Emotion Forecasting: A Transformer-Based Approach

Leire P. Arbaizar, Jorge Lopez-Castroman, Antonio Artés-Rodríguez, Pablo M. Olmos, David Ramírez

Submitted to: Journal of Medical Internet Research
on: July 05, 2024

Table of Contents

Emotion Forecasting: A Transformer-Based Approach

Leire P. Arbaizar[1] MSc; Jorge Lopez-Castroman[1, 2, 3, 4] MD, PhD; Antonio Artés-Rodríguez[1, 5] PhD; Pablo M. Olmos[1, 5] PhD; David Ramírez[1, 5] PhD

[1]Signal Theory and Communications Department Universidad Carlos III de Madrid Leganés ES
[2]Institut de Génomique Fonctionnelle University of Montpellier CNRS-INSERM Montpellie FR
[3]Department of Psychiatry Nimes University Hospital Nimes FR
[4]Gregorio Marañón Health Research Institute Madrid ES

Corresponding Author:
Leire P. Arbaizar MSc
Signal Theory and Communications Department
Universidad Carlos III de Madrid
Edificio Torres Quevedo
Av. de la Universidad, 30
Leganés
ES

Abstract

Background: Monitoring the emotional states of psychiatric patients has always been challenging due to the non-continuous nature of clinical assessments, the effect of being in a healthcare environment, and the inherent subjectivity of existing evaluation instruments. However, mental states in psychiatric disorders exhibit significant variability over time, making real-time monitoring crucial for preventing risk situations and ensuring appropriate treatment.

Objective: Our objective is to leverage new technologies and deep learning techniques to enable a more objective, real-time monitoring of patients. This will be achieved by passively monitoring variables like step count, patient location, and sleep patterns using mobile devices. We aim to predict patient self-reports and detect sudden variations in their emotional valence, identifying situations that may require clinical intervention.

Methods: Data for this project are registered with the Evidence-Based Behavior (eB2) MindCare mobile application, where both passively and self-reported variables are recorded from patients. We utilize daily summaries of these variables. We implement imputation methods based on hidden Markov model (HMM) to address missing data and transformer deep neural networks for time-series forecasting. Finally, classification algorithms are applied to predict several variables, including emotional state and responses to the Patient Health Questionnaire (PHQ-9).

Results: Through real-time patient monitoring, we demonstrated the ability to accurately predict their emotional state, obtaining an accuracy of 0.93 and 0.98 of receiver operating characteristic (ROC) area under the curve (AUC) for emotional valence classification with an XGBoost classifier and anticipate emotional state changes (ROC AUC of 0.87 for change detection one day in advance). Additionally, we showed the feasibility of forecasting general responses to the PHQ-9 questionnaire. Especially good results were obtained for the score prediction of certain questions. For instance, in the case of question 9, related to suicidal ideation, we obtained an accuracy of 0.9 and ROC AUC of 0.768 in predicting the following day's response.
Secondly, from a methodological perspective, we illustrate the enhanced stability of multivariate time-series forecasting when combining HMM pre-processing with a transformer model, as opposed to other time-series forecasting methods, such as the Recurrent Neural Network or the Long Short- Term Memory cells. Concretely, we exploit the capabilities offered by attention mechanisms to capture longer time dependencies.

Conclusions: From a methodological perspective, we found out that the stability of multivariate time-series forecasting improved when combining hidden Markov model pre-processing with a transformer model, as opposed to other time-series forecasting methods (RNN, LSTM...), leveraging the attention mechanisms to capture longer time dependencies and gain interpretability. We show the potential to assess the emotional state of a patient and the scores of psychiatric questionnaires from passive variables in advance. This offers a real real-time monitoring of patients and hence better risk detection and treatment adjustment.

(JMIR Preprints 05/07/2024:63962)

DOI: https://doi.org/10.2196/preprints.63962

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Emotion Forecasting: A Transformer-Based Approach

Abstract

Background:
Monitoring the emotional states of psychiatric patients has always been challenging due to the non-continuous nature of clinical assessments, the effect of being in a healthcare environment, and the inherent subjectivity of existing evaluation instruments. However, mental states in psychiatric disorders exhibit significant variability over time, making real-time monitoring crucial for preventing risk situations and ensuring appropriate treatment.

Objective:
Our objective is to leverage new technologies and deep learning techniques to enable a more objective, real-time monitoring of patients. This will be achieved by passively monitoring variables like step count, patient location, and sleep patterns using mobile devices. We aim to predict patient self-reports and detect sudden variations in their emotional valence, identifying situations that may require clinical intervention.

Methods:
Data for this project were collected using the Evidence-Based Behavior (eB2) application, which records both passive and self-reported variables daily. Passive data refer to behavioral information gathered via the eB2 app through sensors embedded in mobile devices and wearables. These data were obtained from various studies in which eB2 has participated in collaboration with hospitals and clinics. We use hidden Markov models (HMM) to address missing data and transformer deep neural networks for time-series forecasting. Finally, classification algorithms are applied to predict several variables, including emotional state and responses to the Patient Health Questionnaire (PHQ-9).

Results:
Through real-time patient monitoring, we demonstrated the ability to accurately predict patients' emotional states and anticipate changes over time. Specifically, our approach achieved high accuracy (0.93) and a receiver operating characteristic (ROC) area under the curve (AUC) of 0.98 for emotional valence classification. For predicting emotional state changes one day in advance, we obtained a ROC AUC of 0.87. Furthermore, we demonstrated the feasibility of forecasting responses to the PHQ-9 questionnaire, with particularly strong performance for certain questions. For example, in question 9, related to suicidal ideation, our model achieved an accuracy of 0.9 and a ROC AUC of 0.77 for predicting the next day’s response.
Second, we illustrate the enhanced stability of multivariate time-series forecasting when combining HMM pre-processing with a transformer model, as opposed to other time-series forecasting methods, such as the Recurrent Neural Network or the Long Short-Term Memory cells. Concretely, we exploit the capabilities offered by attention mechanisms to capture longer time dependencies.

Conclusions:
We found out that the stability of multivariate time-series forecasting improved when combining hidden Markov model pre-processing with a transformer model, as opposed to other time-series forecasting methods (RNN, LSTM, etc.), leveraging the attention mechanisms to capture longer time dependencies and gain interpretability. We show the potential to assess the emotional state of a patient and the scores of psychiatric questionnaires from passive variables in advance. This offers a real real-time monitoring of patients and hence better risk detection and treatment adjustment.

Keywords:
Affect; Emotional Valence; Machine Learning; Mental Disorder; Monitoring; Mood; Passive Data; PHQ-9; Psychological distress; Time-series Forecasting.

Introduction

The presence of a specific mood status is a necessary criterion for many psychiatric diagnoses, as outlined in the Diagnostic and Statistical Manual of Mental Disorders [1], and self-perceived mood is a fundamental component of assessing mental states in psychiatry [2, 3]. Therefore, precise monitoring and following of mood conditions play a vital role in mental health care. For instance, both positive and negative mood states, as well as their fluctuations, have demonstrated their predictive value for significant outcomes, such as compulsive overeating episodes in bulimia nervosa, adherence to treatment in bipolar disorder, and opioid use disorders [4, 5]. In recent years, technological advancements have facilitated the real-time tracking of individuals' self-reported mood status. One notable advancement is the utilization of smartphone-delivered Ecological Momentary Assessment (EMA), which allows for the analysis of an individual's experiences, behavior, and emotions as they unfold in their natural environments [6]. However, the effectiveness of this method of mood state evaluation largely depends on the individual's level of self-awareness and ability to interact with the EMA platform. In many cases, psychiatric disorders can cause behavioral changes that decrease the likelihood of individuals engaging with an EMA tool, resulting in missing data. Consequently, a crucial research priority is the development of objective behavioral biomarkers for mood states that can be passively sensed without requiring active involvement from individuals.

By harnessing the power of patients' mobile phones and wearable devices, it has become feasible to gather continuous sensor data in a noninvasive manner, providing valuable insights into their daily activity patterns [7]. These models hold the potential to forecast mental health crises and detect abnormal behavioral patterns, facilitating early intervention [8].

In this study, we examine daily summaries of passively collected behavioural data, treating them as multivariate time series. Passively collected data refer to behavioral information recorded through sensors embedded in devices such as wearables and mobile phones, where patient interaction is not required for data capture. This passive data collection is supplemented by actively provided inputs from patients, such as self-reported emotions or responses to questionnaires tailored to the study context. The dataset utilized in this research was obtained using the eB2 application, a platform developed by our research group and implemented across multiple studies. These studies encompass a variety of patient cohorts, providing extensive behavioral data complemented by responses to psychological, quality-of-life, and nutrition-related questionnaires, depending on the specific objectives of each study.

For time-series analysis, we use transformer models, which are particularly well-suited for capturing long-range dependencies within sequential data [9]. Leveraging the advantages offered by transformer models for time series analysis, we aim to uncover underlying patterns in the data collected over time. The attention mechanism in transformers is a fundamental feature that enables the model to prioritize relevant temporal dependencies within the input sequence. By assigning varying weights to different time samples in the sequence, transformers can effectively capture and integrate the most pertinent information for accurate forecasting [10, 11]. Furthermore, transformer models exhibit the ability to handle sequences of variable lengths. This flexibility renders transformers scalable to datasets containing a large number of time points, accommodating diverse time series lengths without sacrificing performance [12]. These attributes position transformers as a promising option for time series forecasting, as they facilitate the modeling of complex temporal patterns and enhance forecasting accuracy.

Related work

This section covers different aspects related to depression diagnosis and tracking, including research in patient emotional state monitoring and the reliability of the PHQ-9 questionnaire for its diagnosis. Moreover, it explores how researchers are utilizing passive observations to obtain standardized and objective insights into patients' self-reported states. Following this, we introduce works that apply machine learning and deep learning methods to analyze such data for enhanced diagnosis accuracy and longitudinal patient tracking. Lastly, we provide some background on transformer-type attention models for time series forecasting.

Depression Diagnosis and Tracking

As numerous studies have shown [13, 14, 15], depression remains the worldwide leading cause of disability. However, conventional diagnosis and tracking approaches primarily rely on self-reported depressive symptoms in clinical settings, methodologies established over half a century ago. These methods typically entail survey completion or face-to-face interviews, offering limited accuracy, ecological validity, and reliability while imposing significant costs for monitoring and scalability [16]. Moreover, the subjective nature of patient and clinician evaluations, combined with the fluctuating nature of mental health conditions over time, emphasizes the need for ongoing, longitudinal assessments to accurately capture these nuances effectively. In terms of assessing depression, the PHQ-9 questionnaire has shown to be reliable for the criteria-based diagnosis of this disorder, alongside giving a valid measure of depression severity. These characteristics plus its brevity make the PHQ-9 a useful clinical and research tool [17].

A set of studies focuses on the relationship between mood variability and some psychiatric disorders. In this work, we used mood in the sense of the subjective variation of the patient's emotional state as expressed by the patient [18]. Research indicates that mood variability including hypomania, cyclothymia, and hyperthymia have been described in 40–50% of patients with depression and that such variability could also characterize anxiety disorders [19, 20]. Over the past two decades, there has been a surge in research linking various patterns of short-term emotional change to adaptive or abnormal psychological functioning, often with conflicting results [21]. Psychiatric decompensations are characterized by specific patterns of emotional fluctuations across time and provide insight into what constitutes optimal and sub-optimal emotional functioning.

Advances in Ecological Momentary Assessment (EMA) and Passive Monitoring

To prevent bias and capture changes in behavior over time and across different contexts, several researchers propose a shift away from relying solely on global retrospective self-reports collected during research or clinic visits in clinical psychology assessment [22]. In Ecological Momentary Assessment (EMA), data is repeatedly collected on subjects' current behaviors and experiences as they happen in their everyday environments [6, 23]. EMA helps reduce memory bias, provides more accurate insights into daily life, and allows for the study of small-scale influences on behavior in real-world settings. Technologies used for EMA range from traditional written diaries and phone calls to electronic diaries and physiological sensors [22]. In this case, an EMA-style monitoring is conducted to rely solely on passively collected variables through mobile phone sensors and wearables, excluding all the patient-reported data. Several studies that reviewed the passive follow-up of patients with different conditions, including bipolar disorder, schizophrenia, and depression, highlighted the potential of passive biomarkers for the monitoring of different types of disorders. Particularly, variables such as

accelerometry, location, audio, and usage data showed a high general performance [24]. Other studies explored the detection of daily-life behavioral markers through mobile phone global positioning systems (GPS) and usage sensors. They showed that features extracted from these sensors provided markers strongly related to depressive symptom severity [21]. In this line, one of the variables that has shown a correlation with the individual's mental state is daily activity [25, 26]. A study focused on detecting emotional state instabilities through passive data, found that three weeks of continuous, passive recordings were enough to reliably predict mood changes, obtaining average and median errors of Mood Instability Scores (MIS) within a margin of 5% [27].

Machine Learning and Deep Learning Approaches for Behavioral Data Analysis

Due to the potential demonstrated by these variables in predicting emotional states and their variability in patients, previous works have focused on the application of machine learning and deep learning algorithms to analyze this data [28]. A study by Ghandeharioun et al. [13] applied machine learning methods to data on sleep behavior, motion, phone-based communication, location changes, and phone usage patterns, to impute missing clinical scores from self-reported measures and predict depression severity from these continuous sensor measurements. Similarly, other studies evaluated the performance of random forest and support vector machine classifiers for binary classification of the PHQ-9 score, resulting in 60.1% and 59.1% accuracy, respectively, demonstrating a proof of concept for the detection of depression from passive features [29]. Following this approach, diagnostic meta-analyses have demonstrated the effectiveness of the PHQ-9 for depression screening using mobile devices through various machine learning techniques [30].

