UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention

Ntekouli, Mandani; Spanakis, Gerasimos; Waldorp, Lourens; Roefs, Anne

DOI
10.1007/978-3-031-63797-1_5

Publication date
2024

Document Version
Final published version

Published in
Explainable Artificial Intelligence

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Ntekouli, M., Spanakis, G., Waldorp, L., & Roefs, A. (2024). Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Explainable Artificial Intelligence: Second World Conference, xAI 2024, Valletta, Malta, July 17–19, 2024 : proceedings* (Vol. II, pp. 75-99). (Communications in Computer and Information Science; Vol. 2154). Springer. https://doi.org/10.1007/978-3-031-63797-1_5

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Nov 2026

Explaining Clustering of Ecological Momentary Assessment Data Through Temporal and Feature Attention

Mandani Ntekouli[1(✉)], Gerasimos Spanakis[1], Lourens Waldorp[2], and Anne Roefs[3]

[1] Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands
{m.ntekouli,jerry.spanakis}@maastrichtuniversity.nl
[2] Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
L.J.Waldorp@uva.nl
[3] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
a.roefs@maastrichtuniversity.nl

Abstract. In the field of psychopathology, Ecological Momentary Assessment (EMA) studies offer rich individual data on psychopathology-relevant variables (e.g., affect, behavior, etc.) in real-time . EMA data is collected dynamically, represented as complex multivariate time series (MTS). Such information is crucial for a better understanding of mental disorders at the individual- and group-level. More specifically, clustering individuals in EMA data facilitates uncovering and studying the commonalities as well as variations of groups in the population. Nevertheless, since clustering is an unsupervised task and true EMA grouping is not commonly available, the evaluation of clustering is quite challenging. An important aspect of evaluation is clustering explainability. Thus, this paper 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. To evaluate the proposed approach, an EMA dataset of 187 individuals grouped in 3 clusters is used for analyzing the derived attention-based importance attributes. More specifically, this analysis provides the distinct characteristics at the cluster-, feature- and individual level. Such clustering explanations could be beneficial for generalizing existing concepts of mental disorders, discovering new insights, and even enhancing our knowledge at an individual level.

Keywords: Clustering · Multivariate Timeseries · EMA · Explainability · Temporal Attention · Feature Attention

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
L. Longo et al. (Eds.): xAI 2024, CCIS 2154, pp. 75–99, 2024.
https://doi.org/10.1007/978-3-031-63797-1_5

1 Introduction

For quite some years now, Ecological Momentary Assessment (EMA) [2,24] has been a popular method in psychology research to collect data in daily life. During an EMA study, individual real-time data is collected over time, capturing various psychopathology-related behaviors, experiences, and symptoms in their natural environments [7]. For each individual, EMA data is then organized as a multivariate time series (MTS). Such a rich structure sets new pathways to employing advanced analytical techniques, enhancing our understanding of the complex patterns in psychopathology.

According to psychopathology theory, personalized models have been traditionally applied. A popular model for EMA data is the vector Autoregressive (VAR) model, where one or more time lags are used to predict variables at subsequent time-points. Because of heterogeneity between individuals, these models focusing on each individual separately allow for interventions and treatments that are tailored to their unique patterns of symptoms [20]. Although, personalized models can reveal quite distinct patterns, they may also reveal commonalities. Such commonalities can be quite valuable, giving insights into more general mechanisms that are valid for particular subgroups. It is, however, quite difficult to ascertain if there are such subgroups for which commonalities hold and also how to discover them.

In an EMA study, collecting data for multiple individuals at the same time could facilitate the understanding of mental disorders by studying the commonalities as well as variations present in the sample. An interesting approach for broadening EMA analysis involves enhancing models with data from more than one individual [15]. More specifically, group-based models, relying on individuals exhibiting similar temporal patterns of symptomatology or behavior, can offer potentially valuable insights for the whole sample [17]. Thus, uncovering homogeneous groups of people is of great importance. Such groups of individuals can be uncovered by clustering applied to EMA data [16]. Nevertheless, since clustering is an unsupervised task and true EMA grouping is not commonly available, the evaluation of the derived clustering labels is quite challenging. Because of the inherent variability of EMA data as well as the complexity of the real-world MTS, it is expected that 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, 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 inter-relationships among theoretically similar subtypes of disorders.

