A STEP TOWARDS ADVANCING DIGITAL PHENOTYPING IN MENTAL HEALTHCARE

by

Emese Sükei

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in

Multimedia and Communications

Universidad Carlos III de Madrid

Advisors:

Antonio Artés Rodríguez

Pablo Martínez Olmos

2022

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

Marie Curie

ACKNOWLEDGEMENTS

Firstly, I would like to warmly thank my advisors, Prof.Dr Antonio Artés-Rodríguez and Dr Pablo Martínez Olmos, who made this work possible. Their patient guidance, insightful feedback, continuous generosity and unwavering support carried me through all the stages of my PhD studies.

Next, my sincere thanks go to Prof. Dr Kristel Van Steen, Dr Tobias Heimann and Dr Matthias Siebert for their supervision, support and tutelage during the secondments. I also appreciate their team's help and friendliness, making my stays not just fruitful but also memorable.

I want to thank the PhD students and post-docs of the Signal Theory group for introducing me to Spanish culture and making me a little less *guiri*, the stimulating discussions about work, and the extracurricular experiences we had. Special thanks to Alex and Lorena for helping me with the Spanish translation of the abstract.

It is impossible to extend enough thanks to my husband and family, who always believe in me, even when I don't, encourage me constantly and put up with me no matter what.

To my friends, this would have been a much more difficult feat without you. Thank you all for reminding me to take breaks and have fun sometimes, even when I've felt overwhelmed by deadlines.

PUBLISHED AND SUBMITTED CONTENT

The following list of works is a short bibliography of articles from journals, conferences, workshops and preprints, included as part of this thesis. Each contribution's complete or partial inclusion in this thesis is indicated. In addition, their presence in the chapters is shown in each of the introductory paragraphs. Lastly, I declare that this source's material included in the thesis is not marked by typographical means or references.

JOURNALS

1. Sükei E, Norbury A, Perez-Rodriguez M, Olmos P, 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 - Partly included in the introduction and Ch. 3

2. Ryu J, Sükei E, Norbury A, H Liu S, Campaña-Montes J, Baca-Garcia E, Artés A, Perez-Rodriguez M. Shift in Social Media App Usage During COVID-19 Lockdown and Clinical Anxiety Symptoms: Machine Learning–Based Ecological Momentary Assessment Study. *JMIR Ment Health* 2021;8(9):e30833. DOI: 10.2196/30833 - Partly included in Ch. 4

3. Moreno-Pino F, Sükei E, Olmos P, Artés A. PyHHMM: A Python Library for Heterogeneous Hidden Markov Models. *arXiv preprint*, 2022. arXiv:2201.06968 - Partly included in the introduction and Ch. 2

4. Sükei E, Romero-Medrano L, de Leon-Martinez S, Herrera López J, Campaña-Montes JJ, M. Olmos P, Baca-Garcia E, Artés A. Assessing WHODAS 2.0 Scores from Behavioral Biomarkers: a Data-driven Approach. *JMIR Preprints*. 25/03/2022: 38231 - Partly included in Ch. 5

CONFERENCES & WORKSHOPS

5. Sükei E, Norbury A, Perez-Rodriguez M, Olmos P, Artés A. Predicting Emotional State Using Behavioral Markers Derived from Passively Sensed Data. *ICML 2021 Workshop on Computational Approaches to Mental Health (CA2MH)*, 2021. [pdf]

ABSTRACT

Smartphones and wrist-wearable devices have infiltrated our lives in recent years. According to published statistics, nearly 84% of the world's population owns a smartphone, and almost 10% own a wearable device today (2022). These devices continuously generate various data sources from multiple sensors and apps, creating our digital phenotypes. This opens new research opportunities, particularly in mental health care, which has previously relied almost exclusively on self-reports of mental health symptoms.

Unobtrusive monitoring using patients' devices may result in clinically valuable markers that can improve diagnostic processes, tailor treatment choices, provide continuous insights into their condition for actionable outcomes, such as early signs of relapse, and develop new intervention models. However, these data sources must be translated into meaningful, actionable features related to mental health to achieve their full potential.

In the mental health field, there is a great need and much to be gained from defining a way to continuously assess the evolution of patients' mental states, ideally in their everyday environment, to support the monitoring and treatments by health care providers. A smartphone-based approach may be valuable in gathering long-term objective data, aside from the usually used self-ratings, to predict clinical state changes and investigate causal inferences about state changes in patients (e.g., those with affective disorders).

Being objective does not imply that passive data collection is also perfect. It has several challenges: some sensors generate vast volumes of data, and others cause significant battery drain. Furthermore, the analysis of raw passive data is complicated, and collecting certain types of data may interfere with the phenotype of interest. Nonetheless, machine learning is predisposed to address these matters and advance psychiatry's era of personalised medicine.

This work aimed to advance the research efforts on mobile and wearable sensors for mental health monitoring. We applied supervised and unsupervised machine learning methods to model and understand mental disease evolution based on the digital phenotype of patients and clinician assessments at the follow-up visits, which provide ground truths. We needed to cope with regularly and irregularly sampled, high-dimensional, and heterogeneous time series data susceptible to distortion and missingness. Hence, the developed methods must be robust to these limitations and handle missing data properly.

Throughout the various projects presented here, we used probabilistic latent variable models for data imputation and feature extraction, namely, mixture models (MM) and hidden Markov models (HMM). These unsupervised models can learn even in the presence of missing data by marginalising the missing values in the function of the present observa-

tions. Once the generative models are trained on the data set with missing values, they can be used to generate samples for imputation. First, the most probable component/state has to be found for each sample. Then, sampling from the most probable distribution yields valid and robust parameter estimates and explicit imputed values for variables that can be analysed as outcomes or predictors. The imputation process can be repeated several times, creating multiple datasets, thereby accounting for the uncertainty in the imputed values and implicitly augmenting the data. Moreover, they are robust to moderate deviations of the observed data from the assumed underlying distribution and provide accurate estimates even when missingness is high.

Depending on the properties of the data at hand, we employed feature extraction methods combined with classical machine learning algorithms or deep learning-based techniques for temporal modelling to predict various mental health outcomes - emotional state, World Health Organisation Disability Assessment Schedule (WHODAS 2.0) functionality scores and Generalised Anxiety Disorder-7 (GAD-7) scores, of psychiatric outpatients. We mainly focused on one-size-fits-all models, as the labelled sample size per patient was limited; however, in the mood prediction case, it was possible to apply personalised models.

Integrating machines and algorithms into the clinical workflow require interpretability to increase acceptance. Therefore, we also analysed feature importance by computing Shapley additive explanations (SHAP) values. SHAP values provide an overview of essential features in the machine learning models by designating the weight of predictability of each feature positively or negatively to the target variable.

The provided solutions, as such, are proof of concept, which require further clinical validation to be deployable in the clinical workflow. Still, the results are promising and lay some foundations for future research and collaboration among clinicians, patients, and computer scientists. They set the paths to advance future research prospects in technology-based mental healthcare.

RESUMEN

En los últimos años, los *smartphones* y los dispositivos y pulseras inteligentes, comúnmente conocidos como *wearables*, se han infiltrado en nuestras vidas. Según las estadísticas publicadas a día de hoy (2022), cerca del 84% de la población tiene un *smartphone* y aproximadamente un 10% también posee un *wearable*. Estos dispositivos generan datos de forma continua en base a distintos sensores y aplicaciones, creando así nuestro fenotipo digital. Estos datos abren nuevas vías de investigación, particularmente en el área de salud mental, dónde las fuentes de datos han sido casi exclusivamente autoevaluaciones de síntomas de salud mental.

Monitorizar de forma no intrusiva a los pacientes mediante sus dispositivos puede dar lugar a marcadores valiosos en aplicación clínica. Esto permite mejorar los procesos de diagnóstico, adaptar tratamientos, e incluso proporcionar información continua sobre el estado de los pacientes, como signos tempranos de recaída, y hasta desarrollar nuevos modelos de intervención. Aun así, estos datos en crudo han de ser traducidos a datos interpretables relacionados con la salud mental para conseguir un máximo rendimiento de los mismos.

En salud mental existe una gran necesidad, y además hay mucho que ganar, de definir cómo evaluar de forma continuada la evolución del estado mental de los pacientes en su entorno cotidiano para ayudar en el tratamiento y seguimiento de los mismos por parte de los profesionales sanitarios. En este ámbito, un enfoque basado en datos recopilados desde sus *smartphones* puede ser valioso para recoger datos objetivos a largo plazo al mismo tiempo que se acompaña de las autoevaluaciones utilizadas habitualmente. La combinación de ambos tipos de datos puede ayudar a predecir los cambios en el estado clínico de estos pacientes e investigar las relaciones causales sobre estos cambios (por ejemplo, en aquellos que padecen trastornos afectivos).

Aunque la recogida de datos de forma pasiva tiene la ventaja de ser objetiva, también implica varios retos. Por un lado, ciertos sensores generan grandes volúmenes de datos, provocando un importante consumo de batería. Además, el análisis de los datos pasivos en crudo es complicado, y la recogida de ciertos tipos de datos puede interferir con el fenotipo que se quiera analizar. No obstante, el *machine learning* o aprendizaje automático, está predispuesto a resolver estas cuestiones y aportar avances en la medicina personalizada aplicada a psiquiatría.

Esta tesis tiene como objetivo avanzar en la investigación de los datos recogidos por sensores de *smartphones* y *wearables* para la monitorización en salud mental. Para ello, aplicamos métodos de aprendizaje automático supervisado y no supervisado para modelar

y comprender la evolución de las enfermedades mentales basándonos en el fenotipo digital de los pacientes. Estos resultados se comparan con las evaluaciones de los médicos en las visitas de seguimiento, que proporcionan las etiquetas reales. Para aplicar estos métodos hemos lidiado con datos provenientes de series temporales con alta dimensionalidad, muestreados de forma regular e irregular, heterogéneos y, además, susceptibles a presentar patrones de datos perdidos y/o distorsionados. Por lo tanto, los métodos desarrollados deben ser resistentes a estas limitaciones y manejar adecuadamente los datos perdidos.

A lo largo de los distintos proyectos presentados en este trabajo, hemos utilizado modelos probabilísticos de variables latentes para la imputación de datos y la extracción de características, como por ejemplo, Mixture Models (MM) y hidden Markov Models (HMM). Estos modelos no supervisados pueden aprender incluso en presencia de datos perdidos, marginalizando estos valores en función de las datos que sí han sido observados. Una vez entrenados los modelos generativos en el conjunto de datos con valores perdidos, pueden utilizarse para imputar dichos valores generando muestras. En primer lugar, hay que encontrar el componente/estado más probable para cada muestra. Luego, se muestrea de la distirbución más probable resultando en estimaciones de parámetros robustos y válidos. Además, genera imputaciones explícitas que pueden ser tratadas como resultados. Este proceso de imputación puede repetirse varias veces, creando múltiples conjuntos de datos, con lo que se tiene en cuenta la incertidumbre de los valores imputados y aumentándose así, implícitamente, los datos. Además, estas imputaciones son resistentes a desviaciones que puedan existir en los datos observados con respecto a la distribución subyacente asumida y proporcionan estimaciones precisas incluso cuando la falta de datos es elevada.

