

Master's Degree in Machine Learning for Health
2023-2024

Master Thesis

“Human behavior modeling through digital phenotyping for suicide prediction”

Josué Pérez Sabater

Antonio Artés Rodríguez

Leganés, September 2024



Attribution – Non-Commercial – No Derivatives

© 2024. This work is licensed under [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/).

ACKNOWLEDGEMENTS

I would like to thank first and foremost my supervisors Antonio Artés and Pablo M. Olmos, who have patiently guided me in my first steps into the world of research at the University. Also, a special thanks to Miguel Ángel Villalba for his continued efforts on providing and explaining eB2 data, and to Jorge López Castromán for his professional help when dealing with the clinical aspect of the work. Similarly, I thank all other engineers and clinicians involved in the teams at eB2 and Hospital Fundación Jiménez Díaz.

Human behavior modeling through digital phenotyping for suicide prediction

Josué Pérez Sabater
Signal Processing Group
Department of Signal Theory and Communications, UC3M
Leganés, Spain
josue.perez@uc3m.es

Antonio Artés Rodríguez
Signal Processing Group
Department of Signal Theory and Communications, UC3M
Leganés, Spain
aartes@ing.uc3m.es

Abstract—This thesis explores the use of passive digital phenotyping data to predict suicide attempts. The research focuses on the integration of (1) a local observation model to partition the original space of diverse temporal variables into local groups, (2) a heterogeneous mixture model to project each local group into a series of low-dimensional latent profiles series, and (3) a change-point detection algorithm that analyzes these profiles to find behavioral changes indicative of suicidal intent. The dataset involved was specifically curated from a clinical study containing sociodemographic, clinical, and behavioral data from 320 individuals with a history of suicidal behavior. This information was used to analyze the relevance of various behavioral variables in predicting suicide attempts and evaluates different model configurations for optimal performance. The findings suggest that combining behavioral variables like step count, distance traveled, app usage time, time spent at home, ambient light and reported emotion provides significant insights into suicide risk. Furthermore, the research explores the potential of incorporating risk factors such as age, relationship status and education level to personalize the model for better prediction.

Keywords—Suicide behavior, passive digital phenotyping, change-point detection, heterogeneous mixture model, local observation model.

I. INTRODUCTION

A. Background

Suicide is a worldwide health and social challenge affecting people of all ages. The importance of this topic has grown in the last decades to the point of naming suicide prevention a global imperative by the World Health Organization [1]. Suicide has been the leading cause of external death in Spain for several consecutive years, according to data from the National Statistics Institute of Spain [2], and it remains one of the top death causes of any type among the young population. In response to this crisis, there has been a surge in public concern about mental health, prompting efforts by the Spanish

local governments to establish plans for suicide detection, prevention and intervention [3].

Thanks to the growing use of personal digital devices —such as smartphones and wearable technology—, it is now possible to passively collect behavioral data (e.g.: use of mobile apps, distance walked, time spent at home, sleep quality...) from patients of interest. This is commonly referred to as *passive digital phenotyping* (PDP) since it does not require any active action from the user. Information obtained through these technologies can help in the characterization and monitoring of psychiatric patients. This is of great importance in the field of suicide prevention, where a record of suicide attempts is considered to be the major risk factor for repeated attempts [4].

In the face of this context, one of the research lines at the Signal Processing Group (GTS, Grupo de Tratamiento de Señales) at UC3M has explored the possibilities of PDP to detect changes in human behavior patterns that may indicate an intention of suicide attempt, among other applications. The SmartCrisis study [5] is one example of a research initiative aimed at developing and testing a personalized suicide prevention strategy. Patients with a history of suicidal behavior were offered to take part in a 6-month follow-up during which PDP data was collected through mobile health applications. Artificial intelligence-based models (explained in Section I.B below) would then be used to analyze the data and return predictions of potential risk events, which were validated against real suicidal events suffered by the patients. The results obtained by the GTS group in previous work hinted to the viability of using PDP data for such purpose [6], [7]. A few approaches have been studied, but the one currently implemented in some of the laboratory partnerships involves a two-stage, artificial intelligence (AI) model composed of a heterogeneous mixture model, which projects all temporal behavior data into a latent space of daily profiles, and a Bayesian change-point detection algorithm that uses the history of these profiles to identify changes in their occurrence.

B. State of the art

Change-point detection (CPD) consists in the identification of abrupt changes in a temporal or sequential series. The aim is to segment the data in partitions that we consider are generated under different parameters, but without knowing *a priori* when these changes occur [8]. A more specific variation is the Bayesian online approach [9], presented as a probabilistic interpretation of the problem where the observed data x_t are considered to be generated from a distribution with unknown parameters θ_t . Parameters of each partition are independent, and observations are assumed to be independent and identically distributed (i.i.d.). The variable *run length*, r_t , is introduced as the time since the last change point (if there is no change in day t , this counting variable would add +1). The run length is jointly modeled with the observed data, according to the expression:

$$p(r_t, x_{1:t}) = \int p(r_t, x_{1:t}, \theta_t) d\theta_t \quad (1)$$

The goal of the CPD is to find the posterior probability of the run length given all past observations, $p(r_t | x_{1:t})$. This is computed with respect to all previous days (i.e., for each day t , the probability that a change occurred is distributed along all previous days). Then, the *maximum a posteriori* (MAP) estimation r_t^* is obtained by finding the most probable run length for each day. Similarly, an instability measure can be obtained from the run length indicating how noisy the data are. If the individual's behavior is similar every day, instability will be low. On the other hand, if observations change significantly from one day to the next, then the instability will increase. In the cases where instability is higher, performing CPD often becomes more challenging. The MAP estimation is further processed into a binary *prediction vector* with 1 when a change point is detected and 0 otherwise. Different methods (detailed in [10]) can be implemented in this step, which involves the use of a threshold. A visual example of the CPD output is offered in Figure 1.

Online CPD has shown promising results in real-world applications such as water quality monitoring [11] and the analysis of epileptic activity [12]. However, not a lot of research has focused on using CPD methods in the analysis of human behavior, which is usually collected through sensors in personal digital devices (smartphones, wearable technologies...) and expressed in high-dimensional, heterogeneous, periodic variables with a strong presence of missing points [13], [14].

The high dimensionality of human behavior variables leads to greater complexity in the distributions used to define the

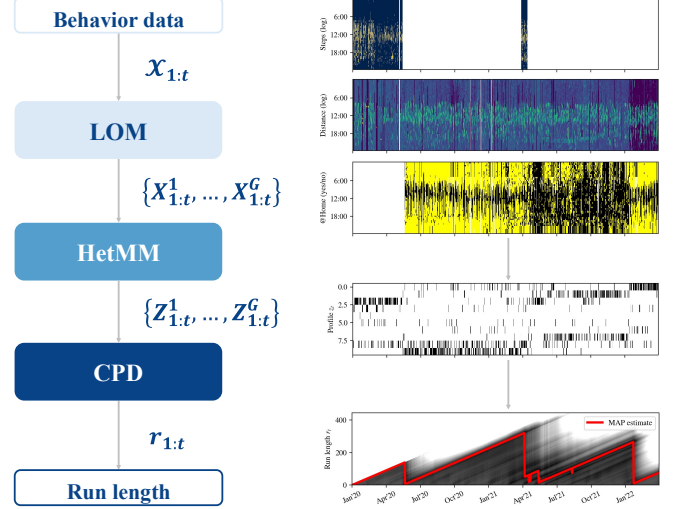


Figure 1. Diagram of the LOM-HetMM-CPD model with the mathematical notation of the relevant data at each stage (left) and a visual, real example (right). Three behavior sources (step count, distance traveled and time at home) are compressed into a latent representation (profile distribution), which is then used to obtain the run length or time since the last change point. Red line indicates its MAP estimation (the most probable run length for each day). Notice how the CPD captures the changes in pattern of the *home* source around April 2021 and January 2022. It also returns a change between April–July 2020, when periods with missing values start or end.

data. As a result, estimating the underlying parameters—and the posterior probability of the run length—becomes a difficult and unreliable process [15]. One solution is to introduce a first stage where raw data are processed in some way that the CPD model is able to work with them. A common strategy is to reduce dimensionality using one of many approaches offered by the fields of machine and deep learning: feature selection, clustering, autoencoding, hidden models, mixture models... Here, the proposed method is to perform a profiling process with a heterogeneous mixture model (HetMM), presented in [16]. This generative model, represented by the equation below, maintains the Bayesian nature of the problem. It assumes that the observed high-dimensional data x_t is generated from a latent, lower-dimensional variable z_t . The latent space is defined by K profiles or components, each with a different set of generative parameters θ_k .

$$p(x_t | \theta_t) = \int p(x_t | z_t) p(z_t | \theta_t) dz_t \quad (2)$$

Each time point (day) is therefore represented by one of these profiles. For instance, if the component activated at time t is $k = 4$, then z_t will have been generated under $\theta_t = \theta_{k=4}$. The pattern of these profiles over time will be analyzed by the CPD model to find changes in the behavior of the subject. The

choice of this technique responds to the ability of the HetMM to jointly handle data of different type (numerical, binary, categorical...) and to deal with partial missing data points (i.e., a time t when some sources are fully or partly missing but others are observed). Previous work offered positive results of this CPD-HetMM integration, specifically in the task of suicide prevention where events of suicidal nature (attempts or self-harm) could be predicted one week in advance [6].