In this paper, to address the need for explanations on clustering results, the proposed methodology focuses on investigating interpretable deep-learning

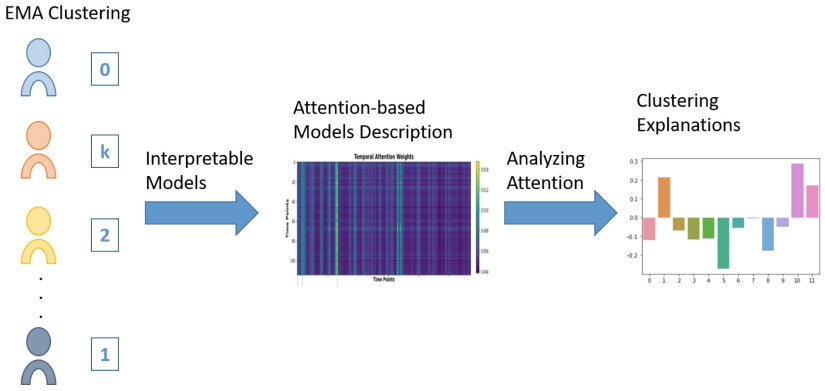


Fig. 1. An overview of our methodological approach focusing on applying attention-based interpretable models to provide clustering explanations.

mechanisms capable of handling MTS data, particularly within the domain of psychopathology. Our approach utilizes advanced interpretative deep-learning models, specifically attention-based mechanisms [25], 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 Fig. 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 [8,22]. While the proposed framework for extracting temporal and feature-level attention is fully described, a significant part of this paper is to examine 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 be proven beneficial for generalizing the existing concepts, uncovering 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.

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 representation for clusters' descriptions as well as the classical explanation methods of the field of Explainable Artificial intelligence (XAI).

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 elements 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 be also 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.

2.2 Explanations on TS Clustering

Beyond statistically analyzing and exploring the structures of each cluster, the majority of the work focuses on extracting the factors (e.g. important variables) that influence the cluster assignments by applying interpretable models. The methodologies employed for explaining time-series clustering vary considerably, despite the low number of publication outputs.

Following the approach of transforming the data to different time-series representations, the work of [4] applies clustering on various interpretable extracted features, including intra-signal (or within feature) and inter-signal (across different features) characteristics, such as variance and correlation, respectively. After data pre-processing 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, like [10,18] was inspired by the general trend of training and explaining classification models to predict the derived cluster labels. On the one hand, according to [18], local interpretability methods, such as Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP) as well as Gradient-weighted Class Activation Mapping, are used with different classification models (e.g. XGBoost) and trained again on statistically-extracted temporal features. Examples of such features were autocorrelation and median difference. On the other hand, the work of [10] 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.

3 Review on Challenges of Explaining MTS Data

Due to the inherent challenges of the topic and limited available published work, the current paper was also inspired by the literature focusing on using interpretable classification models for explaining an output vector, where the classes are the distinct cluster labels. In our setting, that output can be the clustering-derived result. Thus, an additional review of the challenges of applying interpretable classification models on MTS is necessary.

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 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 commonly used, involving an additional classification model trying to predict and explain the clustering labels. For instance, in [1,14], 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 targets.

Regarding classification explainability, beyond using inherently interpretable models, such as linear or decision tree models, post-hoc explainability methods are also commonly explored. These 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 [23]. 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 [19], and SHAP for feature-based explanations developed by [12].

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 were mostly needed, splitting and analyzing data into valuable sub-sequences.

More specifically, LIME and SHAP extensions were necessary to be further introduced, by leveraging the strong foundations of these methods adjusted to address such directions. For instance, TS-MULE is a LIME-based method not limited to uniform segmentation using static windows, but incorporating multiple advanced segmentation algorithms, such as matrix profile and SAX [21]. This way, it is more likely that meaningful subsequences are uncovered, leading to motifs 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 targets.

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 [26]. 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 [3]. 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 [4]. After such transformation, a feature-vector representation is retained, which can be easily used in all existing classification models.

Alternatively, the necessary transformations can come internally through models incorporating data representations and prediction. Recently, neural networks capable of handling multivariate time-series data have been used increasingly. More specifically, recurrent neural networks- (RNN) -based 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 been also introduced to sequential modeling with the ability to uncover and highlight the most important parts or periods of sequences [9,25]. 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.

According to the attributes of attention-based models, our two challenges, 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.

4 Framework for Clustering Explanations

In this section, first, the structure of the examined EMA data is described along with the problem statement of our work. Specifically, the goal of providing explanations on clustering results applied to EMA MTS data is presented in detail. Then, the focus will be on the proposed architecture that utilizes an interpretable attention-based framework and how this can eventually lead to clustering explanations.