Recent studies [30, 31] have focused on longitudinal monitoring of patients highlighting the importance of continuous follow-up and the exploration of temporal behavioral patterns. Aiming to address the lack of clarity on the temporal scale, specificity, and person-specific nature of the associations between smartphone data and affective symptoms, a study was conducted on smartphone-based passive sensing to identify within- and between-person digital markers of depression and anxiety symptoms over time [31]. Here, hierarchical linear regression models and temporal windows were used to understand the time scale at which sensed features relate to mental health symptoms and explore the predictions in the distal, medial, and proximal times. In line with this, other studies employed multilevel modeling to examine the relationships between daily mood and mood variability with symptoms of depression, generalized anxiety, and social anxiety, to confirm the empirical evidence linking EMA of mood variability with psychiatric disorders [33]. The findings showed both common and specific emotional dynamics that defined the severity of affective symptoms.

Our research aligned with previous studies by aiming to predict self-reported emotional states and their fluctuations, as well as PHQ-9 questionnaire scores. However, our methodology diverges from conventional approaches by operating within the natural daily routines of patients, not in an experimental setup. Besides, our sample included patients with a variety of disorders who were expected, but not compelled, to actively report data via the application. This setup introduced challenges, notably the substantial presence of missing data, which we addressed using a hidden Markov model, as documented in prior literature [34]. We build upon this paper, which predicted emotional valence from passive variables, employing HMM to handle missing values and classification methods. With the proposed model, we enhanced the emotional valence prediction achieved in this paper and obtained a more reliable prediction as the time horizon expanded. Additionally, we included the prediction of the scores of the PHQ-9 questions.

To leverage the continuous acquisition of passive variables and work with these temporal data sequences, this study adopted a transformer-based approach. Transformer models, based on attention mechanisms, offer several advantages for time series forecasting. The transformer architecture, as introduced in the paper by Vaswani et al. in 2017 [10], excels at capturing long-range dependencies, making it well-suited for time series data where distant historical information can be crucial. Attention mechanisms within transformers enable contextual understanding, allowing the model to weigh the significance of past elements, thereby improving forecasting accuracy [10, 12].

In addition to the previously mentioned references, several publications delved into the application of transformers to multivariate time-series data [12, 35, 36], showcasing the flexibility and adaptability of the transformer architecture in capturing complex temporal relationships. Initially developed for natural language processing tasks, transformers have demonstrated a seamless transition to time-series forecasting due to their inherent capability to model temporal data. This shift underscores the versatility and robustness of transformer-based models in addressing diverse sequential data tasks beyond language processing. Furthermore, pre-trained transformer models can be fine-tuned for specific forecasting tasks, leveraging insights from diverse datasets, as demonstrated in various transfer learning applications. Finally, attention weights also contribute to their utility in analyzing and forecasting time series data in a more interpretable way [37].

To the best of our knowledge, no prior studies have utilized transformer models specifically for emotion recognition relying solely on behavioral (passive) data from mobile devices. This represents a novel direction in leveraging behavioral data for emotion classification and change detection, particularly in real-world, non-invasive contexts.

In other domains, emotion recognition has been explored using ransformer models applied to various data modalities. For text-based emotion recognition, transformer models such as BERT and GPT proved to be effective in capturing emotional nuances in textual data. For instance, Xie et al. [49] used GPT to encode dialogue features, while Zaidi et al. [51] combined RoBERTa embeddings with wav2vec 2.0 for cross-modal emotion recognition.

For audio-based emotion recognition, speech data have been modeled using architectures like wav2vec 2.0. Luna-Jiménez et al. [48] fine-tuned xlsr-wav2vec2.0 for detailed speech emotion analysis, while Sun et al. [52] combined it with BERT for multimodal integration.

Transformers have also been employed for visual-based emotion recognition, where facial expressions and gestures were analyzed. Huang et al. [47] utilized the multi-head attention mechanism to fuse visual and audio features, effectively capturing both spatial and temporal dynamics.

Multimodal emotion recognition, which combines text, audio, and visual data, has been a common approach. Xie et al. [49] proposed a Crossmodality Fusion Transformer, and Zhao et al. [53] introduced MEmoBERT for cross-modal emotion classification tasks.

In the field of physiological emotion recognition, Transformer-based models like MATS2L [50] and Conformer [54] have demonstrated high accuracy in analyzing EEG and ECG signals for emotion classification.

In contrast to these approaches, our work focused exclusively on behavioral data collected passively through smartphones and wearable devices, without relying on more invasive techniques such as video, voice, or physiological signals like EEG. This distinction offered significant advantages. By aligning with natural patient behavior and environments, our method reduced intrusiveness, ensured scalability, and facilitated seamless integration into daily life, making it particularly suitable for real-world applications.

Objectives

The general goal is to acquire objective indicators of patients' conditions and fluctuations in their emotional well-being from passive biomarkers. For this, we focus on predicting the emotional valence of patients as well as the PHQ-9 score. This approach aims to address the challenge of subjectivity and the absence of continuous monitoring in psychiatry, ultimately aiding in the identification of potentially risky situations. This would facilitate timely intervention and treatment adaptation, thus improving the quality of life of the patients and their environment. We strive to achieve this prediction several days in advance of the actual event. Moreover, the incorporation of attention mechanisms serves the additional purpose of advancing our understanding of behavioral patterns.

Methods

Recruitment. Patient Inclusion and Exclusion Criteria and Indications

Participants were eligible for inclusion in the study if they were at least 18 years old and diagnosed as clinical outpatients with mental disorders by specialists, or if they were attending therapy groups. Among these groups, there were three main categories: High suicidal risk, eating disorder, and common mental disorder (CMD). CMD encompasses a group of distress states manifesting with anxiety, depression, and unexplained somatic symptoms typically encountered in community and primary care settings [38]. Furthermore, patients from other studies, including cohorts of patients affected by cancer, HIV, obstructive sleep apnea, and cardiac conditions, as well as control patients, were included.

Participants were required to own a smartphone running on Android or iOS operating systems, which they connected to a Wi-Fi network at least once per week. Only participants who provided written informed consent for the eB2 study were included.

Patients received instructions from the clinicians at the beginning of follow-up. For most cohorts, only information regarding the application's functionality was provided. However, two groups received specific guidance: patients with eating disorders were required to compulsorily complete meal entries, while patients with CMD were encouraged to regularly log their emotions and periodically complete the PHQ-9 questionnaire. There was no obligation or a set number of required responses for these tasks.

Passive data sources included both mobile and wearable devices. In certain studies (HIV, cancer, and obstructive sleep apnea) wearables were provided to patients. For the remaining studies, if patients had their own wearable, information was extracted from the most reliable data source.

Data

This study was conducted on a sample of 4403 patients from 8 distinct cohorts, each characterized by a different pathology or condition. The patient cohorts included individuals with Common Mental Disorder (CMD) (1785, 40.96%), eating disorder (1477, 33.89%), High Suicidal Risk (413, 9.48%), cancer (84, 1.92%), obstructive sleep apnea (48, 1.1%), HIV (24, 0.55%), cardiology-related conditions (20, 0.46%), as well as control subjects (507, 11.63%).

The data for this study is derived from secondary analyses of multiple clinical studies conducted in collaboration with various hospitals. All studies were designed to monitor patients with distinct health conditions using the eB2 platform. Despite differences in data sources, standardized protocols for data collection were applied across all studies to ensure consistency in data quality. Each clinical study received approval from the relevant Institutional Review Board (IRB) in compliance with ethical standards and the Declaration of Helsinki. IRB approval numbers are indicated in brackets and correspond to the center where approval was obtained for each project and country.

Patients at high risk of suicide were identified through collaborations with the Jiménez Díaz Foundation (FJD, EC005-21), Montpellier University Hospital (CPP Ouest IV 20/18_2), and Clínica Nuestra Señora de la Paz. Patients with common mental disorders were recruited from FJD (PIC148-22), while those with eating disorders were monitored at specialized mental health centers, including Adalmed and ITA clinics. The study also includes cancer patients monitored in partnership with Gregorio Marañón Hospital (EB2COLON2023), CNIO, and Fuenlabrada Hospital; HIV patients from Gregorio Marañón (MICRO.HGUGM.2022-002); cardiology patients from Clínico San Carlos Hospital (19/239-O_P); and patients with obstructive sleep apnea monitored at FJD (PIC163-22). Informed consent was obtained from every participant at the time of inclusion, ensuring adherence to ethical guidelines and participant rights.

All users were Spanish and French. Among them, 57.48% (2531 out of 4403) were female, 40.93% (1802 out of 4403) were male, and gender information was missing for the remaining 1.59% (70 out of 4403). All age groups were well-represented, with a mean age of 46 years (ranging from 18 to 77 years) at the start of the measurement period.

Patient monitoring was conducted using the eB2 MindCare application [39, 40]. The eB2 application operates by harnessing information from diverse sources within the patient's ecosystem. Utilizing phone sensors, Google Fit, and wearable devices, it acquires data at varying intervals, facilitating a nuanced understanding of the patient's daily activities. In parallel with this passive monitoring, patients had the option to input subjective experiences, sleep patterns, and emotional states throughout the day. Emotional states are cataloged within the application, offering 20 options in a spectrum from anger to delight [41, 42].

Daily summaries of this data were the primary focus, although alternative granularities were also considered (hours, minutes…). Therefore, when predicting sequences or emotions at the next temporal moment, this temporal interval is in days. The data collection period for this study spans 8 years, from 2016 to 2023, and the average duration of passive activity sequences for patients is 224 days, with a standard deviation of 200 days.

The data utilized as inputs for training the models were daily summaries of the following passive variables: step count, covered distance, sleep hours, app usage, time at home, number of visited places, and practiced sport, a binary variable indicating whether the patient practiced sports during the day. The targets we aimed to predict from these passive data were patients' emotional valence and PHQ-9 scores. The ground truth values for these targets come from self-reported emotions and questionnaire answers that patients complete through the application. A more detailed description of the different cohorts is included in Multimedia Appendix 1.

Preprocessing

To ensure the variables were sampled at the same frequency, daily summaries were created for each variable. These daily summaries were derived by aggregating the data according to the specific variable.

Sleep hours, distance, steps, and app usage were calculated using passive data collection methods from various sources. Sleep hours were determined based on a prioritization hierarchy: user-entered data is the most prioritized, followed by wearable devices, mobile data, and lastly, a sleep estimation model. Daily distance was calculated using GPS signals from mobile devices, collected every 5 minutes. Step counts were gathered at intervals of 1–5 minutes, depending on the provider. To calculate daily totals, step data from each slot was merged by priority (wearables over mobile). App usage was recorded every 5 minutes, capturing which apps were used and the duration of their use. For these four variables, once the data had been collected and selected from the priority sources, the data were summed to obtain the daily summary of each variable.

To identify home and work locations, the DenStream algorithm was utilized. Location data was collected over a 15-day period to form clusters. Once the clusters were established, incoming location data was tagged in real-time as either home, work, or other. The cluster definitions were updated every 30 days following the initial 15-day period. Based on this clustering process, the time a user spent at home was calculated. Data was collected every 5 minutes and tagged upon entry to indicate the cluster corresponding to "home." The total time at home was measured in seconds and aggregated for daily summaries.

The variable practiced sport was a boolean value indicating whether a user engaged in physical activity lasting at least 15 minutes during the day. This variable was updated whenever new physical activity data was received. Physical activity detection was based on three primary sources: (1) activities automatically labeled by devices (e.g., mobile phones, wearables), (2) manually logged activities by users through the provider's app, and (3) activities logged directly into the eB2 MindCare platform.

To address the variations between sensors and data formats, which resulted in anomalies and noise in the information, a preprocessing stage was carried out. This included removing negative values, thresholding the time-related variables to 24 hours, the time step count to 30,000 per day, and the distance to 500 km. Finally, data were standardized over all the patients' sequences (0 mean and SD 1 for input features). Standardization was performed to ensure all variables had a uniform scale, which improves the efficiency and performance of machine learning algorithms. This is particularly important for models sensitive to differences in feature scales, such as neural networks or distance-based methods. Figure 1 shows the sequence of data preprocessing.