Dependiendo de las propiedades de los datos en cuestión, hemos usado métodos de extracción de características combinados con algoritmos clásicos de aprendizaje automático o técnicas basadas en *deep learning* o aprendizaje profundo para el modelado temporal. La finalidad de ambas opciones es ser capaces de predecir varios resultados de salud mental/estado emocional, como la puntuación sobre el *World Health Organisation Disability Assessment Schedule (WHODAS 2.0)*, o las puntuaciones del *generalised anxiety disorder-7 (GAD-7)* de pacientes psiquiátricos ambulatorios. Nos centramos principalmente en modelos generalizados, es decir, no personalizados para cada paciente sino explicativos para la mayoría, ya que el tamaño de muestras etiquetada por paciente es limitado; sin embargo, en el caso de la predicción del estado de ánimo, puidmos aplicar modelos personalizados.

Para que la integración de las máquinas y algoritmos dentro del flujo de trabajo clínico sea aceptada, se requiere que los resultados sean interpretables. Por lo tanto, en este trabajo también analizamos la importancia de las características sacadas por cada algoritmo en base a los valores de las explicaciones aditivas de Shapley (SHAP). Estos valores proporcionan una visión general de las características esenciales en los modelos de aprendizaje automático designando el peso, positivo o negativo, de cada característica en su predictibilidad sobre la variable objetivo.

Las soluciones aportadas en esta tesis, como tales, son pruebas de concepto, que requieren una mayor validación clínica para poder ser desplegadas en el flujo de trabajo clínico. Aun así, los resultados son prometedores y sientan base para futuras investigaciones y colaboraciones entre clínicos, pacientes y científicos de datos. Éstas establecen las guías para avanzar en las perspectivas de investigación futuras en la atención sanitaria mental basada en la tecnología.

CONTENTS

1 Introduction **1**

1.1 Digital Phenotyping 1

1.1.1 A Mental Health Perspective 2

1.1.2 The Complexity and Nuances of Passive Data 3

1.2 The Passive Monitoring Dataset 4

1.2.1 Study Participants 4

1.2.2 Data Collection 5

1.2.3 Data Description 6

1.3 Overview of Models and Contributions 8

1.3.1 Modelling Passively-Sensed Data 8

1.3.2 Predicting Personalised Mental Health Outcomes 9

1.4 Document Structure 12

2 Modelling Passively-Sensed Data **13**

2.1 The Missing Data Problem 13

2.1.1 Types of Missing Data 13

2.1.2 Common Techniques for Handling the Missing Data 14

2.1.3 Probabilistic Generative Models for Dealing With Missing Data . 16

2.2 Temporal Data Modelling 19

2.2.1 Manual Feature Extraction-Based Approaches 19

2.2.2 Deep Neural Models 20

2.3 Discussion 25

3 Predicting Emotional States **27**

3.1 Introduction 27

3.1.1 Passively Sensed Behavioural Biomarkers 27

3.1.2 Related Work 28

3.1.3 Objectives 30

3.2 Methods 30

3.2.1 Data 30
3.2.2 Probabilistic Generative Models for Dealing With Missing Data . 32
3.2.3 Emotion Prediction Models 33
3.2.4 Personalised Models 34
3.2.5 Evaluation and Interpretability 35
3.2.6 Experiments 35
3.3 Results 36
3.3.1 Generative Models 36
3.3.2 Predicting Emotional Valence 39
3.3.3 Predicting Emotional Arousal-Valence 40
3.3.4 Personalised Models 41
3.3.5 Feature Importance Analysis 42
3.4 Discussion 43
3.4.1 Principal Findings 43
3.4.2 Limitations 44
3.4.3 Conclusions 45

4 Predicting Clinical Anxiety from Social Media App Usage During COVID-19 Lockdown 47

4.1 Introduction 47
4.2 Methods 48
4.2.1 Data 48
4.2.2 Machine Learning Pipeline 49
4.2.3 Performance Evaluation 51
4.3 Results 51
4.3.1 Predicting Clinical Anxiety 51
4.3.2 Feature Importance Analysis 53
4.4 Discussion 54

5 Assessing WHODAS 2.0 Scores from Behavioural Biomarkers 57

5.1 Introduction 57
5.2 A Baseline Approach 58
5.2.1 Data Selection and Preprocessing 59
5.2.2 Feature Selection and Predictive Modelling 60
5.2.3 Results 61

5.2.4 Discussion 63
5.3 A Deep Learning Approach 66
5.3.1 Data Selection and Preprocessing 66
5.3.2 The Proposed Pipeline 68
5.3.3 Experiments 69
5.3.4 Results 72
5.3.5 Discussion 75

6 Discussion and Conclusions 77
6.1 From Digital Phenotyping to Healthcare Solutions 77
6.2 Challenges 79
6.3 Directions for Future Work 81
6.4 Ethical Considerations 81
6.5 Conclusions 82

Bibliography 83

CHAPTER 1

INTRODUCTION

1.1 Digital Phenotyping

Three concepts permeate all aspects of biological life: genotype, phenotype, and environment. The phenotype of an organism refers to the collection of its observable traits. These traits are a product of the organism's genetics and environment and the interplay between the two (GxE interactions). The phenotype, in turn, produces the extended phenotype, defined by Dawkins [42] in 1982, which is the organism's impact on its surroundings to increase the likelihood of survival. Humans' extended phenotype can be found both off- and online (see Figure 1.1).

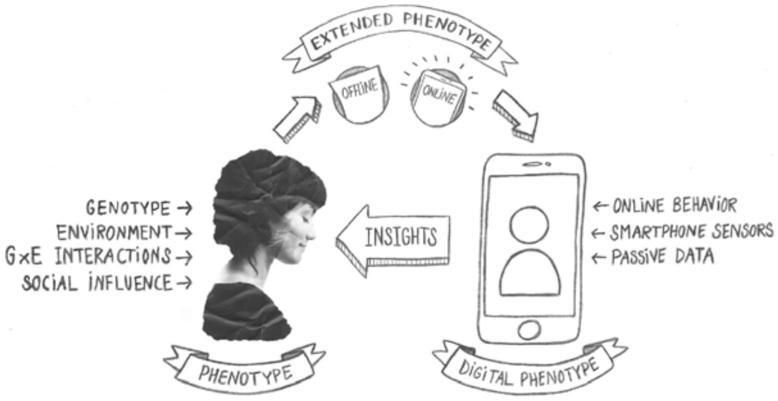


Figure 1.1: Schematic definition of the digital phenotype. Source [1].

Smartphones and wrist-wearables are now ubiquitous and can be harnessed to offer a moment-by-moment quantification of our digital phenotype during naturalistic settings. Using such personal devices allows for the continual collection of a person's activity, such as step count and exercise patterns, health signals, such as sleep and heart rate, and social behaviours (how many calls and text messages are sent). Even though it is more challenging to analyse, the data variability offers a unique opportunity to describe the person's lifestyle and behaviour in-situ [2], [38].

1.1.1 A Mental Health Perspective

Mental health conditions are increasing worldwide. According to the World Health Organisation's (WHO) report, in 2019, 1 in every eight people suffered from a mental disorder, with anxiety and depressive disorders the most common [77]. After 2020, due to the COVID-19 pandemic, a larger than 25% increase in the number of people suffering from anxiety and depression was estimated [144]. There is a significant prevalence of mental health conditions among children and adolescents - around 20% worldwide have a mental health condition, with suicide being the second leading cause of death among 15-29-year-olds [142].

Mental disorders can substantially affect all areas of life, such as school or work performance, social relationships, and the ability to participate in the community. Moreover, they are extremely costly to the global economy: depression and anxiety alone cost the global economy USD 1 trillion each year [143]. Technology can offer positive possibilities in the field of mental healthcare [96]: reduction of costs, availability at all times, defeating stigma, and early prevention.

Traditionally mental health-related assessments are done through face-to-face meetings in clinical settings. The commonly used measurement modalities (e.g., self-report, proxy-report, clinician ratings [130]) are time-consuming and tedious to fill in on follow-up visits [149]. In addition, several self-report methods are prone to biases [3], [5], [68], [154]. When patients are asked to retrospectively report their feelings and the situations which provoked those feelings during their everyday routines, they may not accurately remember [170] or deliberately omit details of previous events [195]. Furthermore, questionnaires usually focus on a summarised version of the events within a particular duration; hence, variations in behaviours and determinants over time and context may be overlooked. Therefore, there is a great need and much to be gained from defining a way to assess the evolution of patients' mental states continuously, ideally in their everyday environment, to support the monitoring and treatments by health care providers.

Ecological momentary assessment (EMA) allows for more continuous evaluation and monitoring of patients without face-to-face appointments. It has the crucial advantage of providing data that is more relevant to daily life [182]. It does not rely on the patient's memory of the events; hence, it is less susceptible to recall bias and may provide insights into the time-varying dynamics of behaviour and its correlations. However, it still requires active patient input leading to refusal and attrition [40], [199]. Thus, developing adequate passive EMA tools may increase retention and help overcome the limitations of active EMA [153].

In the field of mental health, the individuals' unique digital phenotype can provide clues to infer their behaviours, emotions, and feelings and, ultimately, to detect symptoms early and prevent mental health disorders [59], [94], [128], [191] (see Figure 1.2). A smartphone-based approach may be valuable in gathering long-term objective data, aside from the usually used self-ratings, to predict clinical state changes and investigate causal inferences about state changes in patients (e.g., those with affective disorders) [47].

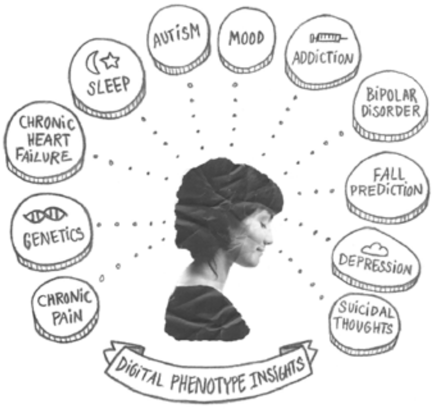


Figure 1.2: Medical insights that can be assessed from the digital phenotype. Source [1].

Solutions for digital phenotyping in the mental health field aim to analyse various aspects of human behaviour. Several studies from the past years have shown that passively-sensed data from smartphones and wrist-wearables were associated with symptoms of schizophrenia, bipolar disorder, and depression [13], [104], [175], [186]. Machine learning offers the possibility to develop approaches for analysing these vast data sources. Integrating these technologies in the clinical workflow may change the nature of identification, follow-up, and treatment of mental disorders. For example, the early identification of behavioural markers of psychiatric disorders may allow clinicians to react early to patients' needs and deliver personalised treatment.

1.1.2 The Complexity and Nuances of Passive Data

Digital phenotyping involves data collection through an application the patient downloads and installs on their smartphone after first participating in the study. It is paramount that participants understand the nature of the collected data and for what purpose it is collected. They can leave the study at any point and delete the application, which will stop any ongoing data collection. We can differentiate two main categories of data: *active* (e.g. taking surveys, contributing audio diary entries) and *passive* (originating from smartphone sensors and logs). While active data collection imposes at least some subject burden, passive data collection does not, as it originates from smartphone sensors (such as GPS and accelerometers) and smartphone logs (such as communication logs and screen activity logs). Thus passive data constitute an objective measurement of different aspects of social, behavioural, and cognitive functioning.