However, insights into these results revealed that relying on categorical profiles can introduce noise when the profile of a day is not modeled with confidence (i.e., the profile distribution for that day is flat), limiting the model’s accuracy. Subsequent research [15] introduced *multinomial sampling* and adapted the CPD to work with probabilities of profiles, rather than a hard profile assignment. Furthermore, the innovative concept of *local observation models* (LOM) was introduced in [7] to better model the integration of high-dimensional, heterogeneous observations: the feature space is divided based on the data’s context, source or statistical type. Each group of observations is transformed into a local, latent representation (one profile representation for each LOM) and different factorization models are proposed to estimate the joint probability of a change point occurring, considering the contributions of each local set. Both modifications have led to improved detection rates, reduced delays, and increased robustness to data heterogeneity. The mathematics behind the derivation of the run length are not the focus of this thesis and thus are not explained here. However, they are described in detail in [15] and [7].

C. Objectives

The main mission behind the work presented here is to deepen our understanding in human behavior and find solutions to real health problems through AI-based applications, particularly in the context of mental health. This may be materialized in two general lines: the exploration of alternatives, modifications or improvements of the models described before, and the innovations on data representation that are expected to enhance the model’s performance. Additionally, the incorporation of sociodemographic information can potentially enable more personalized (and accurate) results. Some specific objectives were defined to help achieve these goals.

- **Curation of dataset.** With the increased number of investigations regarding suicide risk prediction conducted by members at the GTS laboratory, a benchmark database that can serve as reference is expected to facilitate the access to data and the coordination among projects.

The creation of such dataset is proposed, hoping to systematize the organization of information coming from different actors (behavior-monitoring systems, clinical studies, hospital records...) into the same space.

- **Unified model for behavior analysis.** Many models and innovations have been validated and implemented across different projects under the GTS lab. However, there is a lack of a unified version that integrates all of them. The establishment of a single framework that encompasses all model variations and improvements is encouraged. This would allow to combine all past and current work into a single, comprehensive repository to be used as a tool in human behavior analysis.
- **Relevance of behavior variables.** Previous work at GTS mainly involved the use of four sources: app usage, number of steps, distance traveled and time at home. The incorporation of new sources can model new aspects of the user’s behavior, contributing to a more refined profiling. One interesting variable to be studied is the presence of sleep disturbances, which has been linked to a higher risk of suicide behavior [17]. Others may provide complementary information on existing variables, such as the number of phone screen unlocks or the time at work, which are to some extent related to app usage and time at home.
- **Impact by local groups.** Continuing the work in LOM, we propose an analysis of how the observed data can be split according to some fundamental dimensions of human behavior: certain variables collected through PDP methods are part of the same dimension and should be modeled together, creating independent profiles that will then be combined into the CPD. For instance, both the number of steps and distance walked belong to a specific dimension that could be defined as “mobility”, while the time spent on phone apps and the number of screen unlocks may be associated to the “phone use” profile.
- **Personalization based on risk factors.** The database contains information about some risk factors with known clinical importance for suicide, such as sex, age, living with other people, depression, alcohol abuse, other diagnosis, etc. These can be used to further personalize the model in an attempt to reduce the false alarm ratio or increase sensitivity. For example, risk factors can be introduced in the tuning of hyperparameters or in the post-processing of the model predictions.

II. METHODOLOGY

A. Dataset

A custom database was curated as part of this study in order to organize the relevant information in a clear and structured manner. This dataset, which we called *BehaviorCPD*, was designed with the expectation of serving as reference for future work within the GTS laboratory (privacy policies safeguarding data protection prevent the public access to this database). Therefore, the focus was on making it comprehensive and accessible. Several variables that showed potential for further research were included, even if they were not ultimately involved in this thesis. The choice of which variables were relevant was made in collaboration with expert clinicians associated with the research.

BehaviorCPD is built around the patient, which are associated to an anonymized identifier. The database covers a wide variety of patient information, including sociodemographic, clinical, behavioral and emotional data. A total of 320 subjects are included, all of whom have a clinical history of suicidal behavior (i.e., past suicide-related crises or attempts). These subjects had undergone a 6-month suicide-prevention study called SmartCrisis [5]. Data collected during this study (and beyond) was used for the database, and it can be categorized into three groups:

Sociodemographic and clinical (SDC). These variables are static and were gathered from the patient’s hospital records or through a questionnaire at the moment of inclusion into the SmartCrisis study. They encompass sociodemographic data (sex, age, level of studies...), clinical scales (used to assess suicidality, depression, health habits, functioning/disability, sleep quality...) and diagnoses made during the study and obtained from hospital records. Additional relevant data were included during the study regarding the suicide behavior of the patients: 48 events have been annotated and contrasted by clinicians. They correspond to suicide attempts, suicide ideation or self-harm incidents recorded by doctors when the patients arrived at the hospital, either in scheduled appointments or at the emergency room. These events —termed “crisis events”— are of particular importance because they are used as validation of the model predictions. However, following the observations of our clinicians, ideation was not considered a sudden event preceded by behavioral changes, bringing the total count of actual crisis events down to 28.

Active ecological momentary assessment (EMA). This is a common technique used to collect reports from the subject during their daily routine, providing data at moments close to

the patient’s reality. Over 28,000 responses were gathered regularly through an app called MeMind [18], with 34 different questions regarding four main topics: sleep, feelings (including wish to die), appetite and personal relationships. For each report, patients rated their agreement with the given question on a 0–100 scale. Similar EMA methodologies have been successfully used to monitor patients with psychiatric disorders, such as suicidal ideation [19] and depression [20], although its constant intrusion to ask reports may pose a burden on the patient and lead to a decrease in participation along time. On top of that, psychiatric patients may not always report their true feelings, which would undermine the reliability of this source. While EMA data was available in the database, they were not involved in the work presented here.

Passive digital phenotyping (PDP). Behavioral data coming from smartphone or wearable sensors, such as number of steps or app usage, obtained along the patient’s daily routine through an app developed by the group Evidence-Based Behavior (eB2) [21], without any intervention by the user. PDP techniques have the advantage of relieving the user from any action due to their passive nature (a reason why it is sometimes called passive EMA). However, smartphones often kill background processes such as those used by the eB2 app to gather data. This leads to a high presence of missing points, often extending over periods of several days, that affects the quality of data (see Figure 2). Some observations come in the database as daily summaries and are encoded in one or a few dimensions (such as sleep, represented by *duration*, *start time* and *end time*). In contrast, other sources provide all-day information in time slots of 10 or 30 minutes (e.g.: number of steps or app usage). This information can be seen in Table I. It should be noted that a variable called *emotions*, which gathers reports actively made by the user regarding their current emotions, would better fit in the EMA category as it is not passive. Nonetheless, it is provided by the eB2 ecosystem instead of the MeMind app and thus was included here, as well as in the study.