4.1 EMA Data

In the current study, EMA was used as a dynamic approach to collect data on individuals' psychopathology-related behaviors, experiences, and symptoms in real-time and in their natural environments. More specifically, through digital questionnaires on their smartphones, participants were prompted to answer sets of questions. Most questions were answered on 7-point Likert scales; after normalization at an individual level, these data can be considered continuous. As all this data was collected multiple times throughout the day, EMA is typically

structured as MTS data. In this study, the dataset examined consists of real-world EMA pilot data, collected from students of Dutch Universities [13,20]. More specifically, 187 participants were asked to complete 8 questionnaires per day at fixed time intervals for 28 days. According to the domain experts, some variables (e.g., positive affect and negative affect) were averaged across relevant variables (e.g., happy, calm, etc. and sad, angry, etc., respectively), and some variables were not included in the analyses, either due to limited within-person variance or due to relatively less relevance.

The EMA dataset, consisting of the data of $N = 187$ individuals, can be denoted as $X = \{X_1, X_2, .., X_N\}$. For each individual i, X_i is comprised of $V = 12$ variables, each captured for a number of time-points, T_i. The total number of time-points T_i ranged from 112 to 224 and was different across individuals due to differences in compliance. The minimum compliance was set to 50%. Each individual EMA dataset X_i is represented by an MTS dataset, structured as $\{V_1, \ldots, V, T_1, \ldots, T_i\}$.

4.2 EMA Clustering

As already stated, an important task in EMA analysis is clustering individuals based on their MTS data. Clustering entails assigning individuals sharing similar patterns to the same groups, while assigning individuals with different patterns to different groups. This approach enables uncovering distinct profiles among people that might not be immediately apparent. Such information could also provide insights into mechanisms maintaining or causing psychopathology.

A variety of clustering methods can be applied to EMA MTS data, each highlighting different underlying assumptions about the structure of the data [16]. These include traditional clustering techniques, such as k-means and hierarchical clustering, but are adapted to handle the temporal and multivariate nature of EMA data. For instance, clustering methods by incorporating an appropriate distance metric, like dynamic time warping (DTW) and Global Alignment Kernel (gak), are considered capable of advancing the inherent temporal dimensions of the EMA data [5,6]. Alternatively, clustering on EMA data could be approached from a multivariate functional data perspective, by representing the data with smoothed curves. The use of smoothing techniques could possibly facilitate identifying the underlying trends, which might be obscured by noise in the original time series [11]. However, its applicability to EMA data should be explored more carefully.

4.3 Proposed Framework for Clustering Explanations

An overview of the proposed framework of providing explanations on EMA clustering is given in Fig. 2. All components of the framework are described as follows.

Input: EMA Data. The framework begins with the EMA data matrix X as input. X represents the personalized data of $N = 187$ individuals, consisting of $V = 12$ variables and T_i time-points. 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 transformed and filled when necessary, so that all have the same T time-points for further analysis.

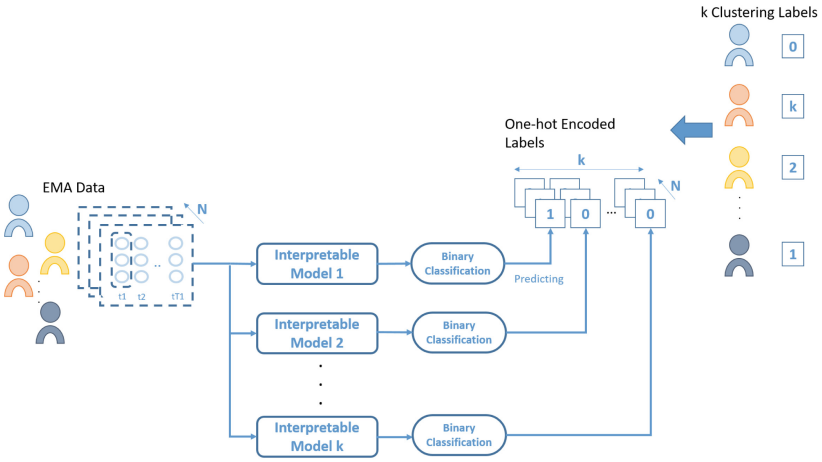


Fig. 2. An overview of the proposed framework for providing explanations on EMA clustering.

Output: Clustering Labels. In this study, we adopt an approach focusing on the explanation of clustering outcomes in EMA data without depending on a particular clustering method. That is, the approach relies on the fact that clustering is performed as a prior step, utilizing only the derived clustering labels in our framework 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.