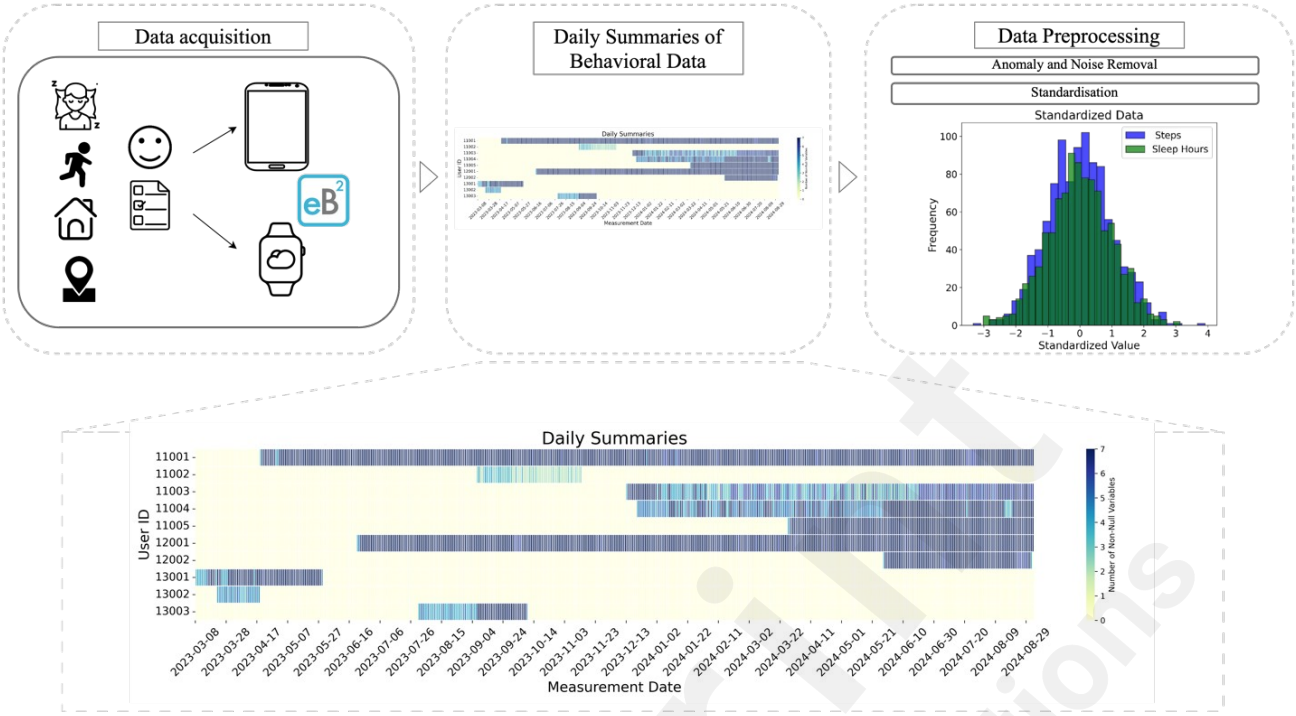


Figure 1. Data preprocessing pipeline: data acquisition, obtention of the daily summaries, and standardization of behavioral data. The sequence shown in the daily summaries displays the temporal sequence of passive data for five different patients. The intensity of the lines indicates the amount of non-missing behavioral data the patient has for that day.

Regarding passive data, the mean percentage of missing data was approximately 60%, with step count (52.6% of total missing) having the fewest missing values, and time at home being the least complete (69.4% total missing). Table 1 shows the percentage of missing values per passive variable grouped by year.

Table 1. Percentage of missing daily data for each passive variable by year.

	Passive Variables							
Year	**Steps**	**Distance**	**Sleep**	**App usage**	**Time home**	**Location clusters**	**Emotions ratio**	**PHQ-9**
2018	60.42	39.64	70.82	90.21	49.18	44.76		
2019	50.32	39.64	74.97	78.45	50.71	48.39		
2020	59.10	54.43	78.14	75.75	65.82	64.40		
2021	60.09	73.48	56.36	52.30	77.99	77.67		
2022	49.89	78.85	53.48	58.16	80.43	79.93		
2023	46.85	77.18	51.84	55.01	78.85	78.47		
Total	52.61	63.43	64.41	67.66	69.47	68.15	96.34	98.00

Regarding targets, Table 1 shows the percentage of missing emotion ratio and PHQ-9 answers in comparison with the passive data. It can be observed that the variables reflecting the patients' mood (self-reported emotions and PHQ-9 answers), which they must enter actively, had the lowest percentage of registered data. This is why we aimed to predict these two variables using passive data, as they were more complete and allowed for continuous monitoring of the patient's condition without relying on the patients to input the data.

In this study, we handled emotions following Russell's classification scheme, which characterizes emotions in a two-dimensional space [43]. James Russell's [56] circumplex model proposed that emotions could be understood along two independent and bipolar dimensions: valence (pleasantness/unpleasantness) and activation (high arousal/low arousal). Independence implied that valence and activation were uncorrelated, while bipolarity suggested that opposite emotions lied at opposite poles of each dimension. For example, happiness and sadness represented opposite ends of the pleasantness spectrum, whereas emotions like tense and sleepy lied at opposite extremes of the activation dimension.

Using this framework, emotions reported by patients were assigned a level of valence as positive (pleasant), negative (unpleasant), or neutral. For instance, positive valence emotions included happiness, enthusiasm, or satisfaction, while negative valence emotions included sadness, anger, or frustration. Neutral valence emotions fell in the center of the pleasantness spectrum, representing a lack of strong affective polarity [57]. Related studies [58] similarly relied on the model proposed by Russell to categorize emotions, either in 1D using valence or in 2D with both valence and arousal. From this scheme, we focused on emotional valence, which was the first prediction target.

The daily valence was determined by the difference between the counts of positive and negative emotions and can take values between 0 and 2 (negative, neutral, and positive valence).

Additionally, the Patient Health Questionnaire (PHQ-9), comprising the 9-item depression module extracted from the full PHQ, was examined as another target variable. According to this questionnaire, diagnosis of major depression is established if "more than half the days" over the past 2 weeks exhibited the presence of 5 or more depressive symptom criteria, with one of these symptoms being either depressed mood or anhedonia [17]. The responses to the questionnaire were derived from the cohort of CMD patients, comprising a total of 597 completed surveys. In the context of this investigation, each of the 9 questionnaire items was treated independently. This approach was adopted due to the diverse nature of the questions, which collectively encompassed various facets of the patients' daily experiences. The simultaneous prediction of all scores posed a considerable methodological challenge in this initial endeavor.

Originally, this score comprised four classes. However, due to the insufficient number of answers for each type and the frequency-based nature of the responses (ranging from "not at all" to "nearly every day"), the intermediate classes sometimes merged with the two extreme groups. Consequently, we decided to classify the answers into two broader classes: 0 (representing "not at all" and "several days" as low frequency) and 1 (representing "more than half of the days" and "nearly every day" as high frequency).

Models

Our global model, illustrated in Figure 2, consists of 3 main sub-models that are trained separately. Hence, we can consider that the global model is trained in three stages: 1) training of the hidden Markov model to deal with missing fields, 2) the autoregressive transformer to

pre-train the model to capture the temporal structure of the data, and 3) the classification model itself.

Initially, the raw data undergo the preprocessing and quality control mentioned previously. To prevent data leakage between the different training phases and biases, the original dataset was divided into three subsets for training and validation purposes, with 30%, 40%, and 30% of the data allocated to the hidden Markov model, transformer, and classification layers, respectively, as shown in Figure 2. The partitions are made by assigning equitable percentages of each cohort to each subset to mitigate biases, assigning unique different users to each of them. The following sections describe in detail the three main stages.

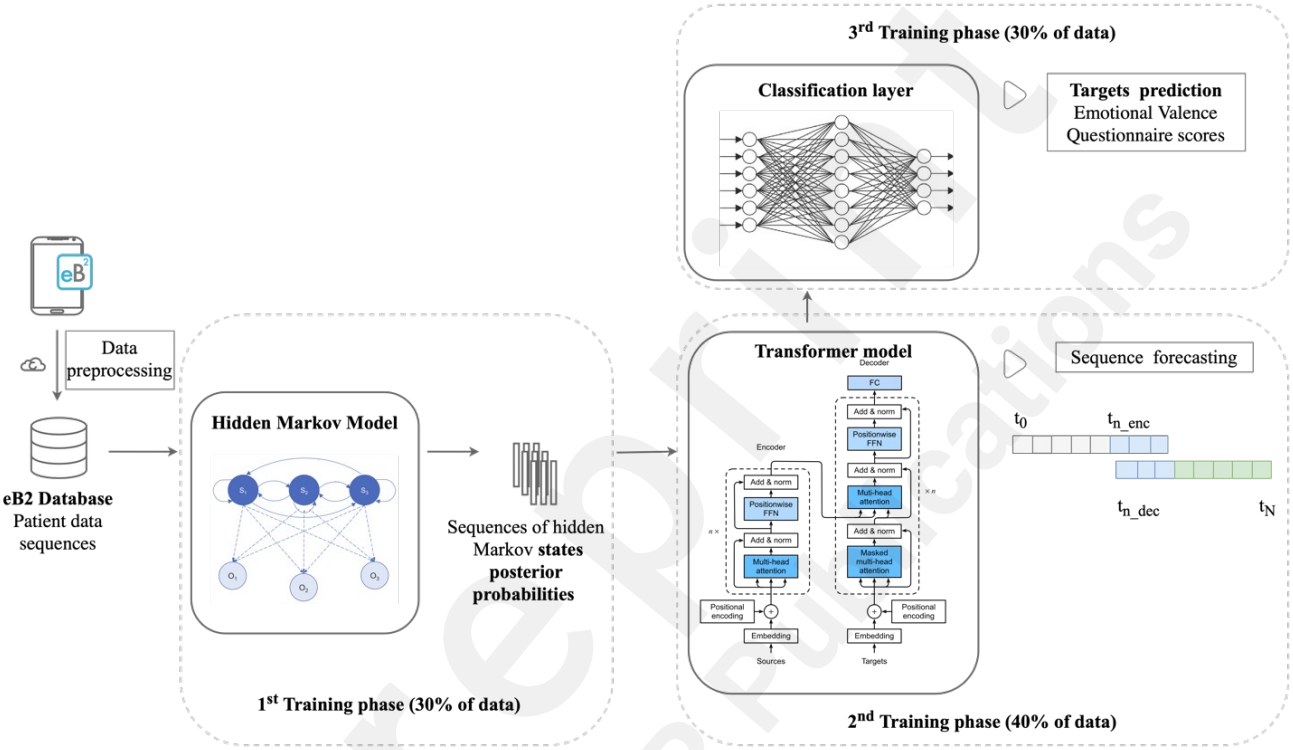


Figure 2. The graphical abstract of the proposed scheme illustrates the underlying architecture, comprising three primary blocks. The first block involves the utilization of Hidden Markov Models (HMM) to address missing data and extract posterior state probabilities. The second block employs a transformer model equipped with an attention mechanism to facilitate pattern recognition and time-series forecasting. Lastly, the scheme incorporates a final classification layer responsible for patient-reported variables, namely Emotional Valence or PHQ-9 score.

Probabilistic Generative Model for Dealing with Missing Data: Hidden Markov Models

Our objective was to evaluate the efficacy of transformer models in enhancing time-series forecasting. However, transformer models do not inherently accommodate missing data, presenting a challenge for our analysis. To address this limitation, we employed hidden Markov models (HMMs). Additionally, we leveraged the latent space representation provided by HMMs, utilizing the posterior probabilities of the hidden states for training the transformer model.

HMMs, commonly employed in time-series analysis, represent a temporal variant of Markov models (MM) [44, 45]. These generative models are characterized by a collection of observable variables and a notable advantage lies in their ability to manage missing data without necessitating prior imputation, achieved through marginalization. In the HMM model, a sequence of observable variables O is generated by a corresponding sequence of internal hidden states S. However, these hidden states are not observed directly. Instead, transitions

between hidden states follow the assumption of a first-order Markov chain. This transition process is defined by a start probability vector π and a transition probability matrix A. Additionally, each observable emission is associated with a probability distribution, conditioned on the current hidden state. These emission probabilities are specified by parameters B. Together, these parameters λ = (π, A, B) fully define the HMM.

The dataset utilized in this study exhibited heterogeneity, encompassing both categorical variables such as practiced sports, as well as continuous variables presumed to take real values (Table 1). For this purpose, we employed the heterogenous-HMMs (HHMMs) implementation from the PyHMM library, which facilitates the utilization of various distributions to manage the emission probabilities of each feature type as depicted in Figure 3 [46, 34].

In this initial phase, the hidden Markov model (HMM) was trained using 30% of the available data, and, once trained, the model was used to infer the state posterior probabilities. At each time step, the set of passive data for a given day and patient was represented by a hidden state posterior probability vector, with a total of 7 hidden states in this instance. Previous studies [34] tested different hidden state configurations, and 7 hidden components were found to effectively capture the underlying patterns in the data. The optimal number of hidden states was determined using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) on a randomly selected subset of sequences with varying levels of missingness [55]. This number of hidden states also led to the best results when a classifier was applied to predict emotions.

The hidden states' posterior probabilities refer to the probabilities of the hidden states given the observed data. These probabilities are calculated using the observations and the model parameters. The sequence of hidden states probability vectors served as the embedding of the daily information for each patient, enabling the training of subsequent models.

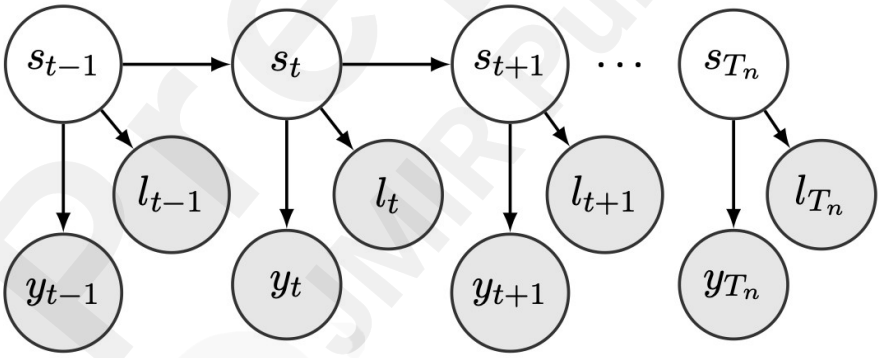


Figure 3. Architecture of Heterogeneous hidden Markov model. Model is described by their hidden states sequence ($s_{0:Tn}$) and continuous observations sequence ($y_{0:Tn}$), discrete observations sequence ($l_{0:Tn}$) [44]. In our case ($l_{0:Tn}$) corresponded to the sequence of discrete observation: practiced sport and ($y_{0:Tn}$) was the sequence of continuous observations: steps, location distances, sleep time, app usage, time home and location clusters count.