These groups can be referred to as static, active and passive variables, respectively, or by the initials indicated above. In total, the dataset contains 873 features and over 80,000 days of PDP information. It also has some important characteristics: firstly, the 28 crisis events describing suicidal behavior have been manually identified by healthcare workers, benefiting from the contextual understanding of these professionals. Similarly, the behavior variables collected by eB2 offer a detailed digital phenotyping of the individual from a natural perspective, including information on the subject’s mobility, use of phone, biological rhythm and

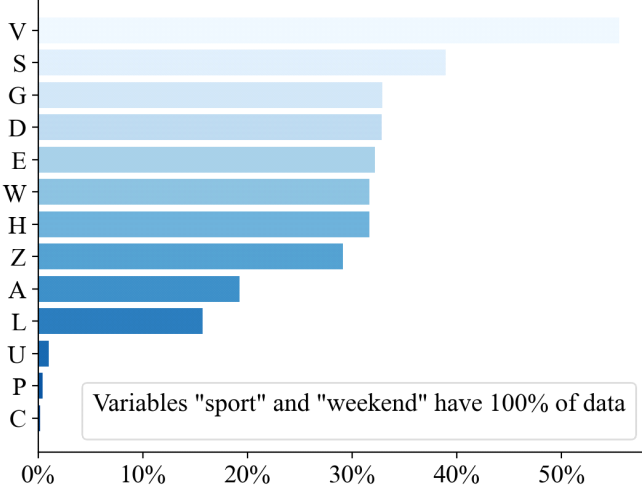


Figure 2. Proportion of available (non-missing) PDP data, considering all patients. Each bar corresponds to a feature in Table I. Variable *activity* (*V*) is the only one surpassing 50% of available evidence, while most others range between 15% and 35%. Many of these missing values belong to days-long periods where the eB2 app could not collect any data. Still, if target events do not happen during such periods, predictions may not be affected that much.

emotions. However, it is the integration of these two —clinical diagnoses and behavior phenotyping— that truly has the potential to uncover the underlying patterns in suicide behavior and aid in its prevention. In fact, both the eB2 and MeMind apps have proved to be feasible methods with potential to monitor suicide risk [22]. On the other hand, an important limitation of the dataset is the imbalance caused by the low number of crisis events. While being expected from a cohort study like SmartCrisis, this disproportion reduces our ability to generalize any findings to the broader population.

B. Models

A common framework was developed to encompass the three main stages of behavior analysis explained in Section I.B. This meant unifying available code scripts of the local observation model, heterogeneous mixture model and change-point detector into a single LOM-HetMM-CPD pipeline, depicted in Figure 1. Implementation of this work was made in Python and an exhaustive documentation was produced, including several Jupyter Notebook files that provide detailed explanations to understand the models and how to use them.

Local observation model. A code was designed to sort behavior PDP observations from each patient, called collectively $\mathcal{X}$, into local groups, creating independent sets of variables $\{X^1, \dots, X^G\}$. The choice of which variables are assigned together will depend on the focus of the study. Some ready-made sorting methods are offered, which may be based on the statistical type (i.e., binary variables in one group,

TABLE I
PASSIVE DIGITAL PHENOTYPING VARIABLES

Name	Type	Dimensions	
A App usage	Numerical	48	Slots of 30 minutes
S Steps	Numerical	48	Slots of 30 minutes
D Distance traveled	Numerical	48	Slots of 30 minutes
P Physical distance	Numerical	48	Slots of 30 minutes
H Time at home	Binary	48	Slots of 30 minutes
W Time at work	Binary	48	Slots of 30 minutes
L Ambient light	Numerical	144	Slots of 10 minutes
G Actigraphy	Numerical	144	Slots of 10 minutes
U Phone unlocks	Numerical	144	Slots of 10 minutes
Z Sleep	Numerical	3	Duration, start time and end time
E Emotions	Mixed	3	Valence, count and ratio
V Activity	Numerical	3	Walking, running and on vehicle
C Circadian data	Numerical	3	Acrophase, MESOR and amplitude
R Sport	Binary	1	Flag
K Weekend	Binary	1	Flag
Q Quality scores	Numerical	5	App, steps, distance, actigraphy/light and unlocks

The type of variable and its dimensions are given. The letters at the beginning of each row are used to identify the source. More detailed descriptions of PDP variables are given in the Appendix.

numerical in another one...) or the context (data that belong to some specific dimension, such as “mobility”, are modeled together). There is also the option to model each variable separately, or to join all of them into a single characterization. These divisions were proposed in [7], but the new code automatizes and streamlines the process. Custom local groups can also be passed to the model, allowing a completely controlled environment that facilitates the analysis of data sources.

Heterogeneous mixture model. For each local group defined in the LOM, an independent profiling process was performed using a heterogeneous mixture model (HetMM). For example, for local group g , the set of variables X^g was used to generate a profile series z_t^g , where each day is assigned to one of K components of the corresponding mixture g . This classification can be seen as a latent projection that

compacts all sources in the set together, effectively reducing dimensionality of the input data. The output is given both as a hard *categorical* result (a series of integers indicating which component of the mixture is activated) and as a *distribution* (the likelihood of having each of the profiles each day). The number of possible components or profiles, K , was set to 10. This value has been used in previous studies with satisfactory results [23]. However, it is assumed that not all subjects should necessarily be modeled with the same number of profiles, and therefore an option to automatically adapt this hyperparameter, rather than using a fixed value, is also available. This enables a more personalized study of each patient, but it was not implemented during our experiments for the sake of simplicity.

Change-point detection. Codes for the CPD stage were based on those provided by the authors of previous work. Two possible CPD models were defined: the original hierarchical approach that can be used with categorical profiles, and the multinomial version presented in [15], adjusted to accept distributions of profiles. A third model called *union CPD* was adapted to integrate several input profiles—coming from the local sources defined in the LOM—and produce a single detection output. Each profile is analyzed with its corresponding CPD type, but then all are combined according to some factorization model that controls how much each local group contributes to the final run length. The same factorization methods explained in [7] were added.

When the MAP estimation of the run length is transformed into a prediction vector, a hyperparameter is involved to mark the decision threshold for marking positives. This threshold can be swept to produce a receiver operating characteristic (ROC) curve, which we used to assess the model trade-off between sensitivity (ability to correctly identify crisis events) and specificity (ability to not raise false alarms, i.e., not returning a positive when there are no events). Moreover, the commonly used metric of area under the curve (AUC) was used as a score to compare different architectures and configurations of the model, and it will be used in the experiments presented in the next section.

C. Experiments

Taking advantage of the diverse data presented in Section II.A, three experiments are proposed. They were designed to help achieve the objectives stated in Section I.C. During these experiments, the change points predicted by the LOM-HetMM-CPD model were compared with information on real suicide behavior exhibited by the patients. These “crisis events” were obtained through the SmartCrisis study and

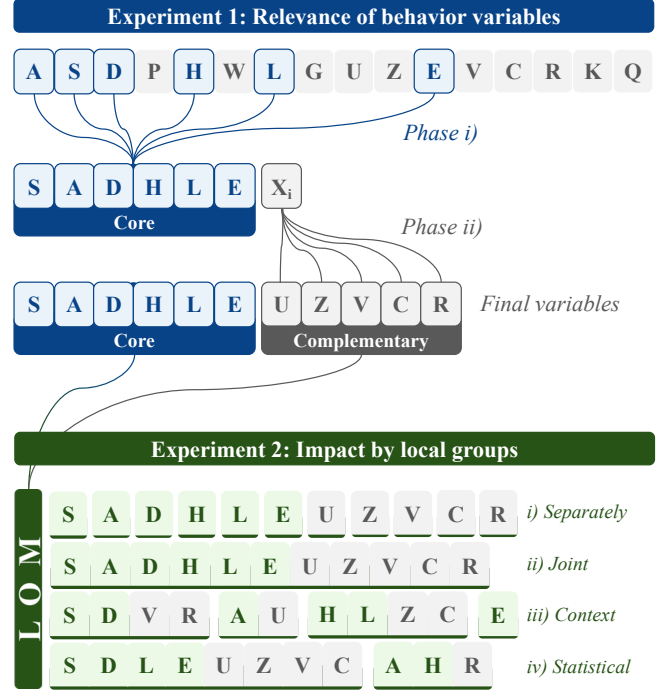


Figure 3. Diagrams explaining the first and second experiments. Refer to Table I to see the one-letter code corresponding to each source. **Experiment 1:** an analysis of feature relevance is made in phases i) and ii). Six core features (with significant correlation with crisis events) and five complementary ones were found. **Experiment 2:** the impact of local groups is compared for different combination methods. Sources that appear together form a local group for which a unique profiling was defined.

conform our validation data, i.e., the events we aim to predict by detecting changes in human behavior. Unfortunately, fourteen patients with crisis events were excluded from the study because such events happened during periods where the eB2 app was not active, and therefore no behavior data at all is available. This brought the number of validation events down to 14.