***k* 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 focuses on predicting one cluster and distinguishing this from the rest of the clusters. For example, regarding the classification Model0, 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

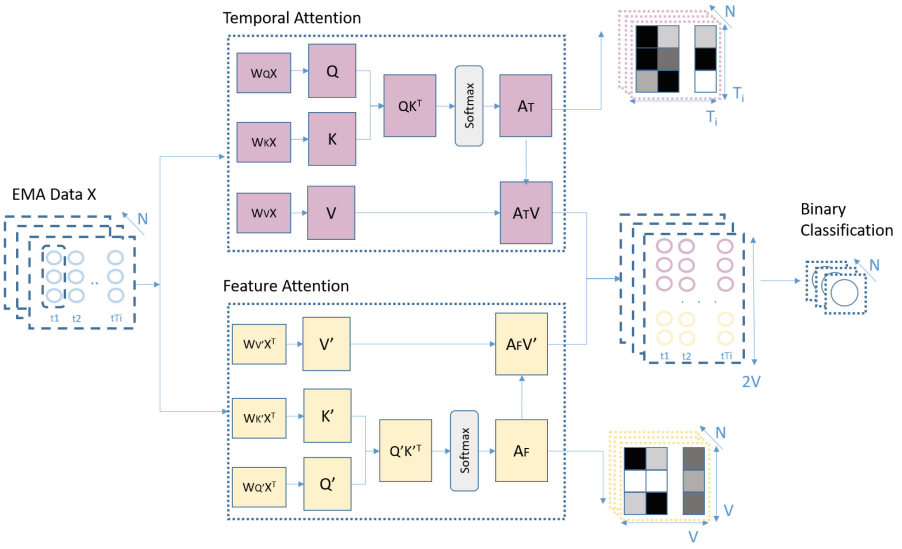


Fig. 3. The main components of each interpretable model.

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 comes into the output. This works by one-hot encoding the clustering labels of the pre-applied clustering and using a different output vector for each model. 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.

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 used are shown in Fig. 3. Each binary classification model is interpretable mainly relying on self-attention mechanisms [25]. By nature, the attention mechanism focuses on the temporal domain of the data, identifying the most important parts of the data with respect 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 [9].

In the current framework, to address the complexity of EMA data, 2-level attention is used in parallel, each focusing on different aspects of the data [8,22]. 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:

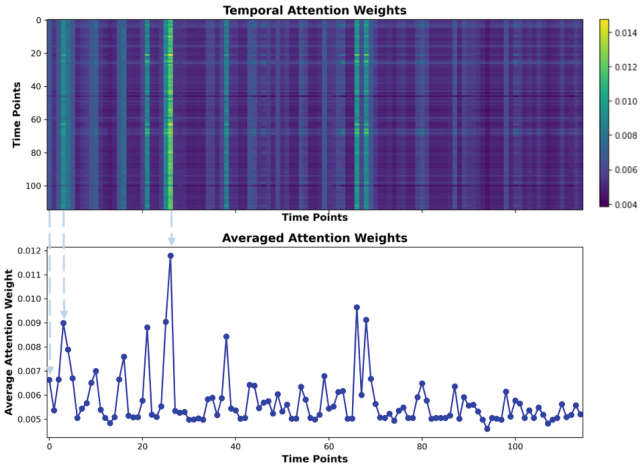


Fig. 4. The averaging process of the full Temporal Attention matrix A_T to $A_{T_{av}}$

– **Temporal Attention**
As shown in Fig. 3, the calculations of the Temporal Attention follow the procedure of a Scaled Dot-Product Attention self-attention block [25]. First, the 3 main components of the attention block, matrices query Q, key K, and value V, are calculated by a linear transformation of the input X data. Then, the 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 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 to 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 of all time-points on each one, 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 noticed in Fig. 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 come after a linear transformation of the transpose of the input 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.

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 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 drove 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 Fig. 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 be also 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 actual 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.

5 Analysis and Results

As already mentioned, in our analysis, a real-world EMA 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, a 3-clustering result derived from a gak kernel k-means was chosen as a good clustering [16]. 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.

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 pre-defined 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 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 1.

Table 1. Comparison of Models Performance

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

This comparison shows the superior performance of the proposed framework over 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. 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.

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 model, Model0. Regarding that 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 Fig. 5.