Being objective does not imply that passive data collection is also perfect. It has several challenges: some sensors generate vast volumes of data, and others cause significant battery drain; analysis of raw passive data is complicated, and collection of certain types of data may interfere with the phenotype of interest [141]. From a data analysis perspective, one of the most challenging aspects is propagating the uncertainties involved at different stages of the data collection due to the differences in how each individual uses their smartphone. For example, where people typically carry their devices varies from individual to individual: the phone might not be on the person at all times, some people might turn their devices off for the night, and so on. These aspects do not invalidate inferences drawn from such data, but they do complicate them.

Missing observations are another critical point to consider in smartphone-based digital phenotyping. Some missingness is expected by design, as sensors have varying sampling frequencies. For example, as GPS drains the phone battery quickly, GPS data can only be sampled at a lower frequency. In contrast, missingness can be caused by sensor non-collection due to technological and behavioural factors. For example, participants may forget to charge their phones, disable the GPS, or uninstall the study application. Due to performance considerations, the operating system may also limit sensor access during specific conditions. Missing data presents various problems: it reduces statistical power, the lost data can cause bias in the estimation of model parameters, it can reduce the representativeness of the samples, and it can have a significant effect on the conclusions that can be drawn from the data. In the case of GPS data, for instance, ignoring the missingness or using simple imputation techniques, such as linear interpolation, led to a 10-fold error variance for daily summary statistics [14].

Finally, as the variety of smartphone data (the availability of information about patient location, movement patterns, activity level, and social engagement) is unparalleled, discerning clinically relevant or meaningful information requires careful consideration and appropriate statistical/machine learning methods [189].

1.2 The Passive Monitoring Dataset

1.2.1 Study Participants

The data used in this study were collected from two ongoing studies involving passive smartphone monitoring of clinical outpatients of two public mental health hospitals, *Hospital Universitario Fundación Jiménez Díaz* and *Hospital Universitario Rey Juan Carlos*, Madrid. Patients were invited to participate in the data collection process by their clinicians. The research followed the code of ethics defined in the Declaration of Helsinki by the World Medical Association.

Patients were included in the study if they were at least 18 years old clinical outpatients diagnosed by specialists at the institutions mentioned above with mental disorders or were attending therapy groups (such as support groups for cyberbullying and relaxation) at these institutes. They had to own a smartphone running on Android or iOS operating systems, which they connected to a Wi-Fi network at least once weekly. None of the patients was paid for participating in the study.

1.2.2 Data Collection

A clinically validated eHealth platform, *eB2 MindCare* [21], [30], was used to collect the participants' passive data. After installing the app, the users undergo an onboarding phase, where they are asked to permit specific data collection streams (the mobile phone's sensors, Google Fit, and wearables such as Fitbit and Garmin), depending on their preferences and what is accessible within the operating system of their devices.

The *eB2 MindCare* app collects data from the different sources (see Table 1.1) at varying intervals. The raw sensory data is transformed into human-understandable digital biomarkers that can be used for further analysis. These biomarkers can then be extracted as 48-half-hour slot daily summaries or the overall daily summary. In the following, we describe some of the biomarkers that were used in this work.

Table 1.1: Features of interested collected by the *eB2 MindCare* app. Source [30].

Sensor	**Collected data**	**Source**
GPS	Location (altitude, longitude, latitude, accuracy)	Smartphone
Light sensor	Environmental light detected	Smartphone
Accelerometer, actigraphy	Physical activity	Smartphone, Fitbit, Garmin, Health, Google Fit
Pedometer/accelerometer	Step count	Smartphone, wearables, Health, Google Fit
Applications	Applications use time	Smartphone
Light sensitive photodiodes, actigraphy, accelerometer	Sleep duration	Wearable, Health, Google Fit
Pedometer, accelerometer	Distance walked	Smartphone, Google Fit, wearable, Health

The step count information is recorded every 5 minutes, and the daily summary value corresponds to the sum of the registered entries. The app use information is gained similarly. The devices register sport-related activities on change, and the daily summary encompasses the total number of times each action was performed.

Distance information is gathered every minute, whereas location data is gathered at 5-minute intervals. Locations are obfuscated with an offset and randomly rotated to protect users' data. The daily travel distance and the number of visited locations are computed from these sources. Time spent at home is computed using clustering based on the most common user locations throughout the day.

There is a hierarchical set-up for hours of sleep for the credibility of different sources; if data is manually introduced by the user or calculated by the phone but confirmed by the user, that value is first considered. Otherwise, the following ordering holds: sleep data by iOS, sleep data by Garmin, sleep data by Fitbit, sleep data calculated from light, app use and steps data, and sleep data calculated by the phone.

The application also allows users to record information about their medication (name, frequency and quantity of intake), quality and duration of sleep, and emotional state (angry, disgusted, scared, sad, overwhelmed, tired, grief, neutral, relaxed, motivated, happy, and delighted) multiple times during the day. This active data source, however, requires manual input and is usually based on the patient's subjective perception.

In addition, clinicians used the *MEmind* [15], [132] electronic health tool to record the patients' socio-demographic and clinical information from all participants. Socio-demographic data included age, gender, household composition, marital status, and employment status. Clinical data entailed the International Classification of Diseases, tenth revision, psychiatric diagnoses grouped into the following categories: (1) anxiety, stress, and trauma-related disorders; (2) unipolar or bipolar mood disorders; (3) personality disorders; (4) substance use disorders; (5) psychotic disorders, and (6) other disorders. The clinicians filled in socio-demographic information together with a first completion of the evaluations when the patients enrolled in the study. The follow-up scores were recorded at an in-person appointment or via a phone call.

1.2.3 Data Description

The mobile sensed data analysed in the different stages of this work were collected between January 2016 - April 2022 from 2300 individuals, yielding over 510k samples (31.5% collected in 2019). Even though the sensory data is recorded at pre-determined frequencies, the missingness problem due to sensor non-collection caused by technological and behavioural factors cannot be avoided. The final data set contained a large percentage of missing observations. Moreover, missingness patterns can appear simultaneously across different attributes.

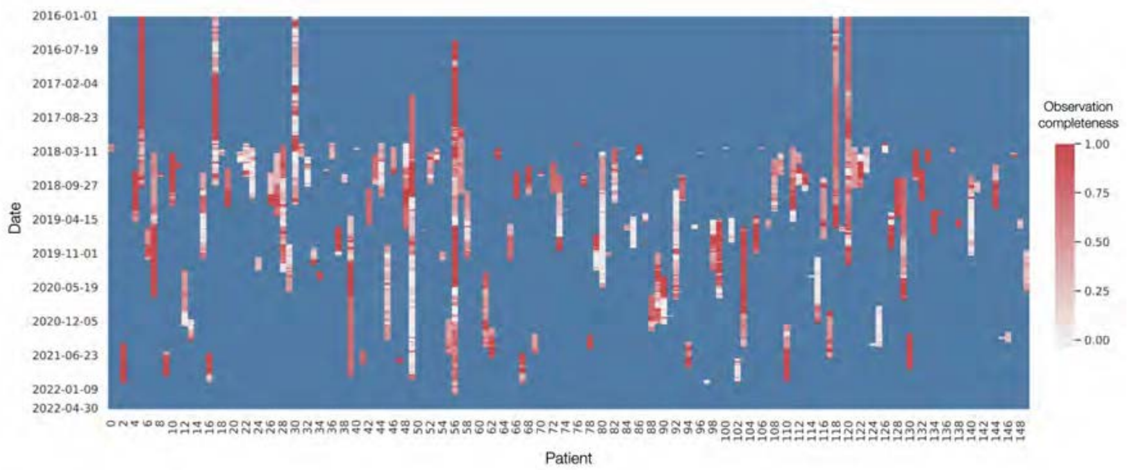


Figure 1.3: An overview of the missingness in the mobile sensing database. Periods of no collection are indicated with blue. White indicates a complete missing day, and the lighter the red shading, the larger percentage of features is missing for that day.

Figure 1.3 illustrates a subgroup of the study population, and the missing pattern in the daily summary of four mobility descriptor features: step count, distance travelled, time spent at home and exercising. The blue areas show periods of no data collection for the specific patient (e.g. they were not enrolled yet, or they dropped out of the study). We can see that there are almost no days with complete observations for all the features of interest, which calls for reliable techniques to deal with the missingness.

Suppose we visualise the univariate distribution of all daily summary variables in a dataset along with their pairwise relationships. In that case, we can understand how the variables are distributed - their range, central tendency, skewness, and outliers. Additionally, we gain insights into whether the distributions vary across subsets defined by other variables.

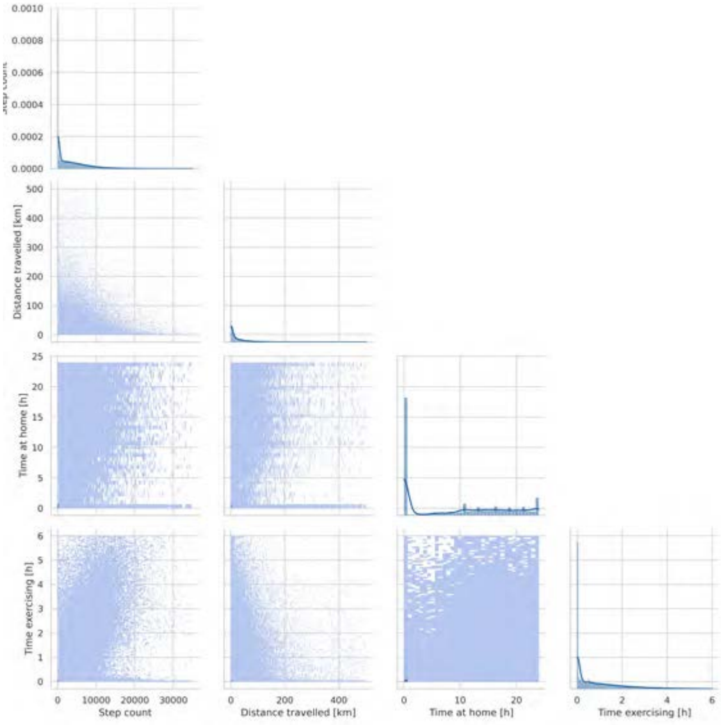


Figure 1.4: Distribution of the daily summary values of the mobile sensed variables. Distance travelled is expressed in kilometres, while the time-related measures in hours.

Figure 1.4 shows the feature distributions on the diagonals using kernel density estimation and their pairwise relationships using histograms. The mobile sensed data is noisy, contains infeasible values, and the distributions are skewed in most cases. Therefore, before applying any modelling techniques, we must clean the data, i.e., process raw data to the usable form. We do this by replacing unfeasible values and extreme outliers (values which lie more than 3.0 times the interquartile range below the first quartile or above the third quartile) with NaNs; hence consider them to missing. We opt for this approach, instead of truncating the variables due to the large distribution shift the latter would cause.

Moreover, the data sources are highly heterogeneous, showing a mixture of numerical (e.g. age, mobile sensed data), binary (e.g. cohabiting status), categorical (e.g. gender, employment status) and other feature types. Heterogeneity is seen in machine learning-based modelling as a particular source of complexity, making it challenging to build credible and consistent models. Therefore, we aim to provide technically robust solutions to structuring such data.