Experiment 1: Relevance of behavior variables. A total of 16 features were present in the dataset: fifteen passive variables plus the active reports about emotions, with dimensions varying from 1 to 144 and diverse statistical types. These are shown in Table I (for a more detailed description of each source, see Appendix). It is expected that some variables will contribute more than others to define the mental state of a patient. For instance, sources used in past research have already shown correlation with attempting suicide, while data from the light sensor or actigraphy may not be particularly relevant. Therefore, the first experiment was made to analyze the impact of each behavior variable on the ability of the union CPD to detect suicide events. However, the approach followed was based on the assumption that no *a priori*

Results of experiment 1: Relevance of behavior variables

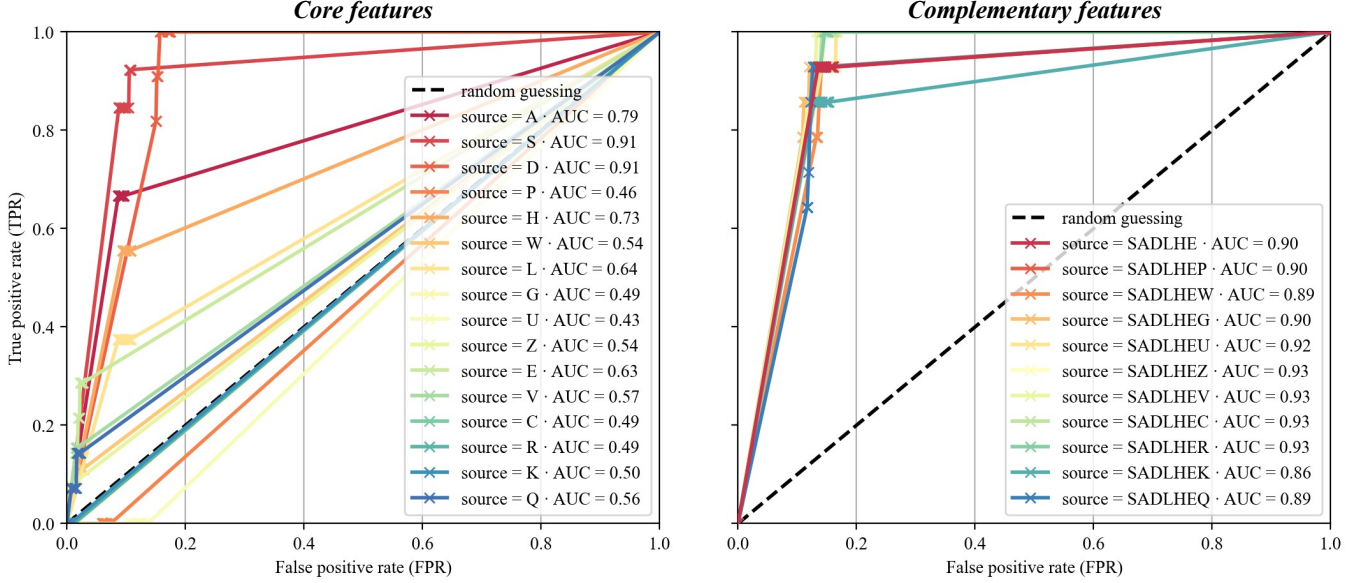


Figure 4. ROC curves for the first experiment analyzing the relevance of behavior variables. Each curve represents the performance of a model using a different behavior source (see the one-letter code in Table 1), and each cross indicates a different value for the model threshold. Despite using an exhaustive grid of threshold values, the points are quite concentrated for each line, indicating low dependency on this threshold and a low number of true events, which accumulates points in a few TPR values. **Phase one** (core features, left): each model tested with a single source. The selected core features (with an AUC score above 0.6) are step count, distance, app usage, time at home, light and emotions. **Phase two** (complementary features, right): core features are now modeled together, and remaining sources are added to them, one at each time. Results improved when these complementary features were used: phone unlocks, sleep time, physical activity, circadian data and sport flag.

knowledge existed, to avoid potential biases. For that, the identification of relevant features was separated in two phases (these steps are illustrated in Figure 3).

- i) *Core features.* An initial sweep was performed where each variable was used by itself as the only input to the model. Results were compared based on the AUC scored and the variables with best results were selected. These would be contemplated as the “core” features containing most of information.
- ii) *Complementary features.* Some observed variables may not be very informative by themselves, but still improve the detection when combined with others. To corroborate if this was the case, a new round of models was trained using all core features plus one of the remaining sources each time. Those that improved the AUC (compared to using only the core features) were considered “complementary” features, i.e., with some informative value that can to some extent help in the detection.

While the focus of this study was on passive sources, the inclusion of some active data (emotions) was tolerated because it belonged to the same provider (eB2) and could offer some additional insights.

Experiment 2: Impact by local groups. Once the set of features of interest (core + complementary) was established, the focus moved on how these variables should be jointly modeled in order to maximize their contribution. The idea behind is that, rather than aggregating all observations together without any order, they could be grouped in a specific way to produce more than one profile distribution, each modeling a different aspect of human behavior. These profiles can then be inputted to the union CPD. However, combining all variables in all possible ways would require too many runs and consume an excessive amount of computation resources and time. Instead, the experiment was simplified and just a few options were tested (depicted in Figure 3), based on [7] and our own intuition:

- i) *Separate.* Each source is represented in an independent profile. The number of profile distributions that are given to the CPD is therefore as many as core + complementary features were selected. This method is actually representing each source in a latent space of $K = 10$ dimensions, which may be suitable for variables that had originally many dimensions (such as *light*) but not so efficient for those that were lower-

dimensional (like *sleep*). Thus, it is not expected to perform especially well.

- ii) *Joint*. All variables are represented jointly in a single profile, achieving the largest degree of compression. In this case, more weight is given to the HetMM stage, whose underlying parameters will need to model a probability distribution adapted to many sources at once. Meanwhile, the work of the CPD is made easier because its input is less complex and less diverse. This option is equivalent to not using the LOM stage at all.
- iii) *Context*. Variables are grouped according to some prior knowledge about their context. The fields proposed here are mobility (steps, distance, activity and sport), phone use (app usage, screen unlocks), biological rhythm (home, light, sleep, circadian data) and mental state (emotions). Especial care was taken to ensure that each group would contain at least one core feature, which are underlined. This was the most promising configuration because it considers the subjacent correlation between variables of the same “human domain” (i.e., related to a same aspect of human behavior).
- iv) *Statistical type*. Variables are grouped according to their statistical type, which can be numerical (steps, distance, light, emotions, unlocks, sleep, activity, circadian data) or binary (app usage, presence at home, sport flag). The advantage presented by this alternative lies on the fact that profiles are no longer heterogeneous (because variables in each HetMM will all follow the same likelihood function, just with different parameters), simplifying the mathematical operations behind the construction of the profiles.

This second experiment was run for different values of hyperparameter λ , involved in the so-called *hazard function* [24] that defines the prior probability of having a change point at any given time instant. The performance of the CPD may be affected by this hyperparameter, which can be tuned to adapt its sensitivity. The four values of λ used are indicated in Table III.