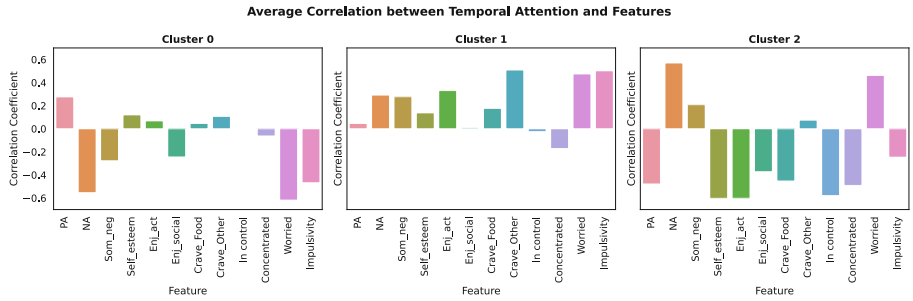


Fig. 5. Cluster-level average correlation effects between the temporal attention weights and the EMA features.

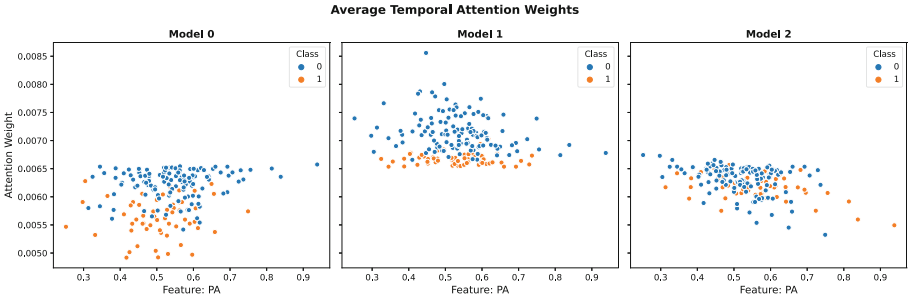
According to Fig. 5, the important effects of features in distinguishing each cluster from the rest are identified. It can easily be seen that the patterns of the correlations of EMA features and temporal attention weights are almost orthogonal for each of the clusters. Furthermore, particular high (low) correlations stand out. In Cluster0, "Negative Effect" (given as NA), "Worried" and "Impulsivity" have strong negative correlations with, 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 stronger 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, it is possible that some individuals' effects deviate from the averaged ones. Thus, explanations at the individual-level should be also 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 actually 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 Fig. 6 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.

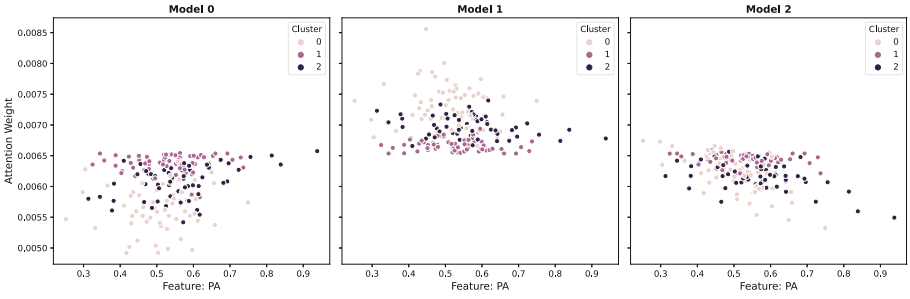
We notice that for Model0 and Model1, individuals belonging to Cluster0 and Cluster1, respectively, get lower (on average) attention scores than the rest. This could possibly reflect the fact that Class1 is always the minority class compared to Class0. Thus, each model gives more attention to the majority class.

Nevertheless, the findings of Model3 are not similarly clear. It is noticeable that the attention weights of the 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.



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



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

Fig. 6. Feature (Positive Affect) Relationship with Temporal Attention weights

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 Fig. 7, 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 a 2d plot, the similarities to Cluster0 are not 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 Cluster2, lower than 0.05, the majority 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. In addition, we can observe individuals of

Cluster1 and Cluster2 exhibiting comparable levels of similarity to more than one cluster, suggesting a degree of ambiguity in their cluster membership. 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.

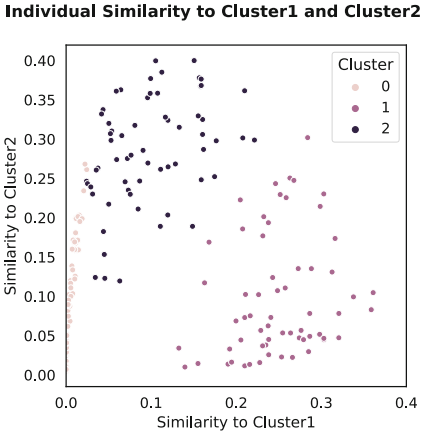


Fig. 7. Similarities of all individuals to Cluster1 and Cluster2.