In the previous section, we mentioned a series of nuances from passive data collected in the wild, which we exemplified here with a concrete real-world dataset. The inherent problems coming from the irregularly sampled, high-dimensional, and heterogeneous time series data, which is often only partially observed, are just one side of the main difficulties we faced throughout this work. On the other hand, the lack of labelled data also reduced the modelling possibilities. Also, the outcomes used as ground truth values are not objective, and there are several aspects of patients' lives that we cannot account for.

1.3 Overview of Models and Contributions

This thesis aimed to apply supervised and unsupervised machine learning methods to model and understand mental disease evolution based on the digital phenotype of patients collected from their smartphones and wrist-wearables and clinician assessments at the follow-up visits, which provide ground truths. A brief overview of the proposed models and principal contributions is provided in the following.

1.3.1 Modelling Passively-Sensed Data

Dealing with Missing Data

The best possible way of handling missing data is to prevent the problem in the first place; however, this is a close to impossible task in real-world passive data collection. As discussed in the previous section, the missing data problem is common in digital phenotyping, and the reasons for missing data is manifold. There are several methods to handling missingness, from simple approaches, such as complete case analysis, single imputation using constants or statistical metrics of the features (mean, median, mode), expectation-maximisation based approaches, multiple imputation techniques, and others [85]. Especially in the scenarios where a significant fraction of values is missing, there is a high variance in the imputation performance and varying impact on predictive performance in downstream ML tasks.

Here, we use simple probabilistic latent variable models for data imputation and feature extraction, namely, hidden Markov models (HMM) [156]. HMMs are a temporal version of mixture models (MM) and are generative models characterised by observable sequences. The discrete states of the HMM are assumed to have been generated by a first-order Markov chain process, and each observation depends only on the paired state. An HMM comprises an initial state probability distribution, a state transition probability distribution, and a symbol emission probability distribution.

Different frameworks that implement these well-known models are publicly available. However, these implementations usually lack the features required to use these models with real-world datasets: missing data inference, ability to manage heterogeneity, semi-supervised training support to increase hidden states' interpretability, and synthetic data generation. Our contribution in this area is to provide an extension of existing HMM implementations via a Python package, PyHMM, including the previously enumerated features.

Temporal Data Modelling

Temporal data modelling is relevant in many fields and has been studied for decades. Solutions vary from simple approaches focusing on manually defined features that summarise the temporal sequences combined with classical machine learning models (linear regression, support vector machines, decision trees) to complex deep learning methods that automate the feature extraction process [87].

Throughout this work, we employ feature extraction and deep learning-based modelling techniques, depending on the data at hand. The first approach is more suitable for small data, is cheaper to perform and generally has easier to interpret outcomes. On the other hand, the second approach is more suitable for high complexity problems, allows for complex features to be learned, but requires significant computational power and lacks straightforward interpretability.

1.3.2 Predicting Personalised Mental Health Outcomes

The main contributions of this thesis are the applications of existing machine learning and deep learning techniques for predicting mental health outcomes of psychiatric outpatients using various sources of mobile sensed data.

A commonly used disease or clinical outcome prediction approach is the one-size-fits-all model [135]. In such models, all available patient cohort data is used to train a global model and perform predictions for each patient. The advantage of these global models is that they can capture the information from the whole cohort; however, they may miss some patient-specific information (phenotype differences, the presence of different diseases).

Albeit, patient-specific models, tailored to each individual, are essential in personalised medicine, as these models can improve predictive performance over global models [32], [90], [101]. A generic framework, derived from the working process of clinicians, consists of two steps: first, a similarity measure among patients has to be defined, and then for each patient, a separate model has to be built based on their similarity cohorts. In the case of longitudinal electric health record (EHR) data, the number of patient visits varies primarily due to patients' irregular visits and incomplete recordings. Thus, one of the main challenges in measuring patient similarity is deriving an adequate representation of each patient without losing his/her historical information.

In this work, we compare global and local models when the dataset allowed (there was more than one outcome label for most patients in the cohort), but in most experiments, the sample size per individual only allowed for global approaches.

Emotional State Prediction

First, we worked on a generic machine learning-based approach for emotional state prediction using passively collected data from mobile phones and wearable devices and self-reported emotions by patients. Emotional state prediction and forecasting could be used as early warning signs in clinical treatment. Detecting prominent affective episodes risk could help catch the early onset of major depressive or manic phases that can be addressed and handled in time, reducing the severity of symptoms and the degree of treatment.

We applied probabilistic latent variable models (MM and HMM) for data averaging and feature extraction on the regularly sampled but frequently missing and heterogeneous time series data. The extracted features were combined with a classifier to provide emotional state predictions. Furthermore, we proposed a personalised Bayesian model to improve the performance, which considers the individual differences in the data by applying a different classifier bias term for each patient.

Probabilistic generative models proved good as pre-processing and feature extractor tools for data with large percentages of missing observations. Models which took into account the posterior probabilities of the MM/HMM latent states outperformed those which did not, suggesting that the underlying behavioural patterns identified were meaningful for individuals' overall emotional state. Moreover, the proposed personalised models demonstrated that accounting for individual differences through a simple hierarchical model substantially improves emotional state prediction performance without relying on previous days of data.

Anxiety and Functional Disability Assessment

Anxiety symptoms during public health crises are associated with adverse psychiatric outcomes and impaired health decision-making. The interaction between real-time social media use patterns and clinical anxiety during infectious disease outbreaks is underexplored. So in a smaller project, we aimed to evaluate the usage pattern of 2 types of social media apps (communication and social networking) among patients in outpatient psychiatric treatment during the COVID-19 surge and lockdown in Madrid, Spain and their short-term anxiety symptoms (7-item General Anxiety Disorder scale) at clinical follow-up.

A machine learning–based approach that combined a hidden Markov model and logistic regression was applied to predict clinical anxiety and nonclinical anxiety, based on longitudinal time-series data that comprised communication and social networking app usage (in seconds) as well as anxiety-associated clinical survey variables, including the presence of an essential worker in the household, worries about life instability, changes in social interaction frequency during the lockdown, cohabitation status, and health status. Patients who reported severe anxiety symptoms were less active in communication apps after the mandated lockdown and more engaged in social networking apps overall, suggesting a different pattern of digital social behaviour for adapting to the crisis. Predictive modelling using digital biomarkers—passive-sensing shifts in category-based social media app usage during the lockdown—can identify individuals at risk for psychiatric sequelae.

The second topic tackled is functional disability assessment based on WHODAS 2.0 outcomes. Functional limitations are associated with poor clinical outcomes, higher mortality, and disability rates, especially in the elderly. Continuous assessment of patients' functionality is essential for clinical practice; however, traditional questionnaire-based assessment methods are very time-consuming and infrequently used. Mobile sensing offers a great range of sources that can assess function and disability daily.

The first part of the work aimed to prove the feasibility of an interpretable machine learning pipeline for predicting WHODAS 2.0 outcomes using passively collected digital biomarkers. One-month long time-series data were summarised using statistical measures (minimum, maximum, mean, median, standard deviation, IQR), creating 64 features. We then applied a sequential feature selection to each WHODAS 2.0 domain (cognition, mobility, self-care, getting along, life activities, participation). Finally, we predicted the WHODAS 2.0 functional domain scores using linear regression using the best feature subsets. Our findings show the feasibility of using machine learning-based methods to assess functional health solely from passively sensed mobile data. The feature selection step provides a set of interpretable features for each domain, ensuring better explainability of the models' decisions.

Additionally, we aimed to include more information about the intra-day variability of the digital biomarkers and also the socio-demographic background of the patients. Therefore, we propose a Long Short-Term Memory (LSTM) neural network-based pipeline for predicting mobility impairment based on WHODAS 2.0 evaluation from the raw digital biomarkers that provide insights into the patients' behaviours on a half-hour scale. We address the missing observation problem utilising hidden Markov models and the possibility of including information from unlabelled samples via transfer learning. Finally, we also show that our multi-modal pipeline can be easily fine-tuned to predict the GAD-7 outcomes.

1.4 Document Structure

The structure of this doctoral manuscript is divided into two main parts: Chapter 2 provides an overview of the theoretical background and proposed inference and learning methods, while Chapters 3-5 present the different applications in the mental health field where these methods were applied. The obtained results and the drawn conclusions are also listed for each project. The thesis ends with an overall summary and discussion of further work possibilities.

CHAPTER 2

MODELLING PASSIVELY-SENSED DATA

2.1 The Missing Data Problem

A commonly faced problem in real-world applications is the occurrence of missing data. By definition, missing data is the data value that is not stored for a variable in the observation of interest [69]. The presence of missing data can significantly affect the conclusions drawn from the data, as it reduces statistical power, the lost data can cause bias in the estimation of model parameters, and it can reduce the representativeness of the samples. Accordingly, several studies have focused on handling the problems caused by missing data and the methods to avoid or minimise the number of misses during data collection in medical research [39], [139] and beyond [20], [102].

2.1.1 Types of Missing Data

Two types of missing data can be differentiated based on assumptions about the mechanism of the missingness: ignorable, with the subtypes missing completely at random (MCAR) and missing at random (MAR), and non-ignorable or missing not at random (MNAR) [164]. As defined by [179], “the missingness mechanism concerns whether the missingness is related to the study variables or not“. Assuming a partially observed dataset in matrix form, $\mathbf{X} \in \mathbb{R}^{n \times p}$, where the n is the sample size and d is the number of variables that have been measured, we denote an observed entry of variable j as $\mathbf{X}_j^{obs}$, and a missing value as $\mathbf{X}_j^{miss}$, with $j = 1, \ldots, d$, and use this notation to describe the missingness mechanism in a probabilistic fashion.

The data is considered MCAR when the probability of missing data is not related to either the observed or the unobserved features:

$$Pr\left[X_j^{miss} \mid X_1, ..., X_p\right] = Pr\left[X_j^{miss}\right]. \qquad (2.1)$$

This implies that the probability of being missing is the same for all the units. If observations are missing by design due to a sensor failure or faulty transmission, they can be regarded as MCAR. When data is missing completely at random, although statistical power may be lost in the design, there is no bias in the estimated parameters due to the absence of the data.

Data is assumed to be MAR when the missingness is not random. However, it is related to the observed features only.

$$Pr\left[X_j^{miss} \mid X_1, ..., X_p\right] = Pr\left[X_j^{miss} \mid X_1, ..., X_{j-1}, X_{j+1}, X_p\right] \tag{2.2}$$

Since MAR is an assumption that is impossible to verify statistically, we must rely on its substantive reasonableness. Although typically randomness is considered not to produce bias, MAR does not mean that the missing observations should completely be ignored.

Suppose the data characteristics do not meet those of MCAR or MAR. In that case, they fall into missing not at random (MNAR). One of the implications of MNAR is that missing entries have a different distribution than the observed ones, even when they otherwise have the same characteristics. The MNAR mechanism of missingness is non-random and cannot be considered ignorable. The data that cause others to be missing are unobserved, and obtaining an unbiased estimate of the parameters can only be obtained by modelling the missing data.