Sequence Forecasting with Transformer Model

Due to the limited data on emotions and PHQ-9 responses, to obtain a more informative representation of the time series, we first performed a phase of self-supervised training, following a forecasting approach. To achieve this, we employed a transformer model for time-series forecasting, leveraging its strengths in handling sequential data. Transformers have shown promising results in capturing long-range dependencies and can provide interpretability regarding the most relevant parts of the sequence for each specific task. Our

transformer followed the basic encoder-decoder structure, as depicted in Figure 4, incorporating elements from the encoder of the Informer model [12].
The encoder within the transformer architecture comprised a series of 3 layers. Each layer incorporated two distinct sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Following each sub-layer was a residual connection and layer normalization. Likewise, the decoder consisted of 3 layers. In addition to the two sub-layers present in each encoder layer, the decoder incorporated a third sub-layer responsible for performing cross-attention over the output generated by the encoder stack. Residual connections and layer normalization were similarly applied around each sub-layer to the encoder. To prevent the decoder from attending to subsequent positions, modifications were made to the self-attention sub-layer within the decoder stack, implementing masking: causal (masked) self-attention.
Through the encoder, we obtained contextual information, and with the decoder, we performed the prediction of future observations, considering both the contextual information and the current observations. To incorporate information about the relative or absolute position of tokens in the time series we combined a positional encoding with the input embeddings at the beginning of both the encoder and decoder stacks [10].

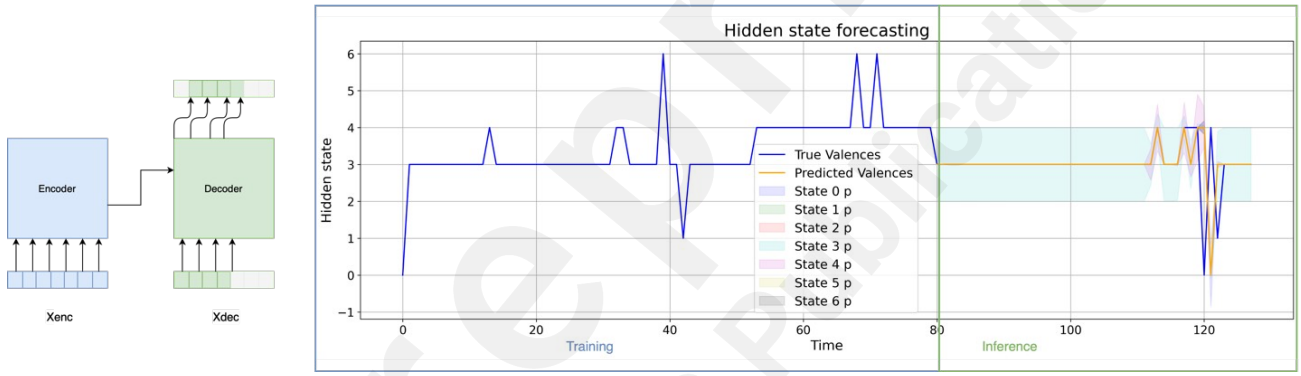


Figure 4. On the left side of the image, the simplified encoder-decoder architecture of the transformer model is shown [10]. On the right side, an example of model training and inference is illustrated. During training, the model learns the parameters to forecast the future state, using the real data as a reference. During inference, the transformer's decoder is responsible for forecasting the future state value in an autoregressive manner. The blue line shows the true states, and the orange line shows those predicted by the model. During inference, the possible future states have associated probabilities, which are illustrated in the graph with different colored margins around the orange line.

The transformer model was trained from scratch in a forecasting paradigm, wherein the transformer's output was compared to the actual output shifted one day ahead. This approach enabled the model to predict future time points with precision. Since our dataset consisted of real-world data, it was uncommon for the patients to have extensive sequences of data. Because of this limitation, for model training, we experimented with different sequence lengths, aiming to strike a balance between model performance and sequence duration. Through experimentation with various sequence lengths for both the encoder and decoder, we determined that a sufficiently large sequence (between 25 and 30 days) was required to adequately capture relationships and facilitate accurate forecasting.
For training we used diverse training schedules, integrating early stopping and dropout techniques to mitigate data overfitting. The loss functions we tested included both mean

squared error (MSE) and mean absolute error (MAE), computed between the predicted and actual future sequences. Optimization was carried out using the Adam optimizer across 80 epochs.

A hyperparameter grid was defined, including values for model dimensions (32, 64, 128), number of attention heads (4, 8, 16), number of encoder and decoder layers (2, 3, 4, 6), feedforward dimensions (128, 256, 512), dropout rate (0.3, 0.5), and learning rate (0.01, 0.001, 0.0005).

The best-performing model, which we present in our results, was trained using sequences spanning 50 days of passive data, with 30 days allocated for the encoder and 20 days for the decoder. During inference, using the preceding 30 days of collected data was sufficient for forecasting, since the model could then predict further into the future autoregressively. The optimal model configuration was achieved with an embedding dimension of 32, 4 attention heads, 3 layers, a feedforward dimension of 128, a dropout rate of 0.3, and a learning rate of 0.001. The results obtained for the different combinations of hyperparameters, as well as a more detailed explanation of the employed architecture, are included in Multimedia Appendix 2.

Emotional State and PHQ-9 Score Classification

For this section, we employed the output of the transformer model to train the subsequent classification layers, aimed at predicting specific targets. The objective was to predict emotional valence and the PHQ-9 scores for the following day.

For the prediction of emotional states, the HMM was first applied to the data sequences to obtain the posterior probabilities of the hidden states of the Markov model, which represented the embeddings of the passive data sequences. Subsequently, forecasting for the following days was conducted using the transformer, trained in the previous phase. These outputs were then utilized to train a model for emotional valence classification.

Several classification models were experimented with, including multilayer perceptron (MLP), ensemble models (EM), support vector classifier (SVC) and extreme gradient boosting (XGB). Among these, random forest classifiers (RF), as part of the EM, and XGB demonstrated superior performance. This constituted the third training phase, which we evaluated through F1-score, accuracy, precision-recall (PR) curves, and ROC curves, along with their respective AUCs. For each input day, these models provided a probability distribution of the predicted emotional valence for the following day, which in this case was a probability distribution among 3 possible outcomes (0, 1, and 2). In this regard, we analyzed both the probability distributions of the valence predicted by the model and the concrete valence estimated for the following day (the one with the highest probability).

As for the PHQ-9 score, to predict the binary score for each of the questions we compared several classifiers by training them with temporal windows of 15, 7, and 3 days using the transformer output (the contextual embeddings), the emotions predicted by the classifier or a combination of the hidden states' posterior probabilities and the predicted emotions.

Results

Time Series Forecasting

In this section, we first compare the time-series forecasting capabilities of the transformer and the HMM, without considering specific target predictions, solely as a comparison of their ability to capture the temporal structure and prediction performance of both models.

For the forecasting of the sequences of the posterior probabilities of the hidden states, which represent the information of the passive observations, our results demonstrate enhanced

stability over time when comparing the transformer model to the HMM model. Specifically, when both models perform autoregressive forecasting, predictions made by the transformer model exhibit greater similarity to the true values as we project further into the future beyond the last observed day.

Figure 5 presents the accuracy of state forecasting, measured as the rate of correctly predicted states for 0-7 days into the future. While both models performed similarly for immediate next-day predictions, the transformer's predictions decayed more gradually than those of the HMM model as the forecasting horizon extended. The chosen configuration for the transformer model involved training with sequences of 50 days -30 for the encoder and 20 for the decoder- during the initial training phase, encompassing all potential patient sequences. Training with shorter sequences yielded inferior forecasting performance.

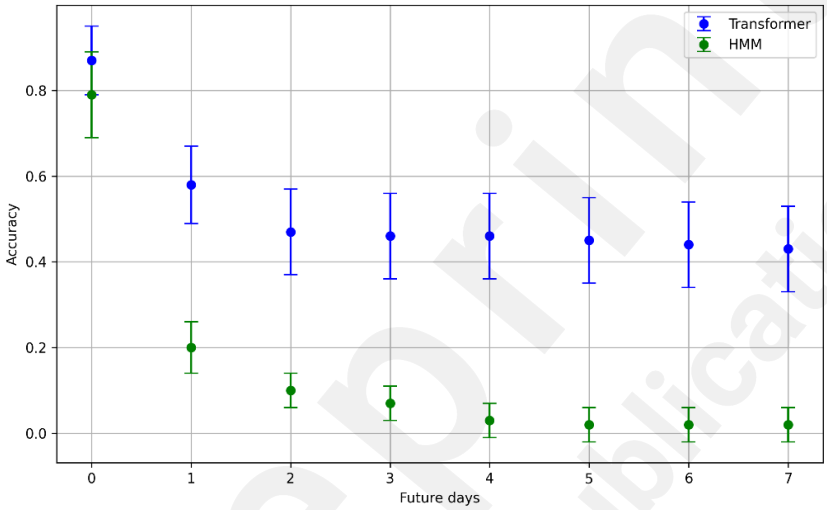


Figure 5: Model comparison in state matches forecasting.

Emotional Valence Forecasting

In terms of emotional valence forecasting, our results focus on predicting emotional state one day ahead. The most effective approach involves employing the XGBoost algorithm with a 7-day window of the transformer model decoder as input for the classification layers.

Alternative approaches were explored for the classification. One included incorporating model-embedded vectors of the hidden states, not the final output of the decoder, to potentially enhance informativeness, yet this yielded comparable results to only using the transformer's decoder output. Another strategy involved retraining the entire model (including the transformer and classifier) or some of the layers via fine-tuning, as opposed to solely training the classifier while keeping the transformer parameters frozen, to improve task performance. However, the results were similar, and given the time and resource requirements for complete retraining, we opted to train the classification layers for the specific task.

Results across different models used for classification exhibited minor variations, potentially attributed to the embedded vectors effectively capturing necessary information for forecasting, irrespective of the classification model's complexity.

Table 2 shows the results for valence classification. Notably, the XGBoost model performed best with a ROC AUC of 0.982 and an accuracy of 0.93. Of particular interest is the discrimination of class number 2 (representing a neutral state), which is challenging to classify due to its scarcity. Despite a slight bias toward negative valence, models achieved robust discrimination for

negative and positive states, and notably good results for the neutral state, one day in advance of the actual event.

Table 2. Metrics for Emotion Classification Models. [a]

	MLP	RF	XGB	XGB-Retrained	XGB-Embedding
Precision	0.86	0.87	0.89	*0.89*	0.84
Recall	0.82	*0.86*	*0.86*	0.85	0.8
F1-Score	0.84	*0.87*	*0.87*	0.86	0.8
Accuracy	0.91	0.93	0.93	*0.94*	0.9
AUC	0.97	0.98	*0.98*	0.98	0.92
PR	0.89	0.92	*0.93*	0.9	0.88

[a] For each metric, the results obtained with the best model are highlighted italics.

Figure 6 illustrates individual patients' emotional series, where the blue line represents the actual emotion for the current day, the orange line represents the forecasted emotion for the following day, and the circles represent the probabilities predicted by the classification model for each emotional valence for the next day. The position and size of the circles correspond to their probability values, with higher probabilities resulting in larger circles. Purple indicates valence 0 (negative), green indicates valence 1 (neutral), and red indicates valence 2 (positive). Regarding emotional valence change detection, analyzing the emotional valence distributions reveals a pattern. During stable periods, the distributions are skewed, meaning that most of the probability mass is concentrated around a single valence outcome. However, near a change, the distributions become closer to a uniform distribution. That is, the probability mass is distributed more evenly among the possible outcomes.

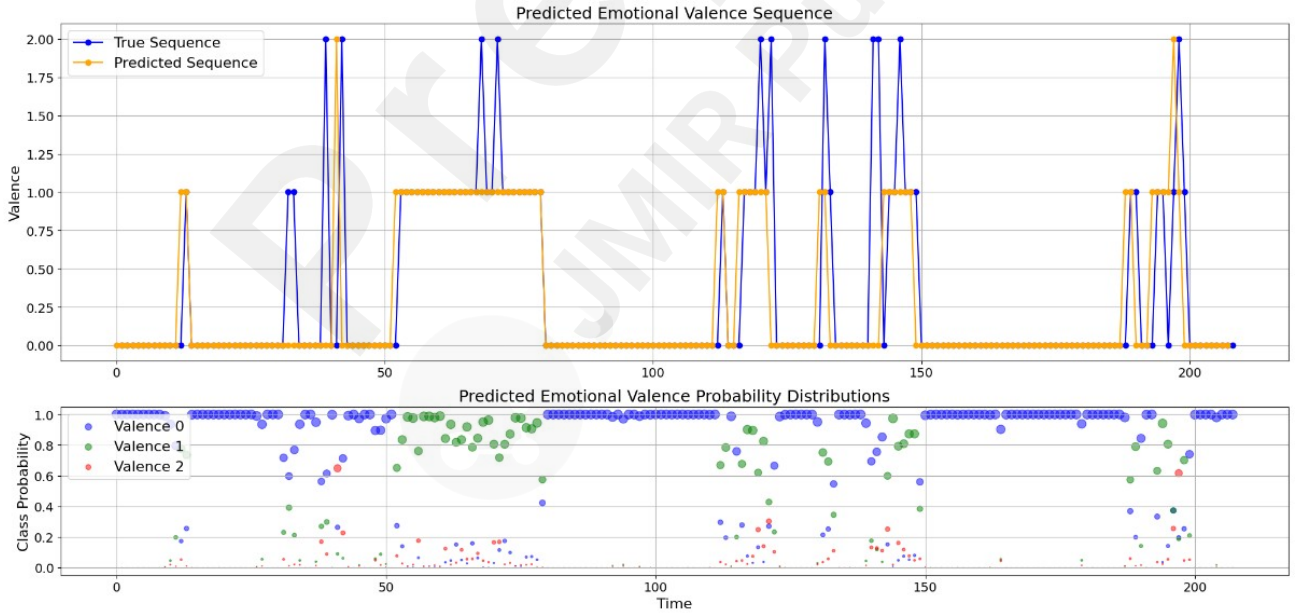


Figure 6. Sequences of real emotions (blue) and those predicted by the model (orange) for the following day. The colored circles denote the probability distribution of the emotional state, purple indicates state 0 (negative valence), green indicates state 1 (neutral), and red indicates state 2 (positive valence). These figures pertain to two patients (ids: 374 and 218) selected from a set of patients with high variability of emotion in their temporal sequence.