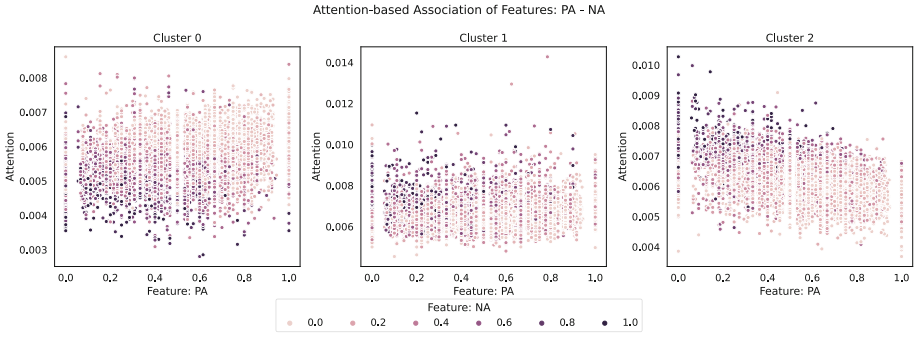


Fig. 8. Features Interaction between "Positive Affect" (PA) and "Negative Affect" (NA) with respect to the Temporal Attention.

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 Fig. 8. For each cluster, each point represents a time-point of the individuals

belonging to that cluster. According to this figure, the 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, 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 characterizing each cluster.

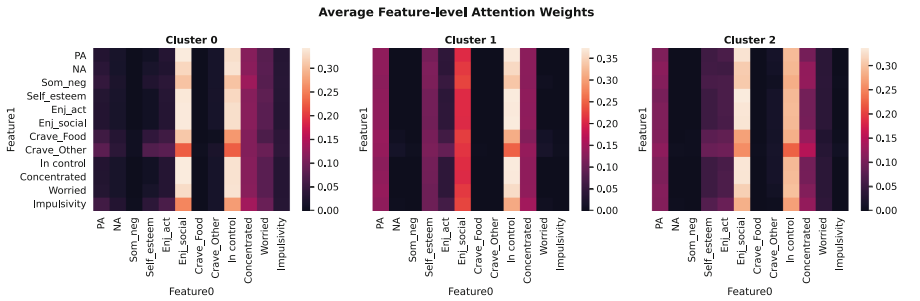


Fig. 9. Cluster-level average feature-level attention weights.

5.3 Cluster-Level Explanations Through Feature-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 Fig. 9. The 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's noticeable that the same features emerge 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 Fig. 10. 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 row, respectively. Each point represents an individual belonging to a particular cluster. As expected from

Fig. 9, 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 for low values of "Crave Food" and high values of "Enjoying Social Activities". By analyzing such 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 of feature interactions could reveal some special characteristics and patterns that are important for distinguishing each cluster.

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 to. The example of the first individual, Individual0, belonging to Cluster2, is used for the rest of the analysis. A detailed summary

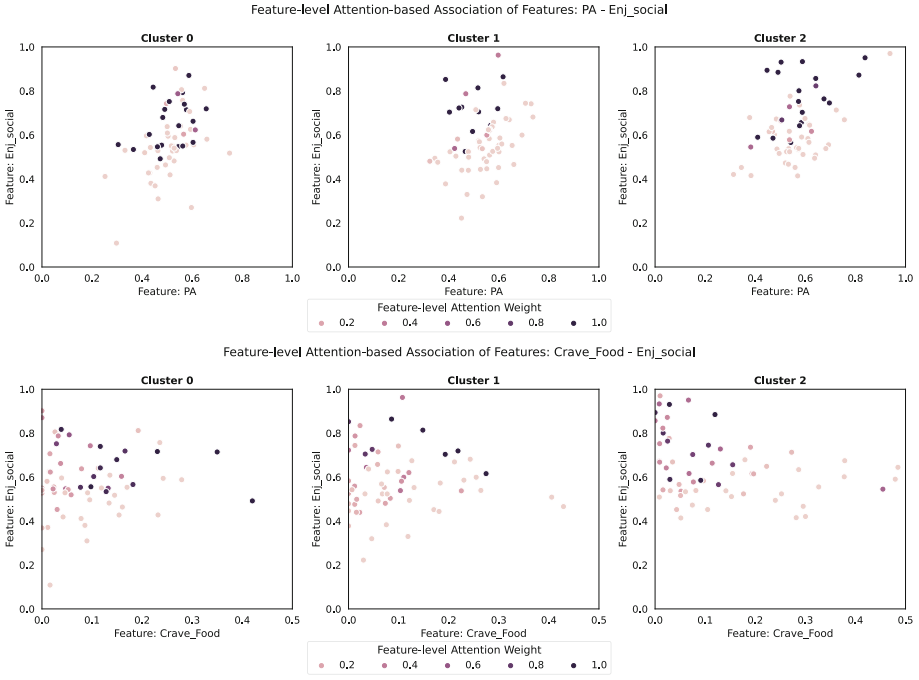


Fig. 10. Feature-level Attention: Unfolding the inter-connection of "Enjoying Social Activities" to two other features: "Positive Affect" and "Crave Food".