Personalization with risk factors. Our third experiment focused on the information contained in sociodemographic and clinical data (SDC) and its potential to enhance model performance. Among all 35 SDC variables contained in the dataset, those with potential to influence suicidal behavior (i.e., risk factors) were identified. This task was made in collaboration with a psychiatrist expert in suicide and by researching the literature [4], [25]. The risk factors selected are shown in

Table II and some of them may be related to several variables in the dataset. These factors were used to divide the sample

TABLE II
STATIC SOCIODEMOGRAPHIC AND CLINICAL VARIABLES

Risk factor	Related variables
Demographic factors	
Sex	
Age	
Living with	<i>Demographic factors are all associated to just one variable. They were collected at the moment of inclusion of the patient in the study.</i>
Civil status	
Work status	
Education level	
Clinical factors	
Suicide behavior	<i>C-SSRS</i> : suicide ideation and attempts. <i>MINI</i> : intensity of suicidality, suicidal behavior disorder
Depression	<i>IDS-CR</i> and <i>QIDS-SR</i> : depression scales <i>MINI</i> : major depressive disorder or episodes, with or without psychosis.
Anxiety	<i>GAD-7</i> : scale of anxiety <i>MINI</i> : social and generalized anxiety disorders
Panic	<i>MINI</i> : panic disorder
Personal factors	
Victim of sexual abuse	<i>CTQ-SF</i> : questionnaire about childhood trauma
Alcohol abuse	<i>CAGE</i> : questionnaire addressing alcohol use <i>MINI</i> : alcohol use disorder

Each risk factor may be associated to more than one variable from the dataset. *MINI* data correspond to diagnoses, while other names correspond to questionnaires or scales. Detailed descriptions of SDC variables are given in the Appendix.

of 320 patients into smaller subsamples, with the idea of comparing the CPD performance based on the demographic or clinical background. However, because the number of subjects (and the amount of validation events) is not large enough, only a few comparisons were possible. The combination of passive sources involved in this experiment was the one considered to perform best in the previous experiments.

III. RESULTS AND DISCUSSION

The creation of the *BehaviorCPD* repository —containing the database, the models and the pipelines to implement them— enabled to assemble the work of passed researchers together,

introduce changes and major improvements in specific parts of the models, boost code efficiency, consistency and readability, and document every part of the process. This work provides a unified, easy-to-use and well-explained tool to implement and interact with the change-point detector. The private repository can be accessed by members of the GTS laboratory and imported as a Python module, facilitating the task of behavior analysis. However, the repository was not made publicly available due to privacy regulation protecting patient data and confidentiality surrounding the algorithms used by eB2.

The results of the **first experiment** show that, from the original 16 behavior variables present in the dataset, only 11 of them contribute to characterizing risk of suicide behavior. As it can be seen in Figure 4, the sources that seem more informative by themselves (i.e., using one source as the only input for the model) are the step count and distance traveled, both with an AUC of 0.91. This score is close to 1, indicating that the model works on a region with high sensitivity and specificity. In fact, for some threshold values, the model was able to detect all 100% of suicide events with a false alarm ratio below 20%, which is considerably notable. The two best sources are followed by the time of app usage and the time at home. This outcome is consistent with the decisions made by previous authors, who used exactly those four sources in their experiments due to their high significance and low missingness.

However, the amount of ambience light and the emotions reported by the users also indicated some connection with crisis events, although their AUC scores fell on the 0.6–0.7 range. Light detected by the phone sensor is possibly linked to the time spent outdoors or in places with abundant natural light. It has been proved that the amount of light one receives affects our emotional state and mental health [26], which may be the reason behind the correlation between this variable and suicide behavior.

Regarding the emotions, it should be reminded that, despite not being technically “passive”, this feature is also part of the data managed by eB2, and we decided to include it in the study to allow some comparison between active and passive sources. Patients can record their emotions in the eB2 app at any point, which is expected to provide rich information about their state in real time. Still, some users may overamplify their reported emotions and others may not report them at all, introducing subjective biases and leading to a not-so-accurate description of their reality. Both situations were actually observed in our dataset, where some patients

TABLE III
RESULTS OF EXPERIMENT 2: IMPACT OF LOCAL GROUPS

	$\lambda = 10^3$	$\lambda = 10^5$	$\lambda = 10^7$	$\lambda = 10^{10}$
Separate	0.86	<u>0.95</u>	0.82	0.85
Joint	<u>0.91</u>	0.90	<u>0.91</u>	<u>0.89</u>
Context	0.89	0.87	0.88	0.86
Statistical type	0.89	0.90	0.87	0.84

The table contains AUC scores obtained for all LOM configurations evaluated in the second experiment. Four values of the hazard function hyperparameter λ were tested. For each of them, the best result is underlined.

introduced dozens or even hundreds of emotions in one single day, while 26% of them did not record any emotion. Moreover, research suggests that some people are hesitant to give honest answers when these are socially undesirable [27], and that repeated measurements of one’s emotions alter the intensity of such measurements [28]. These circumstances may affect the validity of this variable and cause it to perform worse in capturing suicidal behavior, therefore yielding a lower AUC compared to other true passive sources.

Besides these core features, some of the complementary features slightly improved the AUC by two or three points over the reference. The sleep time, physical activity, circadian data, sport flag and number of phone unlocks were all found to contribute positively in the detection. While this increase in AUC may not seem of great significance, the use of such complementary features successfully helped detect 100% of suicide events, even if it increased the percentage of false positives (although it never exceeded 17%). While these false alarms should ideally be minimized, they may be acceptable in the context of suicide prevention if it ensures that all patients with a real risk of attempting suicide are effectively identified and treated. Furthermore, these false positives may still be signs of instability or changes in the subject’s life, and dedicating special attention to the patient during these moments may be of clinical importance and improve the quality of mental health monitoring and assistance.

Regarding the **second experiment**, Table III shows the AUC score achieved by each grouping option (separate, joint, by context and by statistical type) for each value of λ . The LOM successfully produced an independent profile distribution for each of the groups, which were then combined to feed the CPD. The best overall results were attained when all observations were modeled together into one single profile (“joint” configuration), for three of the four values of λ tested (10^3 , 10^7 and 10^{10}). This outcome suggests that the contribution of each source to detect suicide attempts is best exploited

when processed as a single latent variable or profiling. If this were true, the LOM stage might not be a necessary tool, at least in the specific case study proposed here.

The optimal AUC, nonetheless, was achieved for $\lambda = 10^5$ using the “separate” approach, where each source was projected into a different profile representation. This model reached a score of 0.95, surpassing the second best by 5 points. Such a finding was contrary to our original intuition because we expected that modeling every source—including those with just 1 or 3 dimensions—separately as a 10-dimensional profile distribution would complicate the detection task of the CPD. Furthermore, these results are especially surprising considering that the same LOM method performed relatively poorly for all other λ values, being the worst choice in two of them. A reason behind this behavior could be that the positive output for $\lambda = 10^5$ was just an accidental

anomaly and does not truly portray the real properties of the model. However, the “separate” approach also emerged as the best alternative in [7], where it is called “Sources OM”. So, this option should not be disregarded in future research.

The two last methods sorted variables into local groups according to their context or statistical type. Our initial hypothesis proposed that separating sources by their context would better model the individual’s behavior and enhance CPD performance. Nevertheless, while both aggragation strategies performed similarly, none of them improved the joint model and thus our premise could not be verified. Some possible explanations are identified: first and foremost, human behavior is a complex phenomenon conditioned by several internal and external factors that are not easily characterized: personal bonds and experiences, social interactions, mental and physical health... All these probably