To detect the change, we considered calculating the entropy and the Jensen-Shannon (JS)

divergence within a temporal window of the sequence. We calculated the entropy of the valences' probability distributions (Equation 1) for a 3-day window, including the 3 previous days to each event. The Jensen-Shannon divergence (Equation 2) was computed between the present-day distribution and an average distribution over a temporal window, to determine which approach could better determine the shift in emotional state: whether an increment in disorder within the window or a comparison between the actual and previous distribution.

$$H(x) = -\sum p_i \log_2 p_i \quad \textbf{(1)}$$

$$D_{JS}(P||Q) = \tfrac{1}{2} \sum (p_i \log_2 p_i/q_i + q_i \log_2 q_i/p_i) \quad \textbf{(2)}$$

In Figure 7, we display the real changes in emotional valence alongside the JS divergence and, in Figure 8, real changes and change detection through entropy. An increase in entropy within a three-day temporal window, as seen in this case, is associated with a shift in emotional state. Similarly, divergence remains low, close to 0, during stable periods, with most peaks corresponding to real shifts in emotion.

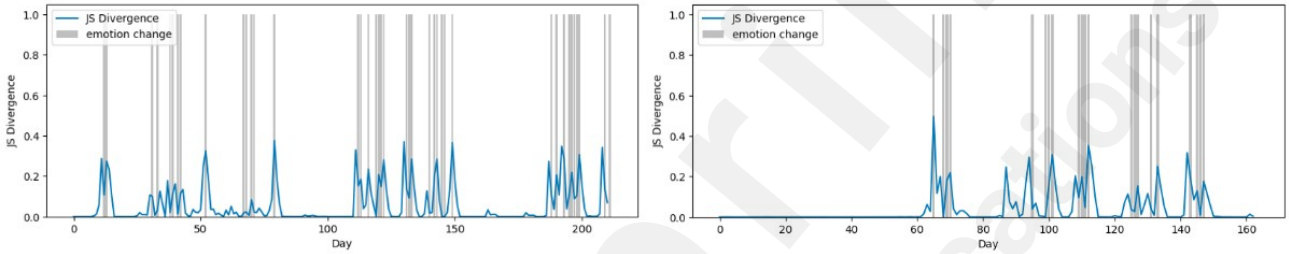


Figure 7. JS divergence calculated in a temporal window (blue curve) and the true emotion changes (grey vertical lines) for two patient sequences.

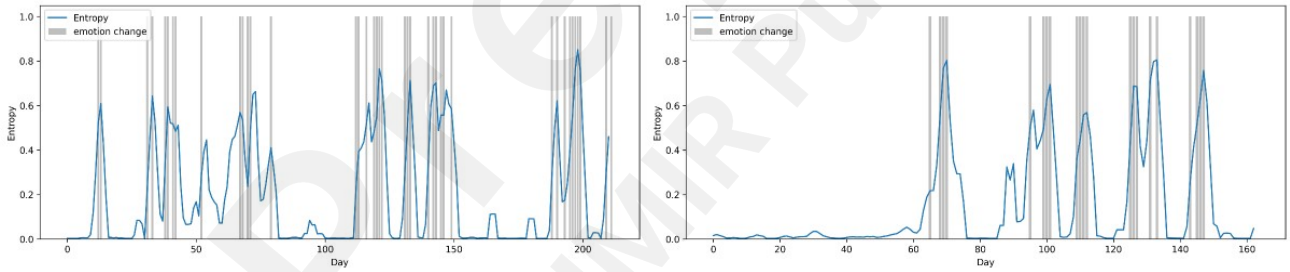


Figure 8. Entropy calculated in a temporal window (blue curve) and the true emotion changes (grey vertical lines) for two patient sequences.

Figure 9 shows the ROC AUC for emotional change detection with entropy and JS divergence. The results of change detection are shown for patients suffering from mental disorders, for patients who have not been diagnosed with mental disorders, and for the total of all study participants to compare detection across different scenarios. For the global case, lower thresholds for both metrics yielded better results, with JS divergence outperforming entropy by 0.13 ROC AUC for change detection.

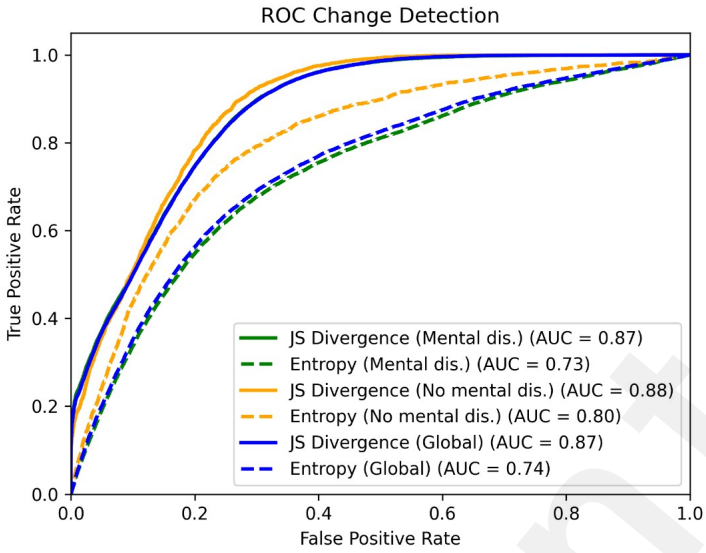


Figure 9. ROC curves for change detection with JS divergence and entropy. The curves are obtained for change detection in patients suffering from mental disorders (mental dis.), in patients without any diagnosed mental disorders (no mental dis.), and for the entire study cohort (global).

PHQ-9 Score Forecasting

Table 3 presents the accuracy and AUC results for the questionnaire score forecasting, delineated by the classification into two classes over a temporal window of 7 days using the decoder transformer outputs. The predictions were obtained using both a Random Forest classifier and an XGBoost Classifier, both of which demonstrated similar results. Among the results presented in Table 3, the difference lies in the input provided to the classifier: either solely the transformer output (hidden states), the transformer output combined with the corresponding emotional state forecasting, or solely the emotional state. Notably, the results across the three configurations exhibit similarity, slightly improved when exclusively utilizing the hidden states or solely the emotion distribution sequence as input.

Table 3. PHQ-9 score forecasting results for questions 1 to 9.[a]

	Transformer Embeddings[b]		**Transformer Embeddings + Emotion**[c]		**Emotion 3 days**[d]		**Emotion 15 days**[d]	
PHQ-9 question	**Accuracy**	**ROC AUC**	**Accuracy**	**ROC AUC**	**Accuracy**	**ROC AUC**	**Accuracy**	**ROC AUC**
1	0.57	0.62	0.57	0.60	*0.59*	*0.63*	0.57	0.61
2	0.6	0.60	0.57	0.59	*0.63*	*0.64*	0.63	0.63
3	0.55	0.51	0.55	0.50	0.55	0.55	*0.58*	*0.62*
4	*0.62*	*0.61*	0.53	0.49	0.53	0.54	0.57	0.57
5	0.54	0.57	0.55	0.60	*0.64*	*0.69*	0.53	0.60
6	0.46	0.46	0.47	0.51	0.6	0.51	0.61	0.65
7	0.52	0.54	0.55	0.54	0.61	0.62	*0.67*	*0.64*
8	0.64	0.52	0.61	0.53	0.68	0.50	*0.72*	*0.56*
9	*0.91*	*0.77*	0.91	0.74	0.91	0.60	0.91	0.74

[a] Results correspond to the classifier performance with different input sets:

[b] Classifier inputs: sequence of embeddings obtained from the transformer.

[c] Classifier inputs: sequence of embeddings obtained from the transformer along with

emotion valence prediction obtained from the emotion classifier.
[d] Classifier inputs: sequence of emotion valence predictions obtained from the emotion classifier.
[e] For each question, the results obtained with the best model are highlighted in italics.

When comparing different questions, certain inquiries demonstrate higher predictability than others. Notably, question 9 consistently yields the most accurate predictions across all configurations. This question pertains to whether patients have experienced thoughts of being better off dead or hurting themselves. In addition, question 8, which addresses difficulties in movement or speech that others may notice, exhibits relatively accurate predictions, along with question 5, which is related to appetite, and question 7, related to concentration problems.

Discussion

Principal Results

In terms of time-series forecasting, our findings indicate that the transformer model surpasses the hidden Markov model (HMM) in predicting future time steps, resulting in more stable predictions. This suggests that attention mechanisms within the transformer model are effective in capturing longer temporal dependencies, leading to improved prediction stability. Such capabilities are particularly beneficial for assessing a patient's state several days in advance, providing valuable insights into their potential behavior and enabling the early detection of high-risk situations.
The findings concerning emotional changes are generally positive, encompassing both global emotion detection results and change detection one day in advance. The variance in outcomes across different machine learning classification algorithms is minimal, highlighting the robustness of the variables and their latent representation. This indicates that accurately predicting a patient's emotional state can be achieved solely through passive variables, with the representation of these variables in the posterior probabilities of the hidden states (obtained with generative models) proving to be informative.
When examining individual patient sequences, it becomes apparent that in many instances, we can correctly anticipate mood shifts in advance. As for change detection using measures such as entropy or divergence, there appears to be a correlation between the disorder in emotional valence probability distribution within a specific temporal window and the subsequent emotional state change. Furthermore, global findings indicate that divergence yields superior results for change detection, suggesting that identifying relative disorder between the current prediction distribution and the past-window distribution holds greater significance than overall disorder within the window, although both contribute to change detection. Similar results are obtained when comparing change detection in patients without any mental disorders and those who suffer from them. This indicates that the detection is accurate in both cases, with slightly better results for the patients without mental disorders. This can be expected, as the literature suggests that these patients exhibit less fluctuation in their emotions compared to patients with mental disorders.
Regarding the PHQ-9 answers, our model aims to predict responses with considerable accuracy across both high and low-frequency classes, allowing for a comprehensive overview of PHQ-9 outcomes. Upon analyzing differences between answers, those better predicted are typically those wherein patients may find it easier to detect their emotions regarding the topic and the frequency of their thoughts. For instance, question 9, corresponding to "Thoughts that you would be better off dead, or of hurting yourself", exhibits the highest classification AUC and accuracy. Conversely, answers such as "Trouble falling or staying asleep, or sleeping too much"

and “Feeling bad about yourself or that you are a failure or have let yourself or your family down” yield poorer results.
Having an approximate score on this questionnaire proves to be a valuable tool, facilitating the monitoring of changes in depressive symptoms over time and guiding treatment decisions. Consequently, it can be employed for screening purposes to identify individuals who may require further evaluation for depression.

Limitations

There is a high percentage of missing data in the passive variables, with the minimum percentage of missing data being 52.6% (519,168/986,909) for the number of steps and the maximum percentage being 69.47% (685,614/986,909) for time at home. Additionally, there is a very high rate of missing data in the active variables we aim to predict, such as emotional valence, which has 96.34% missing data (950,833/986,909).
The recording of emotions is slightly imbalanced, with a higher rate of negative emotions recorded (negative: 45.83%, positive: 30.41%, neutral: 23.74%). Consequently, there is greater sensitivity for detecting negative valence compared to neutral and positive states.
For the PHQ-9, we have limited responses (549), and they are sporadic for most of the patients. The mean interval between two responses is 25.27 days, with a mode of 14.5 days. On average, each patient responds 4.65 times. Thus, predicting the global score based on passive data is challenging.

Conclusions

Our study has yielded several key findings. First and foremost, the utilization of passive variables has led to favorable outcomes in both emotion valence detection and change analysis. This underscores the potential of leveraging passive data sources for monitoring and understanding emotional states.
Moreover, employing temporal methods has enabled accurate prediction of emotional states up to a day in advance, with stable results for subsequent days. This temporal stability in prediction highlights the robustness of our approach and suggests its potential applicability in real-world settings where timely intervention is crucial. Further exploration may elucidate the potential for extending prediction horizons beyond a single day, thereby enhancing the utility of our method in long-term monitoring scenarios.
Furthermore, our model demonstrates promising performance in predicting PHQ-9 scores, providing an approximate understanding derived solely from passive data. This highlights the utility of this approach as a screening tool for identifying individuals at higher risk of having a crisis or whose depressive symptoms are changing, thereby enabling timely intervention or adjustment in treatment.
Future work should focus on the interpretability of the data. This includes exploring the impact of each variable to assess the patient’s condition, as well as the effects of removing or adding new variables (such as heart rate, oxygen saturation, etc.). Additionally, delving deeper into the understanding and extraction of information from the time series could help identify behavioral patterns that determine patient progression, as well as pinpoint specific moments that are most relevant for changes in a patient’s emotional state, ultimately aiding in adapting the treatment.
While there is room for improvement, particularly in refining the predictive accuracy and expanding the scope of our analysis, our findings represent a significant step forward in the development of in-situ support and unobtrusive monitoring strategies for mental health disorders.

Acknowledgements

The work of L. Paz was supported by the Spanish Instituto de Salud Carlos III (PMP22/00032). The work of A. Artés-Rodríguez was partially supported by Spanish Instituto de Salud Carlos III (PMP22/00032), Ministerio de Ciencia, Innovación y Universidades (CPP2022- 009537), and Ministerio de Ciencia e Innovación jointly with the European Commission (ERDF) (PID2021- 123182OB-I00, PID2021-125159NB-I00). The work of D. Ramírez was partially supported by the Office of Naval Research (ONR) Global under contract N62909-23-1-2002, by the Comunidad de Madrid under grant IND2023/TIC-27508 (IRIS) and by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D SORUS project. P. M. Olmos acknowledges the support by Comunidad de Madrid under grants IND2022/TIC-23550 and ELLIS Unit Madrid. Both D. Ramírez and P. M. Olmos are supported by MICIU/AEI/10.13039/501100011033/FEDER, UE, under grant PID2021-123182OB-I00 (EPiCENTER).