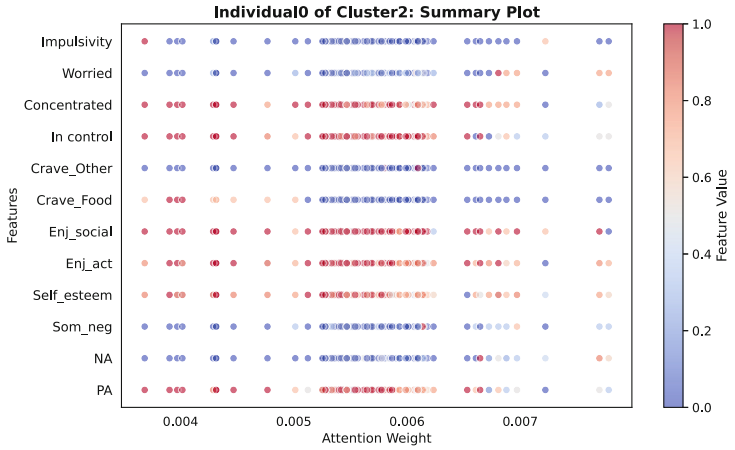


Fig. 11. The summary plot of feature interactions for an individual of Cluster2.

plot including all feature interactions and their attention weights is presented in Fig. 11. Particularly, for each feature on the y-axis, all the time-points of Individual0 are plotted, while ordered according to their attention weights, and colored by the value of each feature. Through this, we can have a thorough 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, 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 Fig. 12. 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 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.

6 Discussion

In this paper, the under-studied 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, both in the temporal and feature-level dimension. Through its multi-aspect design and analysis, 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.

Fig. 12. The attention weights of the interaction between "Positive Affect" and "Enjoying Social" derived from all 3 models.

6.1 The Role of the Multi-aspect Attention

Although attention mechanisms have typically been employed on the time dimension, the current approach of integrating multi-aspect attention mechanisms, focusing on both temporal and feature levels, showed that enhances the performance of the overall model in terms of accuracy. Based on Table 1, this integrated approach outperformed all baseline methodologies, by predicting more individuals in the correct clusters in both training and test sets. Its high performance is indicative of how informative the weights learned by each model are and how these could facilitate distinguishing individuals across clusters.

Moving to interpretability, the multi-aspect attention mechanism offers detailed information on the data patterns that each model considers as important. While temporal attention highlights the significance of particular time-points, 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 like EMA, where the understanding of dynamic patterns is crucial.

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 and feature interactions 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, it is essential to examine the data at the individual level. The individual-level analysis facilitates uncovering important feature interactions and patterns unique to each subject as well as how each person is reflected on the weights learned by models focusing on different clusters.

6.3 The Impact of Utilizing a Pre-defined Clustering Result

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 every possible clustering technique to be potentially prior utilized. Thus, it could be used as an effective tool for evaluating the result given by any clustering method. Such a feature not only enhances the framework's utility but also broadens its applicability, allowing for 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, in the case that clustering labels are not robust, the proposed framework would again provide some explanations on the cluster- and individual-level importance effects, but without these being meaningful or accurate. This may be the case also for the chosen clustering of our EMA dataset. Since the real clusters are not available and the between-cluster similarities are found close to each other, the examined clustering may not be the optimal one. 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, it is important for the clustering results of other clustering algorithms to be investigated. This necessity also comes since in human-centric fields, like EMA studies, developing simulation datasets or objective criteria to evaluate the quality of explanations is quite challenging.

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

7 Conclusion

To summarize, our interpretable framework represents a significant advancement in providing explanations and evaluating an MTS clustering result. By thoroughly 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. This is especially critical in fields like psychopathology, where a better understanding on clustered individuals could be beneficial for personalized interventions and mechanistic understanding.

Regarding future work, the goal is to augment personalized models by using data from similar individuals. Having evaluated various clustering results and uncovered the important data segments of an individual through the proposed framework, incorporating data from similar profiles could significantly improve the predictive performance of personalized models. Such methodology could produce more accurate predictions, enhancing the precision of personalized approaches.