2.1.2 Common Techniques for Handling the Missing Data

First and foremost, preventing the problem of missing data by designing well-defined data collection processes is the best way to deal with missing data in clinical research [46], [171], [201]. Even with a good study design, especially in the case of data collected in the wild (e.g. behavioural biomarkers collected passively in the patients' everyday lives), it is not uncommon to have many missing observations.

Techniques for handling the missing values should be robust to the problems caused by the missing data. Hence mild to moderate violations of the assumptions should produce little to no bias or distortion in the conclusions drawn from the population. However, it is not always achievable to use such techniques. Therefore, many alternative ways of handling the missing data have been developed. Figure 2.1 provides an overview of commonly applied methods for dealing with the problems caused by the missing data.

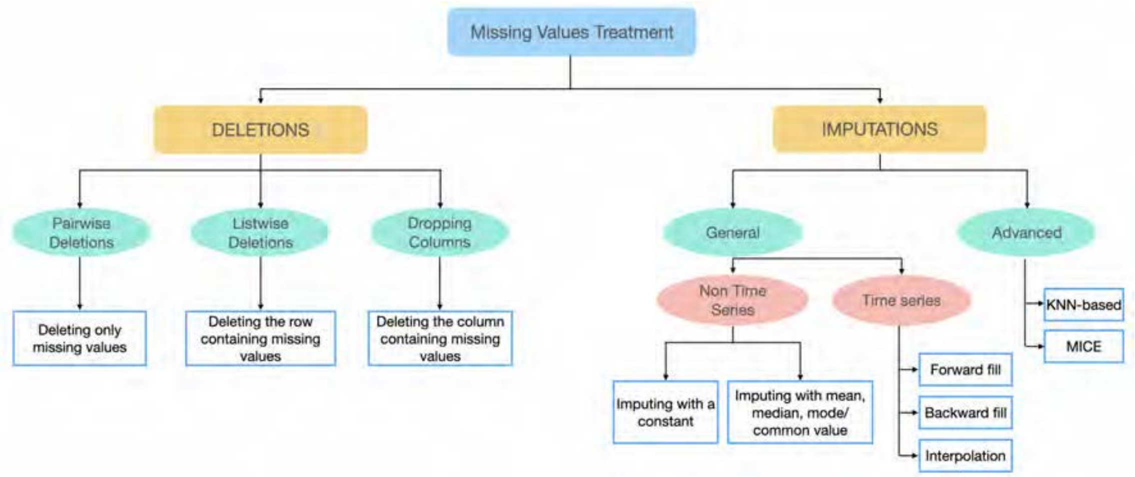


Figure 2.1: Common techniques for handling missing values. Source: Kaggle

Suppose there is a large enough sample and the assumption of MCAR or MAR is satisfied. In that case, deletion may be a reasonable strategy to remove incomplete observations. However, when there is not a large sample, or if there are many missing observations, deletion is not the optimal strategy [93].

Imputing with a constant or the mean or median is a common approach, as it is easy to implement, and under the MCAR assumption, it is fair to consider that the missing values are most likely very close to the mean or median of the distribution (i.e. the most frequent/average observation). These approaches, however, may lead to inconsistent bias [117].

The most widely used imputation techniques are forward and backwards fill when dealing with time-series data. These methods replace the missing value with the last observed value or the next observed value. Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings, so they should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified [107].

Interpolation is as also commonly used for time series. It uses non-missing values from adjacent data points to compute a value for a missing data point based on a polynomial relationship. This approach avoids significantly altering the standard deviation or the shape of the distribution; however, no novel information is added while the sample size has been increased and the standard error is reduced.

Advanced missing data imputation methods can be partitioned into two main categories: global and local missing data imputation methods [33], [51]. Iterative imputation [108], [150], [194] and expectation-maximisation (EM) imputation [89], [173] are considered global strategies. They use information about the whole data set's correlation structure to impute the missing observations encountered in the data set. Local missing data imputation strategies, such as the k-nearest neighbour-based imputation [16], [111], [157], use only similar entries to the missing ones to impute missing values. In general, imputation based on similar observations is more accurate than imputation based on the entire data set's [158], [205].

Multiple imputation [178] is another valuable strategy for handling the missing data. The missing observations are replaced with a set of plausible values in multiple imputation. The benefit of the approach is that it incorporates the uncertainty due to the missing data, which results in a valid statistical inference. Furthermore, it is robust to the violation of the normality assumptions and produces relevant results even in small data sets or in the presence of a large amount of missing data.

2.1.3 Probabilistic Generative Models for Dealing With Missing Data

Probabilistic generative models can learn the underlying distributions in a data set by adjusting the model parameters to best account for the data to maximise the evidence, even in the presence of missing data [45]. Mixture models (MMs) [22] and hidden Markov models (HMMs) [156] are frequently used types of such models.

Mixture Models

MMs comprise a single state with a finite number of components, possibly different distributional types, that can describe different data features, which we'll denote $1, ..., K$. The data can then be modelled in a mixture of several components, each with a simple parametric form (such as a Gaussian). The model is formulated in terms of latent variables, which represent the component each data point was sampled from. A MM assumes the data is generated by the following process: first we sample z, and then we sample the observables $\mathbf{x}$ from a distribution which depends on z, i.e

$$p(z, x) = p(z)p(x \mid z). \tag{2.3}$$

In general, the probability density function (PDF) over x can be computed by marginalising out, or summing out, z:

$$\begin{aligned} p(z, x) &= \sum_{z} p(z)p(x \mid z) \\ &= \sum_{k=1}^{K} Pr(z = k)p(x \mid z = k) \end{aligned} \tag{2.4}$$

The model parameters can be learned from the observed features, referred to as observables, by adjusting the model parameters, which define the observable emission probabilities, such that the MM best accounts for the data in the sense of maximising the evidence using the expectation-maximisation (EM) algorithm [204]. The EM algorithm is a general method for finding the maximum-likelihood estimate of the parameters of an underlying distribution of a given data set. The two steps of the EM algorithm are

- **E-step**. Compute the expectations of the latent variables

$$r_k^{(i)} \leftarrow Pr(z^{(i)} = k | x^{(i)}) \tag{2.5}$$

- **M-step**. Compute the maximum likelihood parameters

$$\theta \leftarrow arg \max_{\theta} \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \left[\log Pr(z^{(i)} = k) + \log p(\mathbf{x}^{(i)} \mid z^{(i)} = k) \right] \tag{2.6}$$

To find the model parameters θ, the two steps are iteratively repeated until a the convergence criteria on the likelihood is met.

Hidden Markov Models

HMMs can be seen as temporal MMs that are commonly used for time-series analysis. These are generative models characterised by a vector of parameters θ, and a set of discrete states $S_{1:T} = (S_1, ...S_T)$, where $S_t \in \{1, 2, ..., K\}$. The states of the HMM are assumed to have been generated by a first-order Markov chain process, i.e,

$$p(S_t \mid S_{1:t-1}, \theta) = p(S_t \mid S_{t-1}, \theta), \quad t = 2, ...T \tag{2.7}$$

and that the observations Y_t are conditionally independent given the paired states

$$p(Y_t \mid S_{1:t-1}, Y_{1:t-1}, \theta) = p(Y_t \mid S_t, \theta), \quad t = 1, ...T \tag{2.8}$$

where $Y_{1:t-1}$ represents the output sequence $(Y_1, Y_2, ...Y_{t-1})$, which is an uni-variate or multivariate time series of observations. Given the above conditionals, the joint distribution of observations and states can be expressed as

$$p(Y_{1:T}, S_{1:T} \mid \theta) = p(S_1 \mid \theta)p(Y_1 \mid S_1, \theta) \prod_{t=2}^{T} p(S_t \mid S_{t-1}, \theta)p(Y_t \mid S_t, \theta) \tag{2.9}$$

and the marginal distribution of the observations in function of the model parameters (likelihood function) as

$$L(\theta \mid Y_{1:T}) = \sum_{s_{1:T} \in \mathcal{S}^{\mathcal{T}}} p(Y_{1:T}, S_{1:T} = s_{1:T} \mid \theta) \tag{2.10}$$

where the summation is over all possible state sequences $\mathcal{S}^{\mathcal{T}}$.

The parameters of HMMs are found using a special case of the EM algorithm, called the Baum-Welch algorithm, that makes use of the forward-backward algorithm to compute the statistics for the expectation step.

If the observable response variable Y has missing values, it can be partitioned into $\mathcal{Y}_{obs} \subseteq Y_{1:T}$, the observed part, and $\mathcal{Y}_{miss} \subseteq Y_{1:T}$, the missing part. We also define an indicator $M_{1:T}$ with values $M_t = 1$ if $Y_t \subseteq \mathcal{Y}_{miss}$, and $M_t = 0$ otherwise.

The full likelihood, which also depends on the hidden states, can be expressed as

$$L_{full}(\theta, \phi \mid \mathcal{Y}_{obs}, M_{1:T}) \propto \textstyle\sum_{s_{1:T} \in \mathcal{S}^{T}} \int p(\mathcal{Y}_{obs}, \mathcal{Y}_{miss}, s_{1:T} \mid \theta)p(M_{1:T} \mid \mathcal{Y}_{obs}, \mathcal{Y}_{miss}, s_{1:T}, \phi)d\mathcal{Y}_{miss} \tag{2.11}$$

When the data is MAR, missingness is ignorable in the inference of θ. If the data is not MAR, hence missingness is not ignorable, we assume conditional independence between

M and Y given the hidden states [180]:

$$p(M_t, Y_t \mid S_t) = p(M_t \mid S_t)p(Y_t \mid S_t) \tag{2.12}$$

Then the full likelihood becomes

$$L_{full}(\theta, \phi \mid \mathcal{Y}_{obs}, M_{1:T}) \propto \sum_{s_{1:T} \in \mathcal{S}^T} p(M_{1:T} \mid s_{1:T}, \phi) \times \int p(\mathcal{Y}_{obs}, \mathcal{Y}_{miss}, s_{1:T} \mid \theta) d\mathcal{Y}_{miss} \tag{2.13}$$

$$\sum_{s_{1:T} \in \mathcal{S}^T} p(M_{1:T} \mid s_{1:T}, \phi) \times p(\mathcal{Y}_{obs}, s_{1:T} \mid \theta \tag{2.14}$$

The number of hidden states can be chosen to allow for intricate patterns of (marginal) dependence between M and Y at a single time point, as well as over time.

Model Order Selection

To select the optimal number of MM components/HMM states, the Akaike information criterion [4] and the Bayesian information criterion [174] can be used. The AIC selects the model that minimises the Kullback-Leibler divergence, i.e.,

$$M^*_{AIC} = arg \min_{M_i} [-2l_{M_i}(\theta^*_i) + 2K_i] \tag{2.15}$$

where $l_{M_i}(\cdot)$ is the log-likelihood of the model M_i, θ^*_i is the maximum likelihood estimator of θ_i, and K_i is the number of parameters of the underlying distribution of the observation process. The best model is the one which has the weakest AIC. This criterion uses maximum likelihood principle, however, unlike the latter, it penalises models with too many variables.

The BIC uses a Laplacian approximation to select the model that maximises the Bayesian posterior probability, i.e.,

$$M^*_{BIC} = arg \min_{M_i} [-2l_{M_i}(\theta^*_i) + K_i \ln N] \tag{2.16}$$

Like in case of applying AIC, the best model is the one with the smallest value of BIC. This criterion penalises stronger over-parameterised models, and is more relevant for over-learning models.