Conflicts of Interest

AA is founder of Evidence-Based Behavior (eB2). The rest of authors and coauthors have no conflicts to declare.

Abbreviations

AUC: area under the curve
CMD: Common Mental Disorder
EM: ensemble models
HHMMs: heterogeneous-HMMs
HMM: hidden Markov model
MLP: multilayer perceptron
PHQ-9: Patient Health Questionnaire
PR: precision-recall
RF: random forest classifier
ROC: receiver operating characteristic
SVC: support vector classifier
XGB: extreme gradient boosting

References

1. American Psychiatric Association, DSM-5 Task Force. Diagnostic and statistical manual of mental disorders: DSM-5™. 5th ed. American Psychiatric Publishing, Inc.; 2013. doi:10.1176/appi.books.9780890425596.
2. Panlilio LV, Stull SW, Kowalczyk WJ, Phillips KA, Schroeder JR, Bertz JW, et al. Stress, craving and mood as predictors of early dropout from opioid agonist therapy. Drug Alcohol Depend. 2019 Sep 1;202:200-208. doi: 10.1016/j.drugalcdep.2019.05.026. Epub 2019 Jul 16. PMID: 31357121; PMCID: PMC6707374.
3. Ortiz A, Grof P. Electronic monitoring of self-reported mood: the return of the subjective? Int J Bipolar Disord. 2016 Dec;4(1):28. doi: 10.1186/s40345-016-0069-x. Epub 2016 Nov 29. PMID: 27900735; PMCID: PMC5127918.
4. Anestis MD, Selby EA, Crosby RD, Wonderlich SA, Engel SG, Joiner TE (2010). A comparison of retrospective self-report versus ecological momentary assessment measures of affective lability in the examination of its relationship with bulimic symptomatology. *Behaviour research and therapy*, *48*(7), 607–613. doi:

10.1016/j.brat.2010.03.012
5. Gershon A, Kaufmann CN, Torous J, Depp C, Ketter TA. Electronic Ecological Momentary Assessment (EMA) in youth with bipolar disorder: demographic and clinical predictors of electronic EMA adherence. J Psychiatr Res. 2019 Sep;116:14-18. doi: 10.1016/j.jpsychires.2019.05.026.
6. Shiffman S, Stone AA, Hufford MR. Ecological Momentary Assessment. Annu Rev Clin Psychol. 2008;4:1-32. doi: 10.1146/annurev.clinpsy.3.022806.091415. PMID: 18509902.
7. LiKamWa R, Liu Y, Lane ND, Zhong L. MoodScope: building a mood sensor from smartphone usage patterns. In: Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '13). 2013. p. 389-402. doi: 10.1145/2462456.2464449.
8. Garriga R, Mas J, Abraha, S. *et al.* Machine learning model to predict mental health crises from electronic health records. Nat Med. 2022 Aug;28(8):1240-1248. doi: 10.1038/s41591-022-01811-5.
9. Zeng A, Chen M, Zhang L, Xu Q. Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence. 2023. p. 11121-11128. doi: 10.1609/aaai.v37i9.26317.
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. arXiv. 2017 Jun. arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
11. Catania F, Spitale M, Garzotto F. Conversational Agents in Therapeutic Interventions for Neurodevelopmental Disorders: A Survey. ACM Trans Access Comput. 2023 Feb;55:1-32. doi: 10.1145/3564269.
12. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106-11115. doi: 10.1609/aaai.v35i12.17325.
13. Ghandeharioun A, Fedor S, Sangermano L, Mohr DC. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. p. 325-332.
14. World Health Organization. Depression. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed 7 May 2024].
15. Angermann CE, Ertl G. Depression, Anxiety, and Cognitive Impairment. Curr Heart Fail Rep. 2018 Dec; 15:398-410. doi: 10.1007/s11897-018-0414-8.
16. De Angel, V., Lewis, S., White, K. et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. npj Digit. Med. 2022 Jan. doi: 10.1038/s41746-021-00548-8.
17. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sept;16(9):606-613. doi: 10.1046/j.1525-1497.2001.016009606.x.
18. LeDoux JE, Hofmann SG. The subjective experience of emotion: a fearful view. Curr Opin Behav Sci. 2017 Oct;19:67-72. doi: 10.1016/j.cobeha.2017.09.011.
19. Bowen R, Clark M, Baetz M. Mood swings in patients with anxiety disorders compared with normal controls. J Affect Disord. 2004;78:185-192. doi: 10.1016/S0165-0327(02)00304-X.
20. Bowen R, Baetz M, Hawkes J, Bowen A. Mood variability in anxiety disorders. J Affect Disord. 2006;91(2-3):165-170. doi:10.1016/j.jad.2005.12.050.
21. Saeb S, Zhang M, Karr CJ,Schueller SM, Corden ME,Kording KP, Mohr DC. Mobile Phone

Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res 2015;17(7):e175. doi: 10.2196/jmir.4273. PMID: 26180009. PMCID: 4526997.
22. Ram N, Brinberg M, Pincus AL, Conroy DE. The Questionable Ecological Validity of Ecological Momentary Assessment: Considerations for Design and Analysis. Res Hum Dev. 2017;14(3):253-270. doi:10.1080/15427609.2017.1340052
23. Coppersmith DD, Harman JL, Dredze M. Heterogeneity in suicide risk: evidence from personalized dynamic models. Behav Res Ther. 2024 Mar;180:104574. doi: 10.1016/j.brat.2024.104574.
24. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. *J Biomed Inform*. 2018;77:120-132. doi:10.1016/j.jbi.2017.12.008
25. Dillon CB, McMahon E, O'Regan G, Perry IJ. Associations between physical behaviour patterns and levels of depressive symptoms, anxiety and well-being in middle-aged adults: a cross-sectional study using isotemporal substitution models. BMJ Open. 2018 Mar 23;8(3). doi: 10.1136/bmjopen-2017-018978.
26. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annu Rev Clin Psychol*. 2017;13:23-47. doi:10.1146/annurev-clinpsy-032816-044949.
27. Morshed MB, Saha K, Li R, D'Mello SK, De Choudhury M, Abowd GD, Plötz T. Prediction of mood instability with passive sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2019;3(3):75. doi:10.1145/3351233.
28. Kolakowska A, Szwoch W, Szwoch M. A Review of Emotion Recognition Methods Based on Data Acquired via Smartphone Sensors. Sensors (Basel). 2020 Nov 2;20(21):6367. doi: 10.3390/s20216367.
29. Wahle F, Kowatsch T, Fleisch E,Rufer M, Weidt S. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. JMIR Mhealth Uhealth. 2016;4(3):e111. doi: 10.2196/mhealth.5960. PMID: 27655245. PMCID: 5052463.
30. Kim S, Lee K. Screening for Depression in Mobile Devices Using Patient Health Questionnaire-9 (PHQ-9) Data: A Diagnostic Meta-Analysis via Machine Learning Methods. Neuropsychiatr Dis Treat. 2021;17:3415-3430. Published 2021 Nov 20. doi:10.2147/NDT.S339412.
31. Stamatis CA, Meyerhoff J, Meng Y, et al. Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. Npj Ment Health Res. 2024;3(1):1. Published 2024 Jan 4. doi:10.1038/s44184-023-00041-y.
32. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: A prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord. 2023;324:170-174. doi:10.1016/j.jad.2022.12.127.
33. Lakhtakia T, Smith SR, Mohr DC, Stamatis CA. Longitudinal associations of daily affective dynamics with depression, generalized anxiety, and social anxiety symptoms. J Affect Disord. 2024;352:437-444. doi:10.1016/j.jad.2024.01.250.
34. Sükei E, Norbury A, Perez-Rodriguez MM, Olmos PM, Artés A. Predicting Emotional States Using Behavioral Markers Derived From Passively Sensed Data: Data-Driven Machine Learning Approach. JMIR Mhealth Uhealth 2021;9(3):e24465. doi: 10.2196/24465. PMID: 33749612. PMCID: 8088855
35. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A Transformer-based Framework for Multivariate Time Series Representation Learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2021

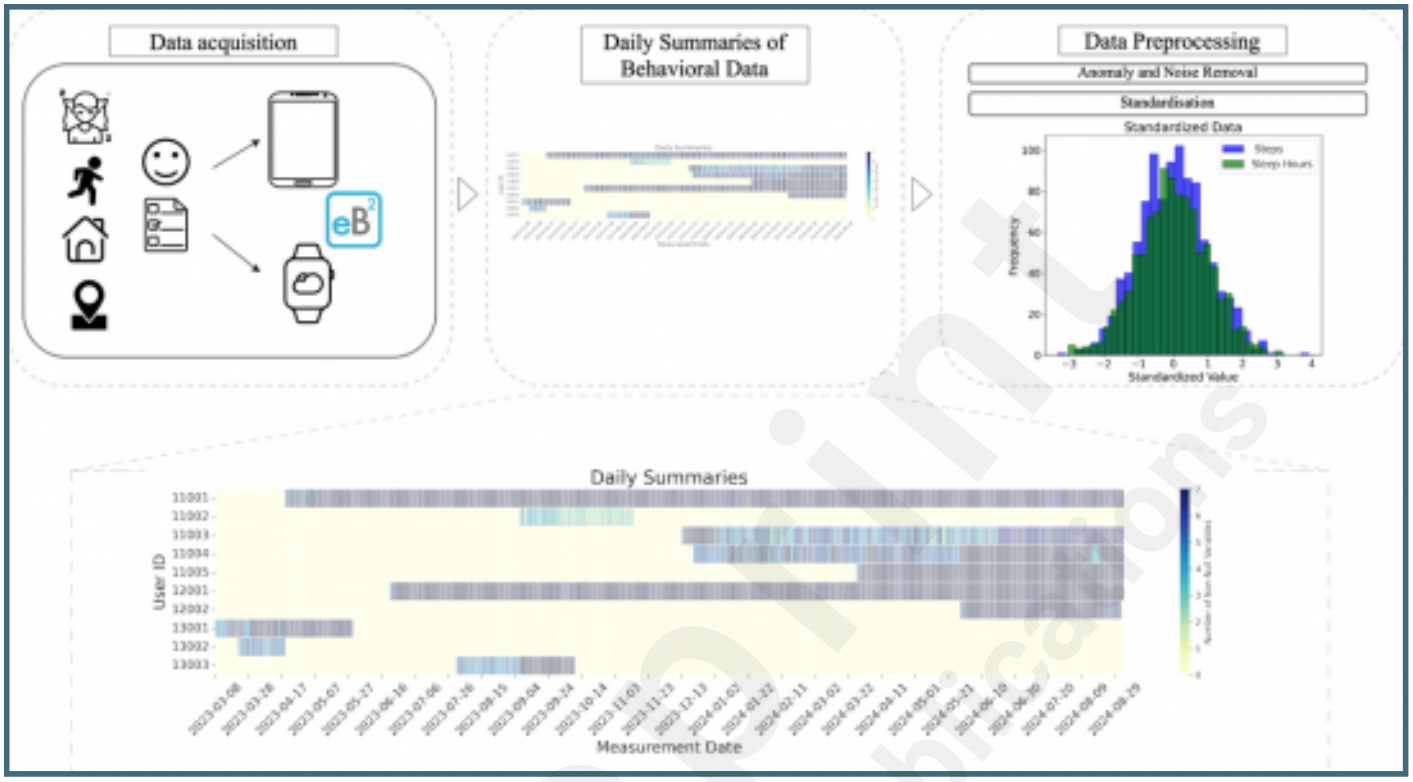
Aug; p. 1194-1204. doi: 10.1145/3447548.3467401.
36. Tong J, Xie L, Yang W, Zhang K, Zhao J. Enhancing time series forecasting: A hierarchical transformer with probabilistic decomposition representation. Inf Sci. 2023 Apr;647:119410. doi: 10.1016/j.ins.2023.119410.
37. Chefer H, Gur S, Wolf L. Transformer Interpretability Beyond Attention Visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021 Jun; p. 782-791.
38. Risal A. Common mental disorders. Kathmandu Univ Med J (KUMJ). 2011;9(35):213-217. doi:10.3126/kumj.v9i3.6308.
39. Berrouiguet S, Ramírez D, Barrigón ML, 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. 2018;6(12):e197. Published 2018 Dec 10. doi:10.2196/mhealth.9472.
40. Carretero P, Campana-Montes JJ, Artes-Rodriguez A. Ecological Momentary Assessment for Monitoring Risk of Suicide Behavior. Curr Top Behav Neurosci. 2020;46:229-245. doi:10.1007/7854_2020_170.
41. eb2.tech. Available from: https://eb2.tech/?lang=en. Accessed 2024 Feb 22.
42. Bonilla-Escribano P, Ramirez D, Sedano-Capdevila A, et al. Assessment of e-Social Activity in Psychiatric Patients. IEEE J Biomed Health Inform. 2019;23(6):2247-2256. doi:10.1109/JBHI.2019.2918687.
43. Russell JA. Core affect and the psychological construction of emotion. Psychol Rev. 2003 Jan;110(1):145-172. doi: 10.1037/0033-295X.110.1.145.
44. Bishop CM, Nasrabadi NM. Pattern Recognition and Machine Learning. Springer-Verlag, New York; 2006.
45. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989 Feb;77(2):257-286. doi: 10.1109/5.18626.
46. Moreno-Pino F, Sükei E, Olmos PM, Aukerṭes-Rodr̈iguez A. PyHHMM: A Python library for heterogeneous hidden Markov models. arXiv preprint. 2022 Jan. doi: 10.48550/arXiv.2201.06968.
47. Huang T, Yang F, Zhao J. Multi-Head Attention Mechanisms for Fusion of Visual and Audio Features in Emotion Recognition. Multimodal Interaction Research. 2020;35(7):1205-1222.
48. Luna-Jiménez A, Huang T, Yang F. Fine-Tuning xlsr-wav2vec2.0 for Speech Emotion Analysis. Speech Emotion Processing Journal. 2021;19(4):87-102.
49. Xie J, Zaidi A, Sun R. Dialogue Emotion Recognition with GPT and RoBERTa Embeddings: Cross-Modal Fusion for Enhanced Accuracy. Proceedings of the 2021 International Conference on Affective Computing; 2021.
50. Götz D. MATS2L: Transformer-Based Emotion Recognition Using EEG and ECG Signals. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2023;31:45-58.
51. Zaidi A, Sun R, Luna-Jiménez A. Cross-Modal Emotion Recognition Using RoBERTa and wav2vec 2.0. Affective Multimodal Learning. 2023;14(2):23-45.
52. Sun R, Zhao L, Götz D. Multimodal Integration of wav2vec 2.0 and BERT for Speech and Text Emotion Analysis. Proceedings of the Emotion Recognition Symposium; 2023.
53. Zhao L, Xie J, Zaidi A. MEmoBERT: A Transformer-Based Framework for Cross-Modal Emotion Classification. Neurocomputing. 2022;480:33-48. doi:10.1016/j.neucom.2022.08.134.
54. Yang F, Zhao L. Conformer-Based Models for Physiological Emotion Classification. Journal of Affective Computing. 2022;9(6):500-520.
55. Dridi N, Hadzagic M. Akaike and Bayesian Information Criteria for Hidden Markov Models.