Acknowledgements. 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).

Disclosure of Interest. Authors have no conflict of interest to declare.

References

1. Bandyapadhyay, S., Fomin, F.V., Golovach, P.A., Lochet, W., Purohit, N., Simonov, K.: How to find a good explanation for clustering? Artif. Intell. **322**, 103948 (2023)
2. Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., Ruwaard, J.: How to predict mood? Delving into features of smartphone-based data. In: Twenty-Second Americas Conference on Information Systems, San Diego (2016)
3. Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A., Bizarro, P.: TimeSHAP: explaining recurrent models through sequence perturbations. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2565–2573 (2021)
4. Bonifati, A., Buono, F.D., Guerra, F., Tiano, D.: Time2Feat: learning interpretable representations for multivariate time series clustering. Proc. VLDB Endow. **16**(2), 193–201 (2022)
5. Cuturi, M.: Fast global alignment kernels. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 929–936 (2011)

6. Cuturi, M., Vert, J.P., Birkenes, O., Matsui, T.: A kernel for time series based on global alignments. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'07, vol. 2, p. II–413. IEEE (2007)
7. Fried, E.I., Cramer, A.O.: Moving forward: challenges and directions for psychopathological network theory and methodology. Perspect. Psychol. Sci. **12**(6), 999–1020 (2017)
8. Hsieh, T.Y., Wang, S., Sun, Y., Honavar, V.: 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, pp. 607–615 (2021)
9. Hsu, E.-Y., Liu, C.-L., Tseng, V.S.: Multivariate time series early classification with interpretability using deep learning and attention mechanism. In: Yang, Q., Zhou, Z.-H., Gong, Z., Zhang, M.-L., Huang, S.-J. (eds.) PAKDD 2019. LNCS (LNAI), vol. 11441, pp. 541–553. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16142-2_42
10. Hwang, H., Whang, S.E.: XClusters: explainability-first clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 7962–7970 (2023)
11. Jacques, J., Preda, C.: Functional data clustering: a survey. Adv. Data Anal. Classif. **8**(3), 231–255 (2014). https://doi.org/10.1007/s11634-013-0158-y
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems, NIPS 2017, vol. 30 (2017)
13. Martínez, A.J., Lemmens, L., Fried, E.I., Roefs, A.: Developing a transdiagnostic ecological momentary assessment protocol for psychopathology (2023)
14. Moshkovitz, M., Dasgupta, S., Rashtchian, C., Frost, N.: Explainable k-means and k-medians clustering. In: International Conference on Machine Learning, pp. 7055–7065. PMLR (2020)
15. Ntekouli, M., Spanakis, G., Waldorp, L., Roefs, A.: Using explainable boosting machine to compare idiographic and nomothetic approaches for ecological momentary assessment data. In: Bouadi, T., Fromont, E., Hüllermeier, E. (eds.) IDA 2022. LNCS, vol. 13205, pp. 199–211. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-01333-1_16
16. Ntekouli, M., Spanakis, G., Waldorp, L., Roefs, A.: Evaluating multivariate time-series clustering using simulated ecological momentary assessment data. Mach. Learn. Appl. **14**, 100512 (2023)
17. Ntekouli, M., Spanakis, G., Waldorp, L., Roefs, A.: Model-based clustering of individuals' ecological momentary assessment time-series data for improving forecasting performance (2023)
18. Ozyegen, O., Prayogo, N., Cevik, M., Basar, A.: Interpretable time series clustering using local explanations. arXiv preprint arXiv:2208.01152 (2022)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)
20. Roefs, A., et al.: A new science of mental disorders: using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. Behav. Res. Ther. **153**(3), 104096 (2022). https://doi.org/10.1016/j.brat.2022.104096
21. Schlegel, U., Vo, D.L., Keim, D.A., Seebacher, D.: TS-MULE: local interpretable model-agnostic explanations for time series forecast models. In: Kamp, M., et al. (eds.) ECML PKDD 2021. CCIS, vol. 1524, pp. 5–14. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93736-2_1
22. Škrlj, B., Džeroski, S., Lavrač, N., Petkovič, M.: Feature importance estimation with self-attention networks. arXiv preprint arXiv:2002.04464 (2020)

23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)
24. Torous, J., et al.: Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. Curr. Psychiatry Rep. **20**, 1–6 (2018)
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, NIPS 2017, vol. 3 (2017)
26. Villani, M., Lockhart, J., Magazzeni, D.: Feature importance for time series data: improving kernelshap. arXiv preprint arXiv:2210.02176 (2022)