Results of experiment 3: Personalization with risk factors

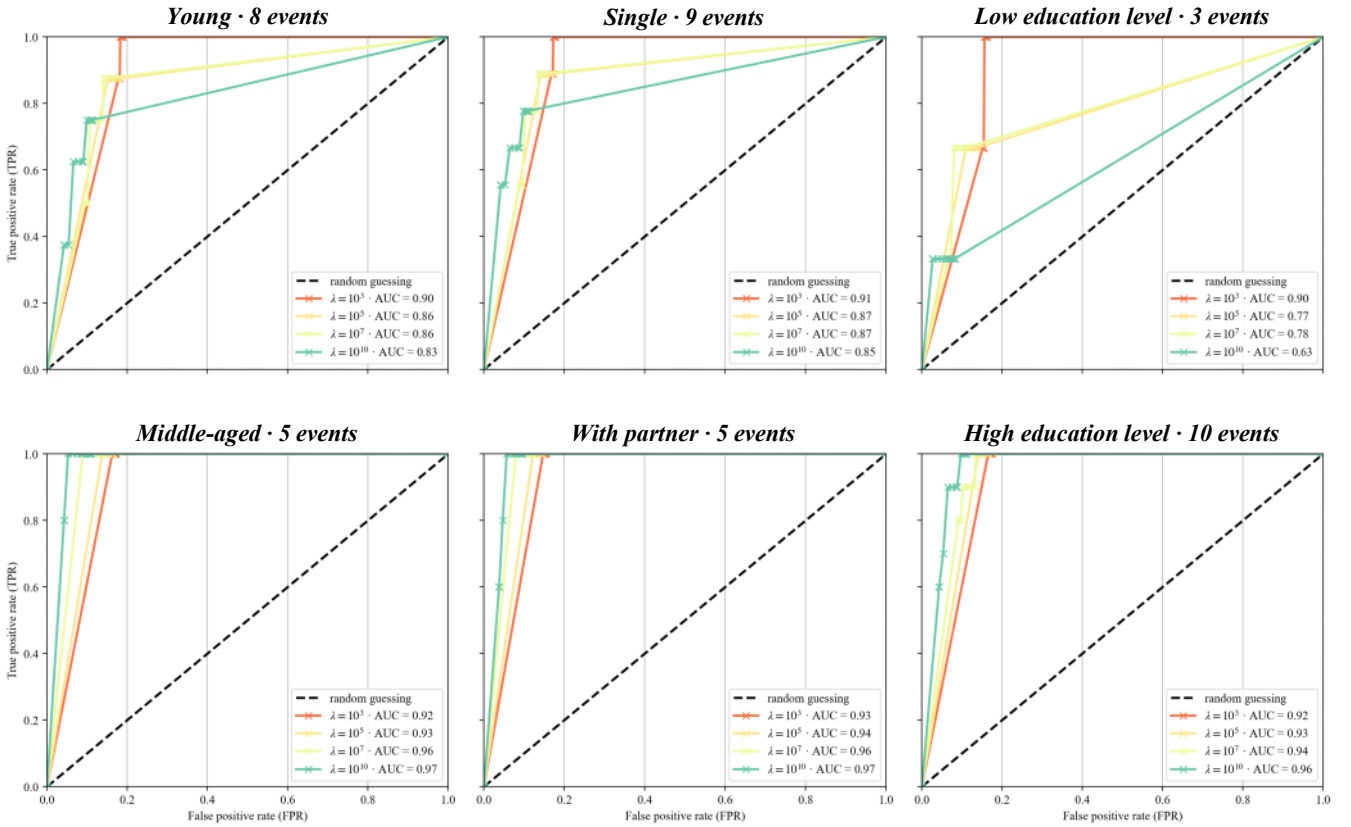


Figure 5. Some examples of the results of the third experiment, which studies the differences in model performance based on risk factors. Three variables are shown: age (left), civil status (middle) and education level (right). Each of them divides the test sample of 320 patients in two subgroups (top and bottom) where the LOM-HetMM-CPD model was run independently, and the resulting ROC curves were plotted. The size of each subgroup can be consulted in the corresponding table of the [Appendix](#). These three risk factors were selected because they display some significant differences between the two studied subgroups: overall, suicidal events enacted by individuals in their midlife, with partner and high education level were easier to predict by the model, reaching 100% of specificity for all values of λ . On the contrary, young, single subjects with less years of study were more challenging to predict, often leading to lower true positive rates and AUC scores. Still, the small sample sizes remain an important limitation to drawing any confident interpretation.

affect the regular activities of a person at many levels. For instance, spending the day with a friend could possibly imply an increase in the distance travelled, a decrease in the use of smartphone and an improvement in the emotions felt by the subject. If these observations are not independent but rather connected by a confounding factor, a joint profiling could capture these underlying relationships and offer a more accurate and comprehensive representation of human behavior.

A second possible reason is that the CPD works better with less input profiles, although this idea is challenged by the fact that the highest AUC was obtained with the “separate” LOM, which introduces 11 profile distributions to the CPD. Anyway, the adaptability offered by the LOM at controlling source modeling is valuable enough to still be considered a useful addition. If local groups do not contribute to improving the results, the “joint” option can always be implemented.

Finally, the **third experiment** encountered a significant limitation: when the original sample was split into subgroups according to the risk factors in [Table II](#), the number of suicide events occurring in those subgroups was often too low to draw any robust conclusions. Despite this, the analysis of ROC curves revealed that certain demographic and clinical factors can influence the predictability of suicide attempts. While some subgroups showed distinct patterns, the limited number of events in many categories makes it challenging to offer definitive interpretations. Some key findings are given:

- Compared to younger patients, middle-aged individuals (30–55 years old) with suicidal behavior were more accurately predicted by the model, exhibiting higher AUC scores and lower false positive rates. The same happened for subjects with partners (whether married or not) in contrast with those who were single. Also, individuals with a lower education level (less than 10 years) were more difficult to predict, but the small sample size limits the generalizability of this finding. The ROC curves for these risk factors are displayed in [Figure 5](#).
- Regarding the variables of work status, living with, anxiety, sexual abuse and panic disorder, the insufficient number of events in their corresponding partitions hindered the task of drawing any conclusions. On a different side, the patient’s sex (limited to “man” or “woman”) and alcohol use presented similar curves regardless of the subgroup, with no meaningful insights offered.
- Similarly, given the nature of the study, all patients had some history of suicidal ideation or attempts, making it

hard to isolate the impact of this factor. And concerning the assessment of depression, opposite trends were observed depending on the specific resource used to identify patients affected by depression (either scores obtained from on-the-spot questionnaires made by a clinician, or diagnoses based on the longer clinical history of the patient). Also, the limited number of patients identified with depression restricted a more in-depth analysis of this group.

In summary, while some variables exhibited different tendencies, the insufficient sample size and the complex interplay of factors make it difficult to establish definitive relationships between demographic and clinical variables and the predictability of suicide attempts.

IV. LIMITATIONS AND FUTURE OUTLOOK

The results achieved by the LOM-HetMM-CPD pipeline in identifying potential risk events by analyzing passively collected behavioral data of suicide patients are promising. Still, some limitations inherent to the methodology and dataset require further exploration.

- **Incompleteness in passive digital phenotyping.** The original number of patients with suicide attempts or visits to the hospital, 28, was reduced to 14 because half of them happened in periods where the eB2 app was not collecting data. In some cases, risk events happened during periods of several missing days in the middle of the follow-up stage, while others occurred beyond the last recorded date, probably when the patient no longer had the eB2 app installed. This situation highlights one of the main drawbacks of passive digital phenotyping: the subjection to factors like battery life, sensor malfunction, network connectivity and user behavior that may affect the completeness and quality of the data. In our case, this led to a very low number of valid clinical events, hindering our ability to draw robust conclusions and limiting the generalizability of our findings.
- **Small sample size.** As it is discussed in [Section III](#), the results of the third experiment—the one focusing on analyzing the model performance under different risk factors—are not easily interpretable nor generalized to the general population. This is because the number of patients with crisis events in the subgroups is very low, which highlights the necessity of a larger sample size. While some notions about differences between population subsets can be suggested, we decided to not draw any firm conclusions based on the available results, and

instead propose further investigation with larger sample sizes to properly assess the effect of demographic and clinical variables.

- **Exploration of model configurations.** The LOM-HetMM-CPD pipeline was designed with flexibility and adaptability in mind in order to make it compatible with diverse data formats and analysis scenarios. This leads to an increase in the number of hyperparameters and possible configurations, from which only a handful were used in this thesis to reduce its scope. Some of the untouched settings are the size of the temporal window to capture events, the factorization method to combine local groups and the method in which the threshold is applied on the run length to transform it into a binary prediction vector. These and more could have an important impact on the model’s performance and should be further explored.

Additionally, we outline here some potential future directions for research, aiming to enhance the model’s accuracy, robustness and clinical applicability in predicting and preventing suicide behavior.

- **Imputation of missing days.** Addressing the issue above, the problem of missing days could be solved with generative models such as variational autoencoders (VAE), which have the ability to reconstruct missing signals. A project within the GTS laboratory is currently exploring the application of a VQ-VAE model [29] in the profiling step, and could potentially address the issue of missing data in behavior signals or profiles.
- **Modeling of low-dimensional variables.** With the introduction of the HetMM stage, the observed variables are described with some likelihood distributions (Gaussian, Bernoulli...) and projected into profiles of a chosen dimension size, in this thesis 10. However, there are low-dimensional sources in the dataset whose contribution could be more effective if they were directly fed to the CPD with no profiling made, such as the sleep data. A proposal is made to investigate this possibility by introducing the normal-inverse-Wishart distribution as the conjugate prior of low-dimensional variables that follow a multivariate normal distribution [30]. By doing so, the input source is not represented as a profile, but rather modeled in a way that can capture correlations

between the different dimensions of the variable while allowing Bayesian inference, making it compatible with the CPD.