Data Imputation

Once the generative models are trained on the data set with missing values, they can be used to generate samples for imputation.

In the case of MMs, first, the posterior distribution $p(z \mid x)$ needs to be inferred for each observation to find which component the observation is most likely to belong to. Just like in Bayesian parameter estimation, we can infer the posterior distribution using Bayes' Rule:

$$p(z \mid x) \propto p(z)p(x \mid z) \tag{2.17}$$

Hence, we evaluate the right-hand side for all values of z, and then renormalise so that the values sum to 1. Then, the missing attributes can be imputed by a sample generated from the most probable component.

When using HMMs, all observation sequences must first be decoded using the Viterbi algorithm on the trained HMM. This method finds the most likely state sequence in the maximum a posteriori probability sense that could have resulted in the given observation sequence. This most probable hidden state sequence can be used for recovering the missing observations with a value generated by the distribution corresponding to the hidden states for each time step.

Sampling from the most probable distribution yields valid and robust parameter estimates and explicit imputed values for variables that can be analysed as outcomes or predictors. The imputation process can be repeated several times, creating multiple datasets, hence, thereby accounting for the uncertainty in the imputed values and implicitly augmenting the data. Moreover, they are robust to moderate deviations of the observed data from the assumed underlying distribution [83] and provide accurate estimates even when the proportion of missingness is high [9], [113].

2.2 Temporal Data Modelling

Time series data are among the most ubiquitous types that capture information in most areas of life. It can be defined as a special type of data set in which one or more variables are measured over time [24]. Mathematically we can define a set of time series as $D = \{x_i\}_{i=0}^{N}$, $x_i \in \mathbb{R}^{d \times t_i}$ and $d, t_i \in \mathbb{N}^*$.

Capturing a sequence of observations indexed by time stamps allows insights into the evolution of the measured quantity. The exponential increase in the volume of data has generated a tremendous opportunity for modelling this type of data with machine learning (ML) methods to automate tasks such as discovering recurrent patterns, correlation analysis, classification, clustering, outlier detection, segmentation, forecasting, and data simulation.

2.2.1 Manual Feature Extraction-Based Approaches

Classical machine learning models (supervised or unsupervised) can only use a well-defined set of feature vectors and not deal directly with data sequences. Therefore, feature extraction must be performed before further modelling can be done. Feature extraction seeks to transform an initial input raw data sequence to generate a new set of features containing meaningful information about the sequence (time, frequency, statistical trends), depending on the nature of the raw input data, the context and domain of the task.

In principle, one might decide to map the set of time series into a design matrix of N rows and M columns by choosing M data points from each time series x_i as elements of a feature vector. However, for more comprehensive insights time series are often characterised with respect to the distribution of the observations, correlation properties, stationarity, entropy, and nonlinear time series analysis [56]. Therefore the feature vector x_i can be constructed by applying feature extraction methods (e.g. statistical measures like in Figure 2.2) $f_j : x_i \rightarrow x_{i,j}$ to the respective time series x_i, which results into a feature vector $x_i = (f_1(x_i), f_2(x_i), \ldots f_M(x_i))$.

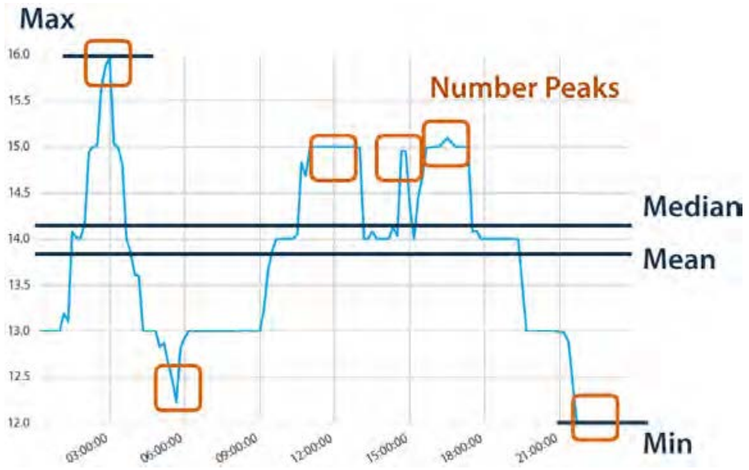


Figure 2.2: Statistical feature extraction from time series. Source: tsfresh

The design and performance of the downstream models are greatly affected by feature extraction, as reducing the information into a lower-dimensional feature space might result in the loss of relevant information, while extracting too many irrelevant features can impair the ability of the methods to generalise. Hence, these feature extraction techniques are often coupled with feature selection methods [65], [100] to find the most relevant subset of features for training the ML models.

2.2.2 Deep Neural Models

Deep learning models provide means to learn temporal dynamics in a purely data-driven manner thanks to their ability to find the appropriate complex nonlinear mathematical functions to turn input into an output.

Recurrent Neural Networks

Recurrent neural networks (RNNs) [165] are one of the first deep neural architectures designed for sequence learning. The outputs of RNNs are not only influenced by the weights associated with inputs like in the case of standard feed-forward neural networks (NN), but the hidden state captures information and allows contextual decisions based on prior inputs and outputs. Over time, as the sequence is processed, the hidden state gets updated. The architecture is shown in Figure 2.3.

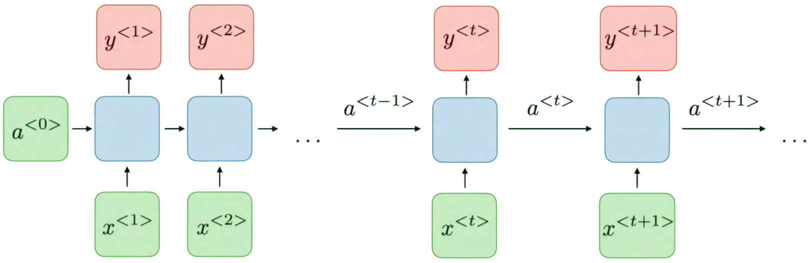


Figure 2.3: Architecture of a traditional RNN. Source: stanford.edu

The activation and the output at each time step t are expressed as

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>}]b_a \tag{2.18}$$

and

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y \tag{2.19}$$

where W_{ax}, W_{aa}, W_{ya}, b_a, b_y are shared temporal coefficients and g_1, g_2 activation functions.

The loss function $\mathcal{L}$ of all time steps is defined based on the loss at every time step as follows:

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}(\hat{y}^{<t>}, y^{<t>}) \tag{2.20}$$

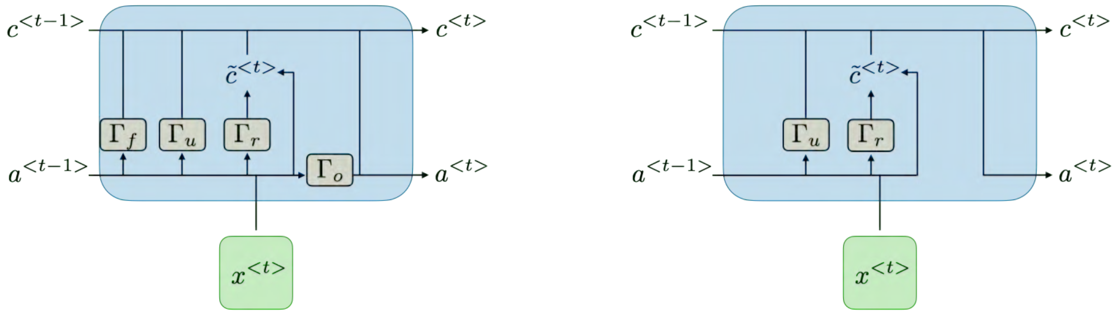
During training, backpropagation is also done through time. The derivative of the loss $\mathcal{L}$ with respect to the weight matrix at timestep T is expressed as

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{t=1}^{T} \left. \frac{\partial \mathcal{L}^{(T)}}{\partial W} \right|_{(t)} \tag{2.21}$$

Although RNNs outperform statistical methods [79], they suffer from two significant drawbacks: exploding or vanishing gradients, which are likely to generate instability, thus influencing the reliability of weight updates, and the inability to exploit information from arbitrarily long time series but only a few time steps [106].

Therefore, gated versions of RNNs, namely long short-term memory (LSTM) [81] and gated recurrent unit (GRU) [35] cells, were proposed to deal with these issues. Figure 2.4 shows the architecture of both units.

Each ordinary node in the hidden layer is augmented with a memory cell $c^{<t>}$. Combining these cells with the different gates ensures that the gradient can pass across many time points without vanishing or exploding. In GRU networks, during training the mod-



(a) Long Short-Term Memory (LSTM).

(b) Gated Recurrent Unit (GRU).

Figure 2.4: Basic architecture of LSTM and GRU cells. Notation: Γ_u = update gate, Γ_r = relevance gate, Γ_f = forget gate, Γ_o = output gate. Source stanford.edu

els learn when to update the hidden state, hence, "remember" important pieces of information, by using the gating unit. In LSTMs, the hidden state $a^{<t>}$ acts as the model's short-term memory, whereas the memory cell $c^{<t>}$ acts as the long-term memory.

These networks are improved versions of RNNs, able to capture long dependencies thanks to the changes introduced in the computation of hidden states and outputs using the inputs. Consequently, most state-of-the-art applications use the LSTM or GRU model as a basis for their design.

Temporal Convolutional Networks

Convolutional neural networks use tied weights to compute a function of a determined local neighbourhood for each input and return an output. They are commonly used for two-dimensional inputs, such as images, but can also be applied to sequence modelling. They do not have to maintain long-term hidden states, so they are an easier-to-train alternative to RNNs.

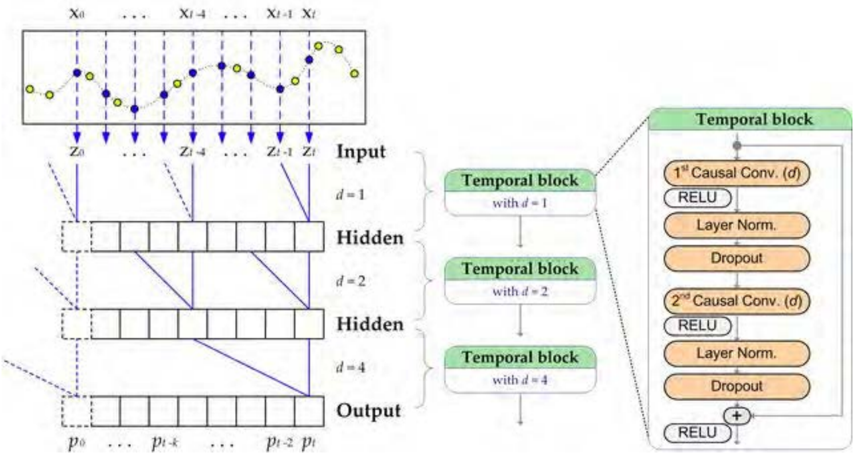


Figure 2.5: The TCN architecture. Source: [12].