IEEE Signal Process Lett. 2018 Dec 14;26(2):302-306. doi:10.1109/LSP.2018.2886015.
56. Russell JA. A Circumplex Model of Affect. J Pers Soc Psychol. 1980;39(6):1161-1178. doi:10.1037/h0077714.
57. Posner J, Russell JA, Peterson BS. The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. Dev Psychopathol. 2005;17(3):715-734.
58. Mitsios M, Vamvoukakis G, Maniati G, Ellinas N, Dimitriou G, Markopoulos K, Kakoulidis P, Vioni A, Christidou M, Oh J, Jho G. Improved Text Emotion Prediction Using Combined Valence and Arousal Ordinal Classification. arXiv preprint arXiv:2404.01805. 2024 Apr 2

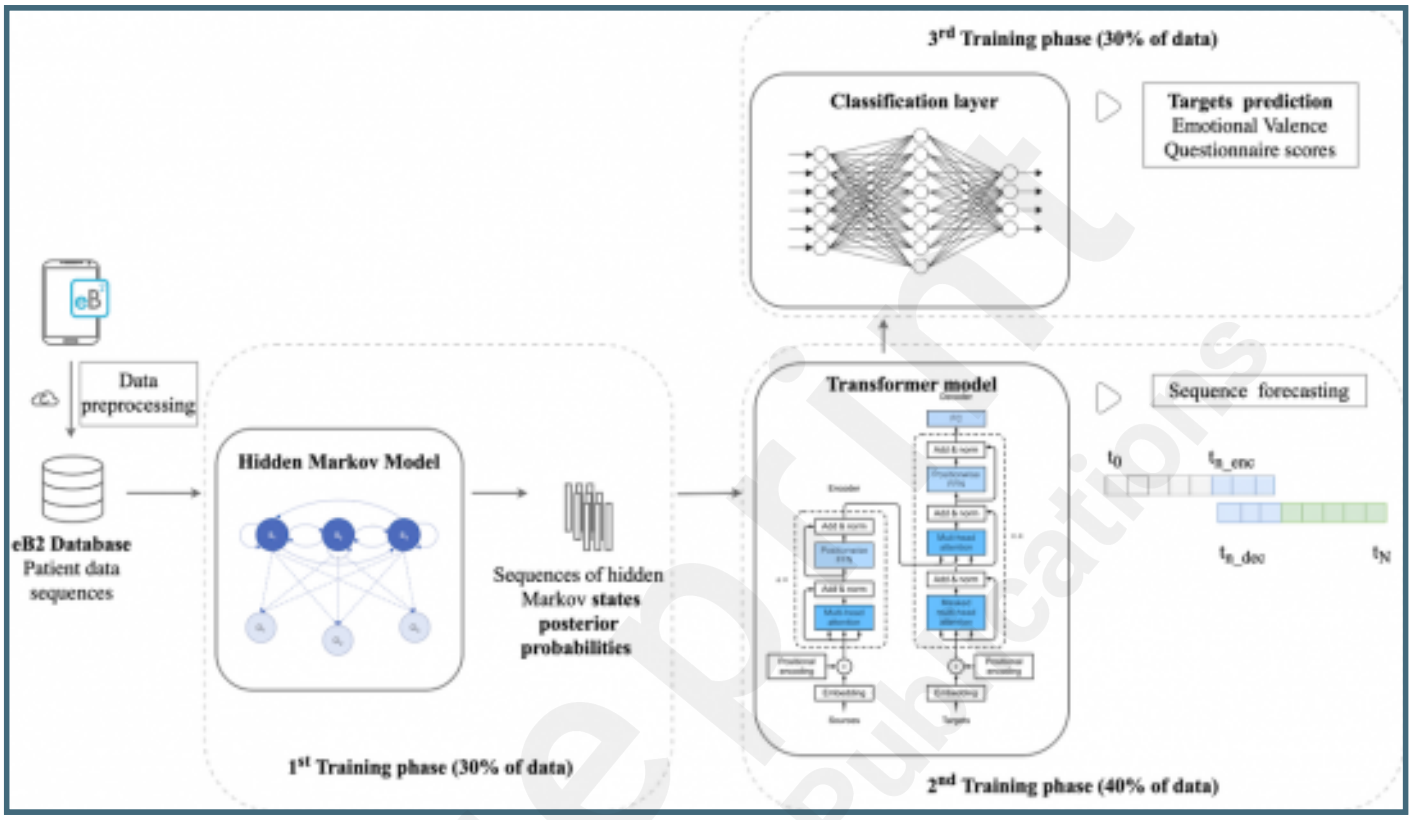
Supplementary Files

Figures

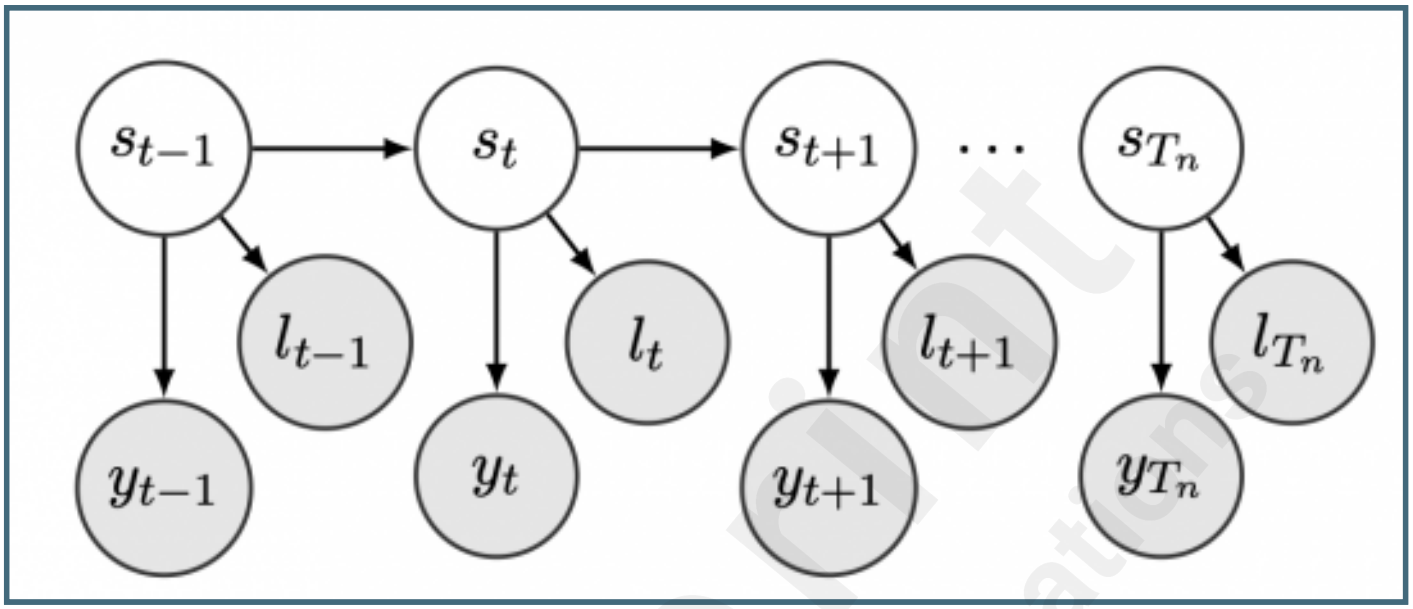
Data preprocessing pipeline: data acquisition, obtention of the daily summaries, and standardization of behavioral data. The sequence shown in the daily summaries displays the temporal sequence of passive data for five different patients. The intensity of the lines indicates the amount of non-missing behavioral data the patient has for that day.



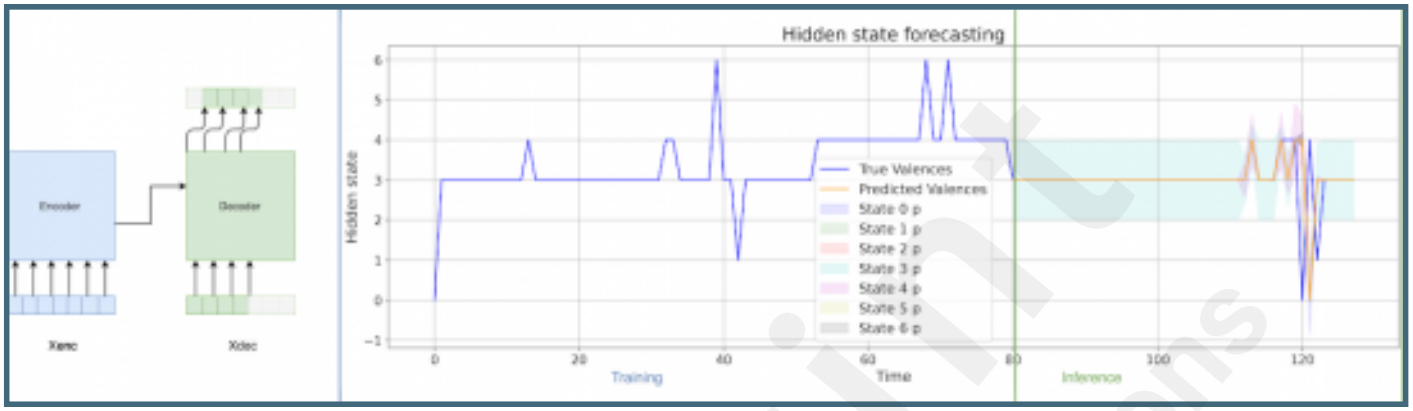
The graphical abstract of the proposed scheme illustrates the underlying architecture, comprising three primary blocks. The first block involves the utilization of Hidden Markov Models (HMM) to address missing data and extract posterior state probabilities. The second block employs a transformer model equipped with an attention mechanism to facilitate pattern recognition and time-series forecasting. Lastly, the scheme incorporates a final classification layer responsible for patient-reported variables, namely Emotional Valence or PHQ-9 score.