- **Incorporation of subsequent suicide events.** The dataset constructed for this work contains only the first suicide-related event of each patient, if any. Posterior events were not considered because the prior probability of having a change point may be affected by suicide attempts within the previous days, and this effect is not accounted for by our model. This means that some of the predictions of the CPD may be correct but are evaluated as false alarms if they occur after a previous event. However, if two suicide episodes are separated far in time, they can be assumed to be independent from each other and the prediction given by the detector would then be considered valid. This perspective was contemplated and would have been implemented, but dates regarding subsequent events were ultimately not provided by the hospital collaborators. The idea remains therefore as an avenue to be tested in future work.

V. CONCLUSION

The presented thesis highlights the potential of using passive behavior data, analyzed through the LOM-HetMM-CPD pipeline, to predict suicide attempts. The high specificity of the model indicates the validity of using passive phenotyping in behavior characterization. The study finds that specific behavioral variables, particularly when modeled jointly, offer effective predictive signals. However, the limitations posed by a strong presence of missing values and small sample size underscore the need for further research with larger and more complete datasets. This was especially manifested when attempting to assess the impact of risk factors, which were used to split the general sample into subgroups that contained insufficient true positives to confidently claim any findings.

Future directions include exploring the imputation of missing data, refining the modeling of low-dimensional variables, and incorporating subsequent suicide events for a comprehensive understanding of behavioral patterns. Overall, the insights from this study contribute to the development of more effective suicide prediction and prevention strategies, and it will serve as reference for continued research on the topic.

REFERENCES

- [1] A. Parekh and M. Phillips, *Preventing Suicide: A Global Imperative*. 2014.
- [2] “Nota de Prensa: Estadística de Defunciones según la Causa de Muerte. Año 2023. Datos provisionales.” INE. Accessed: Aug. 06, 2024. [Online]. Available: <https://www.ine.es/dyngs/Prensa/es/pEDCM2023.htm>
- [3] T. Sufrate-Sorzano *et al.*, “Health Plans for Suicide Prevention in Spain: A Descriptive Analysis of the Published Documents,” *Nurs. Rep.*, vol. 12, no. 1, pp. 77–89, Feb. 2022, doi: 10.3390/nursrep12010009.
- [4] M. Beghi, J. F. Rosenbaum, C. Cerri, and C. M. Cornaggia, “Risk factors for fatal and nonfatal repetition of suicide attempts: a literature review,” *Neuropsychiatr. Dis. Treat.*, vol. 9, pp. 1725–1736, Nov. 2013, doi: 10.2147/NDT.S40213.
- [5] S. Berrouiguet, M. L. Barrigón, J. L. Castroman, P. Courtet, A. Artés-Rodríguez, and E. Baca-García, “Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: the Smarterises study protocol,” *BMC Psychiatry*, vol. 19, no. 1, p. 277, Sep. 2019, doi: 10.1186/s12888-019-2260-y.
- [6] P. Moreno-Muñoz, L. Romero-Medrano, Á. Moreno, J. Herrera-López, E. Baca-García, and A. Artés-Rodríguez, “Passive detection of behavioral shifts for suicide attempt prevention,” Nov. 14, 2020, *arXiv: arXiv:2011.09848*. doi: 10.48550/arXiv.2011.09848.
- [7] L. Romero-Medrano and A. Artés-Rodríguez, “Multi-Source Change-Point Detection over Local Observation Models,” *Pattern Recognit.*, vol. 134, p. 109116, Feb. 2023, doi: 10.1016/j.patcog.2022.109116.
- [8] E. S. Page, “A Test for a Change in a Parameter Occurring at an Unknown Point,” *Biometrika*, vol. 42, no. 3/4, pp. 523–527, 1955, doi: 10.2307/2333401.
- [9] R. P. Adams and D. J. C. MacKay, “Bayesian Online Change-point Detection,” Oct. 19, 2007, *arXiv: arXiv:0710.3742*. doi: 10.48550/arXiv.0710.3742.
- [10] L. Romero Medrano, “Change-Point Detection Methods for Behavioral Shift Recognition in Mental Healthcare,” 2022. Accessed: Sep. 04, 2024. [Online]. Available: <https://hdl.handle.net/10016/36642>
- [11] A. Ba and S. A. McKenna, “Water quality monitoring with online change-point detection methods,” *J. Hydroinformatics*, vol. 17, no. 1, pp. 7–19, Jul. 2014, doi: 10.2166/hydro.2014.126.
- [12] R. Malladi, G. P. Kalamangalam, and B. Aazhang, “Online Bayesian change point detection algorithms for segmentation of epileptic activity,” in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov. 2013, pp. 1833–1837. doi: 10.1109/ACSSC.2013.6810619.
- [13] E. Reinertsen and G. D. Clifford, “A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses,” *Physiol. Meas.*, vol. 39, no. 5, p. 05TR01, May 2018, doi: 10.1088/1361-6579/aabf64.
- [14] P. A. Bloom *et al.*, “Identifying factors impacting missingness within smartphone-based research: Implications for intensive longitudinal studies of adolescent suicidal thoughts and behaviors,” *J. Psychopathol. Clin. Sci.*, p. No Pagination Specified–No Pagination Specified, 2024, doi: 10.1037/abn0000930.
- [15] L. Romero-Medrano, P. Moreno-Muñoz, and A. Artés-Rodríguez, “Multinomial Sampling of Latent Variables for Hierarchical Change-Point Detection,” *J. Signal Process. Syst.*, vol. 94, no. 2, pp. 215–227, Feb. 2022, doi: 10.1007/s11265-021-01705-8.
- [16] P. Moreno-Muñoz, D. Ramírez, and A. Artés-Rodríguez, “Change-Point Detection on Hierarchical Circadian Models,” Mar. 22, 2019, *arXiv: arXiv:1809.04197*. doi: 10.48550/arXiv.1809.04197.
- [17] R. A. Bernert and T. E. Joiner, “Sleep disturbances and suicide risk: A review of the literature,” *Neuropsychiatr. Dis. Treat.*, vol. 3, no. 6, pp. 735–743, Dec. 2007.
- [18] M. L. Barrigón *et al.*, “User profiles of an electronic mental health tool for ecological momentary assessment: MEmind,” *Int. J. Methods Psychiatr. Res.*, vol. 26, no. 1, p. e1554, Mar. 2017, doi: 10.1002/mpr.1554.
- [19] E. K. Czyz, C. A. King, and I. Nahum-Shani, “Ecological assessment of daily suicidal thoughts and attempts among suicidal teens after psychiatric hospitalization: Lessons about feasibility and acceptability,” *Psychiatry Res.*, vol. 267, pp. 566–574, Sep. 2018, doi: 10.1016/j.psychres.2018.06.031.
- [20] S. D. Targum, C. Sauder, M. Evans, J. N. Saber, and P. D. Harvey, “Ecological momentary assessment as a measurement tool in depression trials,” *J. Psychiatr. Res.*, vol. 136, pp. 256–264, Apr. 2021, doi: 10.1016/j.jpsychires.2021.02.012.
- [21] S. Berrouiguet *et al.*, “Combining Continuous Smartphone Native Sensors Data Capture and Unsupervised Data Mining Techniques for Behavioral Changes Detection: A Case Series of the Evidence-Based Behavior (eB2) Study,” *JMIR MHealth UHealth*, vol. 6, no. 12, p. e197, Dec. 2018, doi: 10.2196/mhealth.9472.
- [22] A. Porras-Segovia *et al.*, “Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study,” *J. Affect. Disord.*, vol. 274, pp. 733–741, Sep. 2020, doi: 10.1016/j.jad.2020.05.067.
- [23] M. L. Barrigón *et al.*, “One-Week Suicide Risk Prediction Using Real-Time Smartphone Monitoring: Prospective Cohort Study,” *J. Med. Internet Res.*, vol. 25, no. 1, p. e43719, Sep. 2023, doi: 10.2196/43719.
- [24] O. C. Ibe, “Chapter 4 - Special Probability Distributions,” in *Fundamentals of Applied Probability and Random Processes*, 4th ed., O. C. Ibe, Ed., Boston: Academic Press, 2014, pp. 141–143. doi: 10.1016/B978-0-12-800852-2.00004-3.
- [25] S. Masango, S. Rataemane, and A. Motojesi, “Suicide and suicide risk factors: A literature review,” *South Afr. Fam. Pract.*, vol. 50, no. 6, pp. 25–29, Nov. 2008, doi: 10.1080/20786204.2008.10873774.