Temporal convolutional networks (TCN) [12], depicted in Figure 2.5, are built on two essential properties: no information leakage from future to past and the ability to process sequences of any length and map them to the same length sequence. The first property is achieved by using so-called causal convolutions, meaning that only past elements of the sequence are convolved to get the output. In contrast, the second property holds due to a 1D fully-convolutional network architecture, which results in each hidden layer having the same length as the input layer. TCNs can be built to have a long memory; hence, consider far past inputs when making a prediction when deep networks are augmented with residual layers and dilated convolutions [75].

Several additional approaches combining CNNs for feature extraction and RNNs for capturing temporal dependencies in the data have also been proposed [37], [76], [177], [188].

Attention

A major drawback with the above architectures is that the input sequences are represented as a single vector, which can cause information loss due to the large compression. Especially in sequence-to-sequence models, where the so called encoder-decoder architectures are employed, it is a highly complex task for the decoder to reproduce the original signal from the lower dimensional representation.

The basic idea behind the attention mechanism [10] is to allow the model to pay attention to specific parts of the input that are considered important, rather than looking only at the final output. Instead of encoding the input sequence into a single fixed context vector, the attention model develops a context vector explicitly filtered for each output time step.

The attention output is a sequence of vectors $c^{<t>}$ called context vectors. In the case of RNNs, the context vectors at time t are defined as a simple linear combination of the activations $a^{<t'>}$ weighted by the attention values $\alpha^{<t,t'>}$:

$$c^{<t>} = \sum_{t'} \alpha^{t,t'} a^{t'} \tag{2.22}$$

with $\sum_{t'} \alpha^{<t,t'>} = 1$.

The attention weights are learned using an additional fully-connected shallow network. Computationally they are described as

$$\alpha^{<t,t'>} = \frac{\exp e^{<t,t'>}}{\sum_{t''=1}^{T_x} \exp e^{<t,t''>}} \tag{2.23}$$

A large $\alpha^{<t,t'>}$ attention weight causes the RNN to mainly focus on the activation $a^{<t'>}$, when predicting the output $y^{<t>}$, while a small one causes the activation to be ignored.

Besides helping the models learn the most essential information from the sequences, attention weights can also be visualised and inspected to gain insight into which parts of the input the model thinks are most relevant for generating the corresponding output.

Transformers

The more recently introduced transformer architecture [197] allows for sequence-to-sequence modelling. Although it was initially introduced to tackle problems in natural language processing, they have been repurposed for other temporal modelling problems [200].

Transformer models consist of an encoder-decoder structure composed of multiple identical blocks. Each encoder block is built from a multi-head self-attention module and a position-wise feed-forward network (FFN), while each decoder block inserts cross-attention models between the multi-head self-attention module and the position-wise feed-forward network (FFN). Thus, the need for RNNs is eliminated.

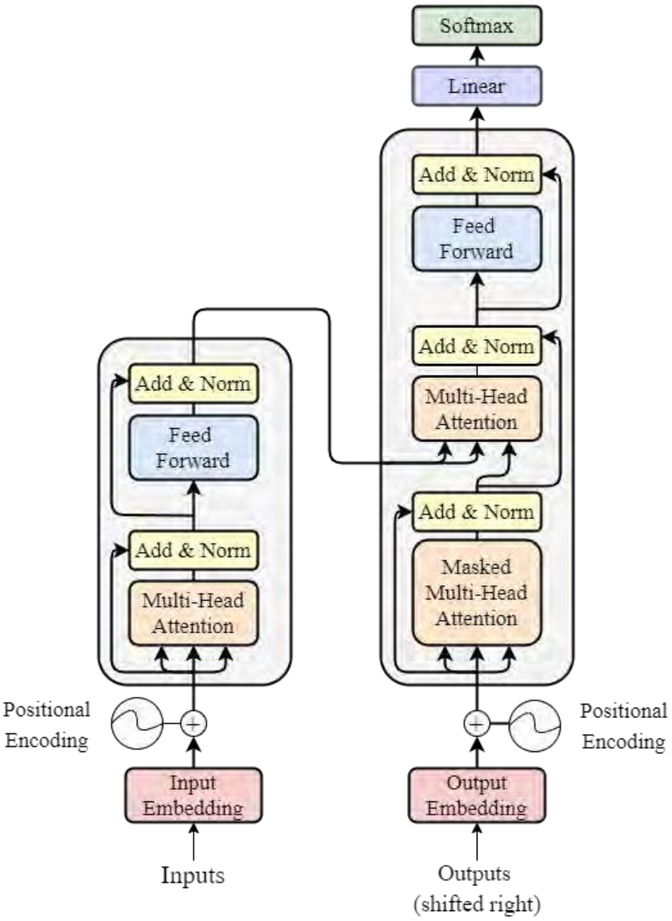


Figure 2.6: The transformer architecture. Source: [197]

While in the case of RNNs, the sequential data is processed element-wise, in transformers, the entire sequence is processed at once, and the positional encoding serves to preserve the sequential information. Processing the whole sequence at once and computing attention weights between the observations enables the model to capture long-term dependencies in the sequence representation.

2.3 Discussion

Passively collected mobile sensed data is noisy and often non-randomly missing; hence, developing robust imputation techniques is a nontrivial problem. However, developing such methods is vital if this information is used to support clinical decision-making.

Imputing missing data using statistical measures such as the mean, median, or even interpolation fails when the percentage of missing data is very high. These approaches can reduce variability in the data set and introduce bias. However, probabilistic generative models can learn the underlying distributions in a data set by adjusting the model parameters to best account for the data in maximising the evidence, even in the presence of missing data. Although these models are simple, they can easily combine different data distributions and can accurately impute values that are far from the overall mean of the data.

This work used probabilistic generative models not only for data imputation but for latent state (feature) extraction too. We will show in the following chapters how in specific set-ups, the latent representation provided by such models can be used as a proxy for the noisy temporal data for different prediction tasks.

Feature extraction from time-series data is time-consuming and complex, challenging the machine learning-based analysis of such data sets. The recent unfolding of deep learning has led to a vast increase in time series models. The ability of deep neural networks to alleviate the preliminary feature engineering requirements and capture not only linear relationships but also complex patterns from high dimensional data makes them powerful assets for time series modelling.

However, deep learning methods need large amounts of training data; they require more substantial computational power, such as graphics processing units (GPUs) for training, and are time-consuming. In contrast, feature extraction-based models are still popular because of their ease of use and robustness, making them particularly suitable for non-expert users. Besides, deep neural networks lack interpretability, which is especially problematic in the medical field. Several post-modelling interpretable models were proposed to analyse feature importance; however, these usually ignore sequential dependencies [129]. Alternatively, analysing the attention weights in models with strategically placed attention layers provides insights into the relative importance of features at each training time step.

Throughout this work, we tried to design data-driven pipelines for predicting different clinical outcomes. When the data sets at hand allowed, we addressed the problems using deep learning-based methods. At the same time, in other cases, we followed the more traditional feature extraction-based route. The next three chapters provide a complete overview of the contributions of different clinical applications.

CHAPTER 3

PREDICTING EMOTIONAL STATES

The work presented in this chapter was published in *JMIR Mhealth Uhealth* 2021;9(3): e24465. DOI: 10.2196/24465. Moreover, it was presented at the *CA2MH 2021* workshop, and during poster sessions at *EEML2021* and *MLSS2021*.

3.1 Introduction

3.1.1 Passively Sensed Behavioural Biomarkers

The subjective experience of mood is one of the most valuable sources of information about an individual's mental health [146]. Self-reported mood is a critical component of the mental status exam interview, which is to psychiatry what the physical exam is to other fields of medicine [70]. Furthermore, clinicians routinely ask questions about mood during clinical encounters. The presence of a specific mood state is a required criterion for many psychiatric diagnoses according to the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (e.g., depressed mood to diagnose a major depressive episode; elevated, expansive, or irritable mood for a manic episode). Mood is a predictor of psychiatric outcomes, and mood changes can be a harbinger of psychiatric decompensations. Therefore, accurate monitoring of mood states is a crucial component of mental health care. For example, both valences of mood states [147], and their variability [7] have been shown to predict important outcomes, such as several binge-eating episodes in bulimia nervosa [7] and treatment adherence in patients with bipolar disorder and opioid use disorders [63], [147].

Until recently, information about mood was only available to clinicians by directly questioning patients in person, either over the phone or via telepsychiatry video platforms. However, technological advances over the last few decades have allowed real-time monitoring of patients' self-reported mood states. Smartphone-delivered ecological momentary assessment (EMA), also known as experience sampling, "assesses individuals' current experiences, behaviours, and moods, as they occur in real-time and in their real-world settings" [27]. However, despite these technological advances, this form of mood state assessment relies on an individual's current level of insight, willingness, and ability to interact with the EMA platform. Many psychiatric disorders cause behavioural changes that may decrease an individual's likelihood of interacting with an EMA tool (demotivation, apathy, and survey fatigue), causing missing data, not at random. Therefore, identifying objective behavioural biomarkers of mood states that can be passively sensed without patient participation is a research priority.

Through patients' mobile phones and other wearable devices, continuous sensor data can be collected in a non-invasive manner, providing valuable information about everyday activity patterns. The possibility of inferring emotional states by analysing smartphone use data [6], [105], [123], GPS traces of movement [29], [124], social media data [43], and even sound recordings [112], [115] has become growing research focus over the past decade. Such approaches can analyse individuals' emotional patterns, enabling better self-management of one's activity and behavioural choices. Moreover, for patients with mental illnesses and their caregivers and health care providers, these models could provide a means to predict mental health crises and maladaptive behavioural patterns and allow for early intervention.

3.1.2 Related Work

In recent years, numerous studies have demonstrated the potential of exploiting mobile sensing data to infer users' emotional states and well-being. In an older study, LiKamWa et al. [105] developed MoodScope, a statistical inference model for predicting the users' daily mood average based on the circumplex mood model [166], [167], from communication history and app use patterns. They collected data from 32 participants over two months and reported an initial accuracy of 66%, which improved over time for personalised models.

Jaques et al. [86] conducted a study using physiological signals, location, smartphone logs, and survey responses from 206 college students over a month to model students' happiness. They applied classical machine learning methods, such as support vector machines (SVMs), random forests (RFs), neural networks, logistic regression (LR), k-nearest neighbour, naive Bayes, and Adaboost to perform the classification task and reported 70% accuracy. Another study focusing on predicting college students' stress and mental health status was conducted by Sano et al. [169]. They compared lasso regression and SVM with linear and radial basis function kernels for two classification tasks: low or high stress and low or high mental health categories. They reported over 70% accuracy and showed a significant performance increase when data from wearable sensors (such as skin conductance and temperature) were used, compared with behavioural data derived from phone sensing.

Umematsu et al. [192] compared non-temporal (SVM and LR) and temporal (long short-term memory [LSTM]) machine learning methods to forecast the stress level of the upcoming day using a predefined number of days of previous data (physiological signals, mobile phone use, location, and behavioural surveys). A more recent study by Morshed et al. [131], using the StudentLife [198] and Tesserae [121] data sets, demonstrated that mood instabilities (computed from the mapping of moods on the photographic affect meter scale [152] to arousal and valence values) are predictable from features derived from passive sensor measurements.

