delft University of Technology (TU Delft). During her studies, she conducted a research internship at the Neuroscience Department of Erasmus Medical Center in Rotterdam, where she worked on techniques for visualizing brain activity and developed computational methods for processing this data. Building on the same topic, she completed her thesis: “Investigating brain function and anatomy through ICA-based functional ultrasound imaging”, under the supervision of Dr. Borbala Hunyadi, Dr. Ir. Pieter Kruizinga and Dr. Ir. Christos Strydis.

In 2020, Mado began her PhD at the Department of Advanced Computing Sciences (DACS) at Maastricht University, under the supervision of Dr. Jerry Spanakis and Professor Dr. Anne Roefs. Mado’s PhD project is part of the 10-year NWO Gravitation program, New Science of Mental Disorders (NSMD), which aims to fundamentally advance our understanding of mental health disorders. In this project, the focus of her research is on developing innovative machine learning models to analyze Ecological Momentary Assessment (EMA) data, aiming to enhance predictions in psychopathology. Her work also delves into the integration of clustering techniques, predictive modeling, and explainable AI for clinical use. Broadly, her research interests include machine learning, time-series analysis, personalized medicine, and transfer learning.

During Mado’s PhD journey, she was a member of the Cognitive Systems Research Group of DACS and the FSE STEM Graduate School. At a national level, she was also affiliated with the Netherlands Research School for Information and Knowledge System (SIKS).

LIST OF PUBLICATIONS

7. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Enhanced Boosting-based Transfer Learning for Modeling Ecological Momentary Assessment Data”. In: ML4ITS2023 - 3rd Workshop on Machine Learning for Irregular Time Series: Advances in Generative Models, Global Models and Self-Supervised Learning. ECML. 2024

6. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention”. In: Explainable Artificial Intelligence. Ed. by L. Longo, S. Lapuschkin, and C. Seifert. Cham: Springer Nature Switzerland, 2024, pp. 75–99. isbn: 978-3-031-63797-1

5. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Exploiting Individual Graph Structures to Enhance Ecological Momentary Assessment (EMA) Forecasting”. In: 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW). IEEE. 2024, pp. 158–166

4. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Model-based Clustering of Individuals’ Ecological Momentary Assessment Time-series Data for Improving Forecasting Performance”. In: BNAIC/ BeNeLearn 2023: Joint International Scientific Conferences on AI and Machine Learning. 2023

3. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Evaluating multivariate time-series clustering using simulated ecological momentary assessment data”. In: *Machine Learning with Applications* 14 (2023), p. 100512

2. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Clustering individuals based on multivariate EMA time-series data”. In: The Annual Meeting of the Psychometric Society. Springer. 2022, pp. 161–171

1. M. Ntekouli, G. Spanakis, L. Waldorp, and A. Roefs. “Using explainable boosting machine to compare idiographic and nomothetic approaches for ecological momentary assessment data”. In: International Symposium on Intelligent Data Analysis. Springer. 2022, pp. 199–211

SIKS DISSERATATION SERIES

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

	40	Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
	41	Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
	42	Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
	43	Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
	44	Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
	45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
	46	Jan Schneider (OU), Sensor-based Learning Support
	47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
	48	Angel Suarez (OU), Collaborative inquiry-based learning
2018	01	Han van der Aa (VUA), Comparing and Aligning Process Representations
	02	Felix Mannhardt (TU/e), Multi-perspective Process Mining
	03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
	04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
	05	Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
	06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
	07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
	08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
	09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
	10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
	11	Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
	12	Xixi Lu (TU/e), Using behavioral context in process mining
	13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
	14	Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
	15	Naser Davarzani (UM), Biomarker discovery in heart failure
	16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
	17	Jianpeng Zhang (TU/e), On Graph Sample Clustering
	18	Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

	36	Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

	13	Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
	14	Selma Čaušević (TUD), Energy resilience through self-organization
	15	Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
	16	Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
	17	Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
	18	Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
	19	George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
	20	Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
	21	Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
	22	Alireza Shojaifar (UU), Volitional Cybersecurity
	23	Theo Theunissen (UU), Documentation in Continuous Software Development
	24	Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
	25	Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
	26	Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
	27	Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
	28	Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
	29	Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
2024	01	Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
	02	Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
	03	Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
	04	Mike Huisman (UL), Understanding Deep Meta-Learning
	05	Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025
01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)