- [26] J. Morales-Bravo and P. Navarrete-Hernandez, "Enlightening wellbeing in the home: The impact of natural light design on perceived happiness and sadness in residential spaces," *Build. Environ.*, vol. 223, p. 109317, Sep. 2022, doi: 10.1016/j.buildenv.2022.109317.
- [27] D. Ciuk, A. Troy, and M. Jones, "Measuring Emotion: Self-Reports vs. Physiological Indicators," Apr. 16, 2015, *Rochester, NY*: 2595359. doi: 10.2139/ssrn.2595359.
- [28] O. Johar and A. M. Sackett, "The Self-Contaminating Nature of Repeated Reports of Negative Emotions," *Basic Appl. Soc. Psychol.*, vol. 40, no. 5, pp. 293–307, Sep. 2018, doi: 10.1080/01973533.2018.1496336.
- [29] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," May 30, 2018, *arXiv*: arXiv:1711.00937. doi: 10.48550/arXiv.1711.00937.
- [30] E. Goffinet, M. Lebbah, H. Azzag, and L. Giraldi, "Predictive distributions of random variables following a multivariate Gaussian distribution with Normal Inverse Wishart prior," *LIPN*, Dec. 2021. Accessed: Sep. 05, 2024. [Online]. Available: <https://hal.science/hal-03962723>
- [31] A. J. Rush *et al.*, "The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression," *Biol. Psychiatry*, vol. 54, no. 5, pp. 573–583, Sep. 2003, doi: 10.1016/S0006-3223(02)01866-8.
- [32] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, May 2006, doi: 10.1001/archinte.166.10.1092.
- [33] D. Bernstein, L. Fink, and D. Bernstein, "Childhood Trauma Questionnaire: A retrospective self-report manual," 1998. Accessed: Aug. 31, 2024. [Online]. Available: <https://www.semanticscholar.org/paper/Childhood-Trauma-Questionnaire%3A-A-retrospective-Bernstein-Fink/4d8f82acd3fefeb318c04295985fa0362d0c1672>

APPENDIX

This appendix contains detailed descriptions of the variables involved in the experiment: passive digital phenotyping variables used in the first experiment (relevance of behavior variables) and static sociodemographic and clinical variables in the second one (personalization with demographic data).

DETAILED DESCRIPTION OF PASSIVE DIGITAL PHENOTYPING VARIABLES

Name	Type	Dimensions	Description
A App usage	Numerical	48	Time (s) that the user has used their phone, given in slots of 30 minutes.
S Steps	Numerical	48	Number of steps taken by the user.
D Distance	Numerical	48	Distance (m) traveled, measured from the GPS (includes any kind of transportation).
P Physical distance	Numerical	48	Distance (m) traveled while doing physical activity (running, walking...), computed from device sensors.
H At home	Binary	48	Whether the user was at home* during the time slot.
W At work	Binary	48	Whether the user was at work* during the time slot.
L Light	Numerical	144	Amount of light detected by the phone sensor, used to estimate sleep. Given in slots of 10 minutes.
G Actigraphy	Numerical	144	Data from phone's accelerometer.
U Unlocks	Numerical	144	Number of times the phone is unlocked.
Z Sleep	Numerical	3	Sleep time encoded in <i>duration</i> , <i>start time</i> and <i>end time</i> , all in seconds. Times before 0:00 are negative.
E Emotions	Mixed	3	User's report on their own emotions, encoded in three variables. • <i>Valence</i>: internal prediction by eB2 categorized as "negative", "neutral" or "positive". • <i>Count</i>: number of times the user inputs an emotion in the app. Emotions can be positive (euphoria, motivation, happiness, relaxation), neutral (indifference, other) or negative (sadness, loneliness, stress, anxiety, anger, pain). • <i>Ratio</i>: cumulative sum of those emotions: add +1 if they are positive, 0 if neutral or -1 if negative.
V Activity	Numerical	3	Time (s) spent in walking, running and on a vehicle.
C Circadian data	Numerical	3	Estimation of circadian rhythm through three variables: • <i>Acrophase</i>: peak time of the circadian rhythm, given as seconds that correspond to a time during the day. • <i>MESOR</i>: value where samples oscillate (a rhythm-adjusted mean). • <i>Amplitude</i>: distance between the peak and MESOR values.
R Sport	Binary	1	Flag to indicate whether the user practiced sport or not. Sport includes walking, running or biking during more than 15 minutes (according to Google Fit).
K Weekend	Binary	1	A flag to indicate whether the date is weekend or not (it does not depend on human behavior but may help model the periodic nature of working days and weekends).
Q Quality scores	Numerical 0 to 1	5	Measurement (from 0 to 1) given to some variables regarding the reliability of the data (how confident eB2 is about the data). There are quality scores for app data, steps, distance, actigraphy/light and unlocks.

* Places (home, work) are computed by analyzing location clusters and finding the most probable location for them. Some users may not have their home and/or work locations defined.

DETAILED DESCRIPTION OF STATIC SOCIODEMOGRAPHIC AND CLINICAL VARIABLES

Risk factor	Size of subgroups	Description of variables*
Demographic factors		
Sex	107 · 210	Sex of the subject: man or woman.
Age	90 · 167 · 44	Age of the subject, which was categorized into young (<30), adult (30–55) and elder (>55).
Living with	54 · 234	Who the subject lives with: alone or with family/friends/shared.
Civil status	202 · 114	Single/separated/divorced/widowed or married/living together for over 6 months.
Work status	37 · 64 · 98 · 98 · 13	Student, unemployed, temporal/permanent disability, active or retired.
Studies	33 · 267	The level of studies, separated in low (<10 years) or high (≥10 years).
Clinical factors		
Suicide behavior	<i>Yes · No</i>	
		<i>Columbia-Suicide Severity Rating Scale</i> , which can be used to separate the sample in four different ways (intensity of ideation was given as a scale from 0 to 5, and it was considered positive when it was above 2):
	141 · 84	• Ideation during last month.
	221 · 34	• Ideation along lifetime.
	74 · 245	• Suicide attempts during last month.
	244 · 73	• Suicide attempts along lifetime.
	77 · 232	MINI**: intensity of suicidality above 17, or suicidal behavior disorder.
Depression	60 · 260	<i>Inventory for Depressive Symptomatology (clinician-rated)</i> and <i>Quick Inventory of Depressive Symptomatology (self-reported)</i> , two scales that evaluate depression in the moment of inclusion, assessed by the clinician and the own patient. Values above 36 and 15, respectively, are associated to severe depression [31].
	95 · 100	
	72 · 239	MINI: major depressive disorder or episodes along lifetime, with or without psychosis features.
Anxiety	88 · 118	<i>Generalized Anxiety Disorder-7</i> , scale where patient may be considered to have anxiety for values ≥10 [32].
	9 · 311	MINI: social or generalized anxiety disorders.
Panic	26 · 294	MINI: panic disorder.
Personal factors		
Victim of sexual abuse	<i>Yes · No</i>	
	55 · 116	<i>Childhood Trauma Questionnaire – Short Form</i> , which contains one scale about sexual abuse. Patients with moderate or severe abuse were considered positive (value above 8) [33].
Alcohol abuse	55 · 131	<i>CAGE</i> questionnaire, with two categories indicating the presence of alcohol dependency and/or problems that were jointly used to identify positive cases.
	11 · 309	MINI: alcohol use disorder.

* In some variables, the total number of patients once the sample was divided is less than 320 due to entries with NaN values or categories with too few samples that were not included. Some similar categories were merged into one to facilitate larger subsamples.

** MINI stands for Mini-International Neuropsychiatric Interview, a questionnaire to obtain diagnoses for a list of mental health disorders.