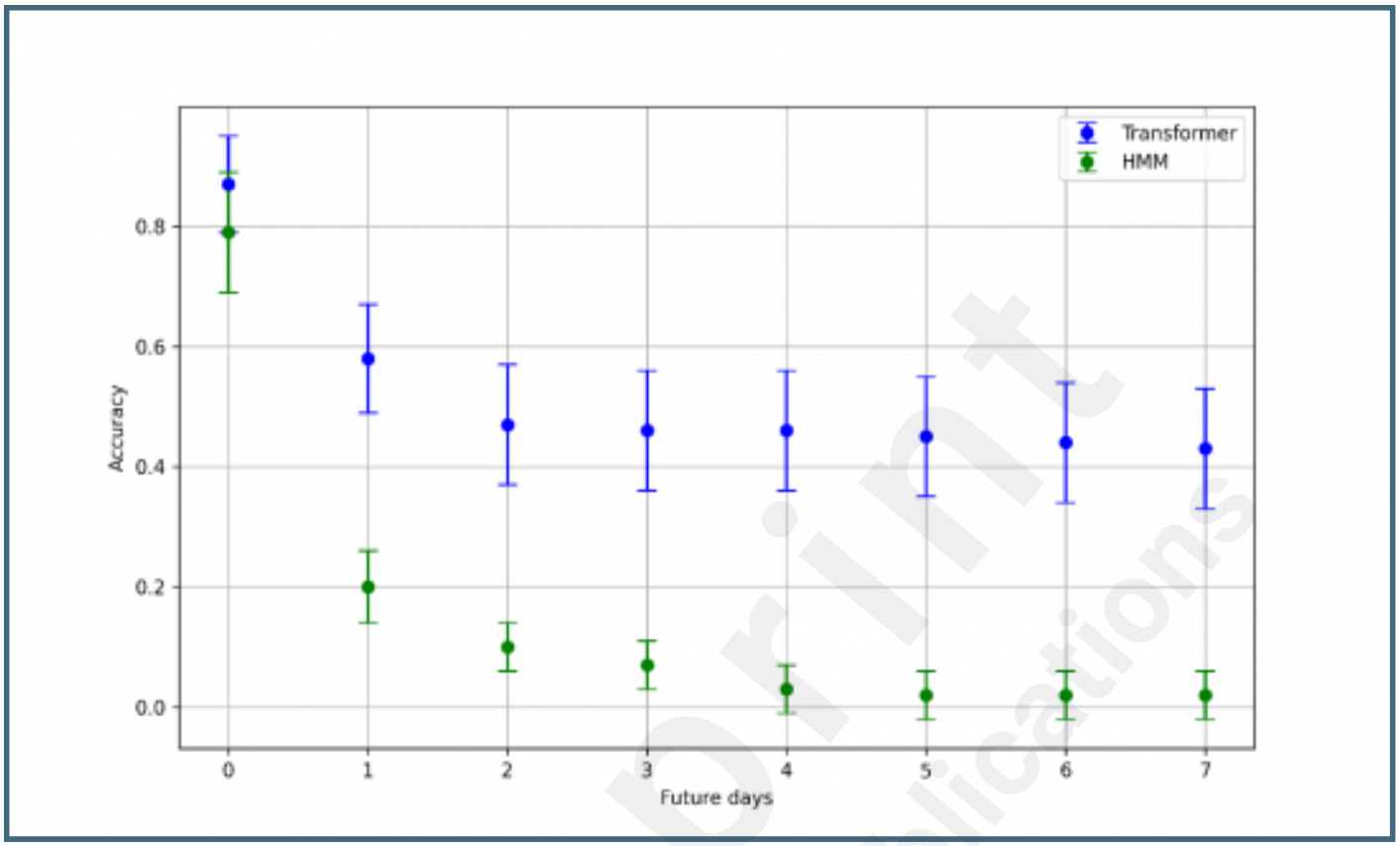
Architecture of Heterogeneous hidden Markov model. Model is described by their hidden states sequence (s0:Tn) and continuous observations sequence (y0:Tn), discrete observations sequence (l0:Tn) [44]. In our case (l0:Tn) corresponded to the sequence of discrete observation: practiced sport and (y0:Tn) was the sequence of continuous observations: steps, location distances, sleep time, app usage, time home and location clusters count.



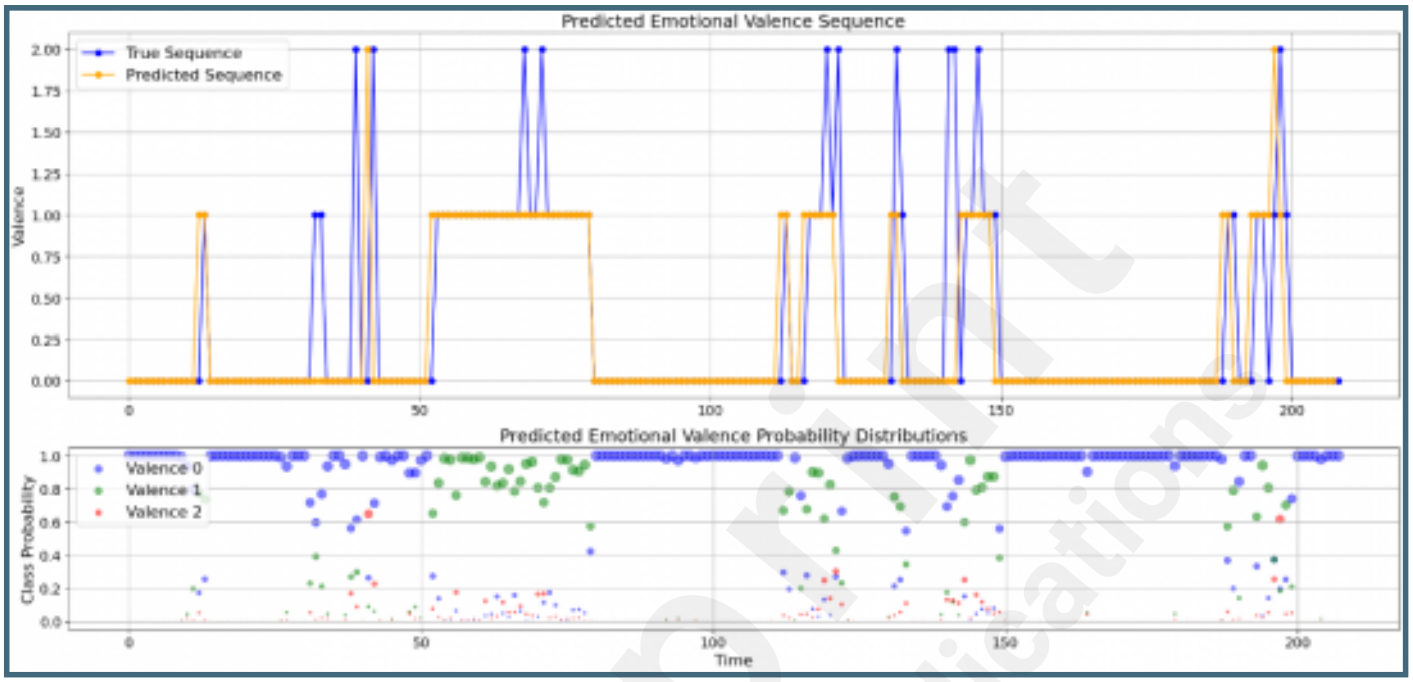
On the left side of the image, the simplified encoder-decoder architecture of the transformer model is shown [10]. On the right side, an example of model training and inference is illustrated. During training, the model learns the parameters to forecast the future state, using the real data as a reference. During inference, the transformer's decoder is responsible for forecasting the future state value in an autoregressive manner. The blue line shows the true states, and the orange line shows those predicted by the model. During inference, the possible future states have associated probabilities, which are illustrated in the graph with different colored margins around the orange line.



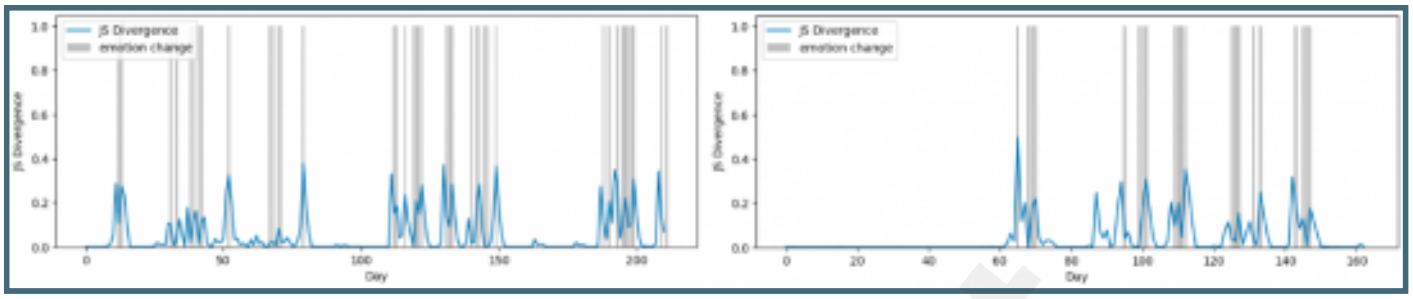
Model comparison in state matches forecasting.



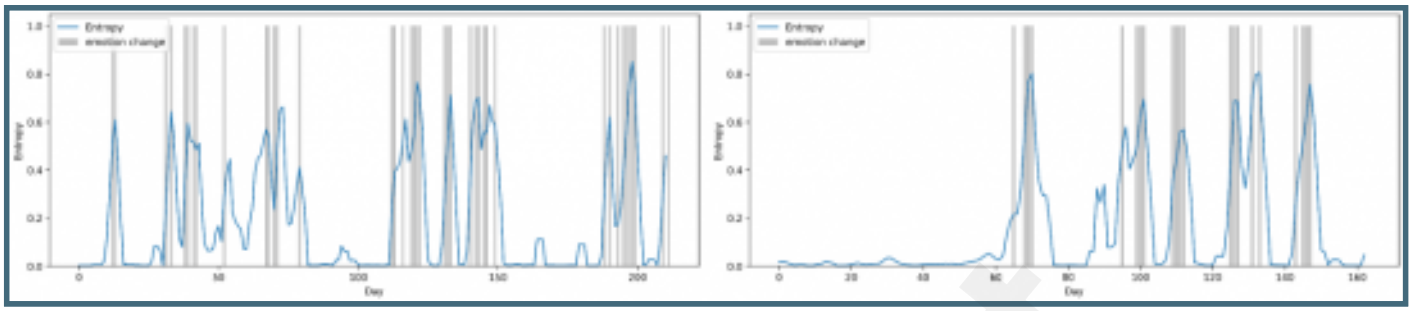
Sequences of real emotions (blue) and those predicted by the model (orange) for the following day. The colored circles denote the probability distribution of the emotional state, purple indicates state 0 (negative valence), green indicates state 1 (neutral), and red indicates state 2 (positive valence). These figures pertain to two patients (ids: 374 and 218) selected from a set of patients with high variability of emotion in their temporal sequence.



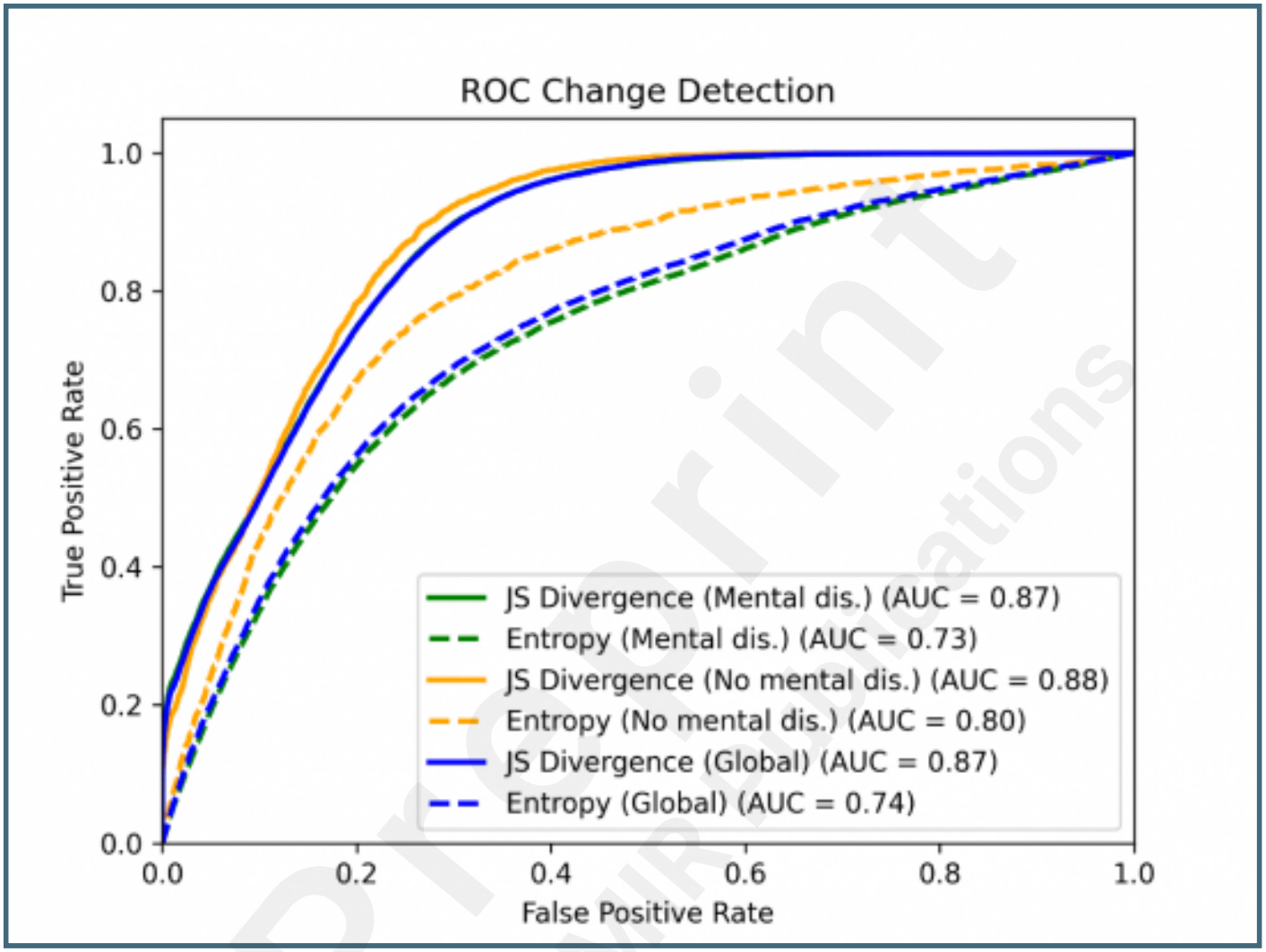
JS divergence calculated in a temporal window (blue curve) and the true emotion changes (grey vertical lines) for two patient sequences.



Entropy calculated in a temporal window (blue curve) and the true emotion changes (grey vertical lines) for two patient sequences.



ROC curves for change detection with JS divergence and entropy. The curves are obtained for change detection in patients suffering from mental disorders (mental dis.), in patients without any diagnosed mental disorders (no mental dis.), and for the entire study cohort (global). PHQ-9 Score Forecasting.



Multimedia Appendixes

Data Appendix: Cohort Overview and Patient Profiles.
URL: http://asset.jmir.pub/assets/df517827c64ddbea5e02cd08625d7780.docx

Technical Appendix: Model Description and Implementation Details.
URL: http://asset.jmir.pub/assets/9af5dc96d4aa679e9f58cab739708b64.docx