In a large-scale study conducted by Servia-Rodríguez et al. [176], the researchers used passive sensing data and self-reported moods collected for about three years from 18,000 users to build a predictive model for users' moods. They trained a deep neural network of stacked restricted Boltzmann machines for a 2-class classification problem (positive and negative mood). They reported above 60% prediction accuracy for weekdays and 70% for weekends. An LSTM recurrent neural network (RNN)–based analysis, performed by Suhara et al. [183], showed that applying a temporal model for forecasting severe depressive states outperformed non-temporal models. Their study relied on a large-scale longitudinal data set of self-reported information about mood, activity, and sleep of 2382 self-declared depressed people over 22 months.

Cho et al. [34] conducted a prospective observational cohort study to evaluate the mood of 55 patients with major depressive disorder and bipolar disorder types 1 and 2. They collected light exposure data passively via mobile phones of patients and self-reported daily mood scores. Activity, sleep, and heart rate data were collected using activity trackers. This information was then processed into 130 features based on circadian rhythms, and mood prediction was performed using the RF method. Their approach showed good sensitivity and specificity for mood state and episode prediction.

Taylor et al. [185] focused on building personalised models for forecasting the next day's mood (good or bad), health (fair or poor), and stress intensity (low or high). The multitask learning-based approach used data about the physiology and behaviour of 206 undergraduate students and the weather of the current day, collected for 30 days. Their results showed that tomorrow's well-being could be predicted with 78% to 82% accuracy using a personalised model based on the present day's data. Busk et al. [28] proposed a hierarchical Bayesian approach for forecasting mood for up to 7 days from smartphone self-assessments of 84 patients diagnosed with bipolar disorder. Their best-performing model used a history of 4 days of self-assessment, indicating that short-term historical mood is a significant predictor.

Another recent observational study by Darvariu et al. [41] combined user-reported emotional information, passive sensing data, and visual context information from individuals' surroundings in the form of images to develop deep learning techniques for emotional state inference. Their findings showed context-dependent associations between self-reported emotional states and the objects surrounding the individuals.

These studies provide insight into the potential of using mobile sensor data to infer individuals' mental well-being. However, none of these studies reported working with a data set consisting of observations from a non-experimental setting or dealing with large amounts of missing data. Moreover, in most of these studies, the problem they are trying to solve is a 2-class classification problem. Here, the problem is approached from a more refined perspective (i.e., predicting emotional state in valence and arousal dimensions).

3.1.3 Objectives

This study applies machine learning algorithms to predict mood states based on passively sensed behavioural patterns. Specifically, we aim to assess which behavioural features provide the most important information about daily emotional valence. The study was conducted using data collected via a clinically validated eHealth platform (*eB2 MindCare*) [21], [30]. This app is designed to run unobtrusively in the background of an individual's smartphone, and it automatically and continuously gathers information about behaviour via both the individual's smartphone and wearable devices. It also provides an electronic diary-type interface for users to register information about their emotions and other important events.

3.2 Methods

3.2.1 Data

Daily summary values of 6 passively collected observations were considered: step count, distance travelled, hours of sleep, phone use, time spent at home, and the number of locations visited. An additional binary variable indicated whether the patient practised sports during the day.

A subset of 943 users (patients and non-pathological subjects) was selected with at least 30 days of passively sensed data in the eB2 database between January 2019 and March 2020. The number of recorded days per patient varied from 30 to 487, with a mean of 190 (SD 122). Demographic information was available only for 871 users. All the users were Spaniards. Of these, 63.5% (553/871) were female, and 25.1% (219/871) were male, and gender information was not available for the remaining 11.4% (99/871). All age groups were adequately represented in the data set, with a mean age of 41 years (range 18-77 years) computed at the beginning of the measurement period. The patient population came from 2 main categories: 61.3% (534/871) were outpatients from external psychiatric consultancy, and 22.1% (192/871) were suicidal high-risk outpatients. The remaining 16.6% (145/871) users were non-pathological. Note that neither demographic nor diagnostic information was used in the rest of the study.

A well-known framework for dealing with emotional experience characterises emotions in a 2-dimensional space defined by Russel [166], [167]. The arousal and valence are combined, with valences ranging from highly negative to highly positive and arousal ranging from low to high. Daily emotional valence and arousal metrics were determined using raw emotion data entered by patients. Valence was then computed as the sign of the difference between positive and negative emotion counts, whereas arousal was determined based on the categories in the study by Scherer [172].

The left subfigure in Figure 3.1 shows the projection of emotions to the arousal-

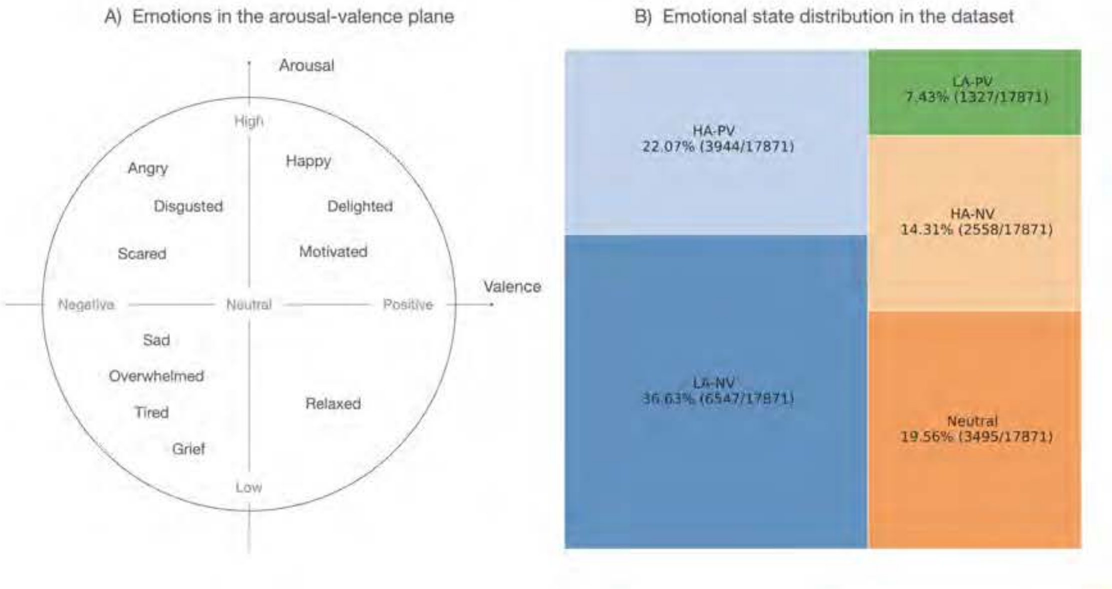


Figure 3.1: Projection of emotions into the arousal-valence plane and their distribution in the data set. HA-NV: high arousal-negative valence; HA-PV: high arousal-positive valence; LA-NV: low arousal-negative valence; LA-PV: low arousal-positive valence.

valence plane. The emotions listed on the graph are those patients can register via the eB2 app. As the right subfigure in Figure 3.1 shows, there is a significant imbalance between the different emotional labels. The majority corresponded to negative emotional valence (9105 entries), followed by positive emotions (5271 entries) and only 3495 neutral entries in the entire data set. Moreover, as emotions are self-reported, with users not being prompted to fill in this information, these entries are scarce compared with passively sensed behavioural data.

As data have been collected from several sources and received in different formats, the raw daily summary data have many anomalies, unwanted information, and noise. The presence of noise in the data can degrade the performance of machine learning methods. Therefore, it is essential to preprocess the data before using it as an input to any machine learning algorithm. The first step of preprocessing was removing any negative values, thresholding the time-related variables to 24 hours, the step count to 30,000 steps per day, and the distance to 500 km. Data were then standardised over all patient sequences, making each input feature 0 mean (SD 1).

Moreover, the data set contained a large percentage of missing observations (Figure 3.2): approximately 84% of the observations were partial, a bit over 5% was complete, and the remaining 10% were entirely missing. Slightly less than 10% of the observations were labelled by an emotion entry. A total of 271 patient sequences were observed for all seven summary variables. Close to half of them did not have information about the time spent at home and the number of locations visited. The app use information was also wholly missing for 226 patients. In addition, 114 patients had more than 30 consecutive

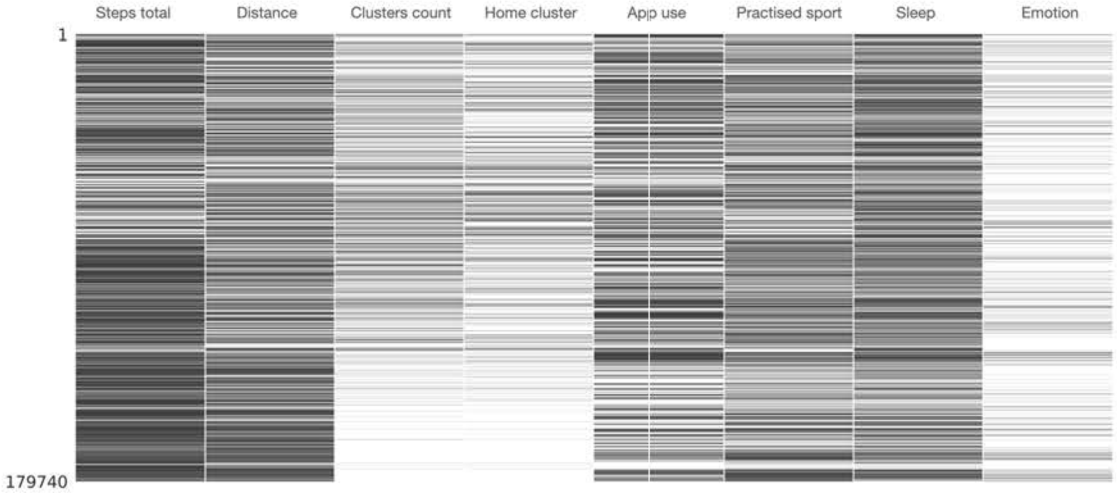


Figure 3.2: The distribution of missing data in the selected observation sequences. Black indicates the presence of observations, and white the lack of them.

days of completely missed observations (range 31-372).

3.2.2 Probabilistic Generative Models for Dealing With Missing Data

Imputing missing data using statistical measures such as the mean, median, or even interpolation fails when the percentage of missing data is very high. These approaches can reduce variability in the data set and introduce bias. However, probabilistic generative models can learn the underlying distributions in a data set by adjusting the model parameters to best account for the data in the sense of maximising the evidence, even in the presence of missing data. Mixture models (MMs) [22] and hidden Markov models (HMMs) [156] are frequently used types of such models.

MMs comprise a finite or infinite number of components, possibly different distributional types, that can describe different data features. The data can then be modelled as a mixture of several components, where each component has a simple parametric form (such as a Gaussian). The model is formulated in terms of latent variables, which represent the component each data point was sampled from and learned from the observed features, referred to as observables, by adjusting the model parameters, which define the observable emission probabilities, such that the MM best accounts for the data in the sense of maximising the evidence.

HMMs are temporal MMs that are commonly used for time-series analysis. These are generative models characterised by a set of observable sequences. The discrete states of the HMM are assumed to have been generated by a first-order Markov chain process, and each observation depends only on the paired state. An HMM comprises an initial state probability distribution, a state transition probability distribution, and a symbol emission probability distribution. Both MMs and HMMs were trained using the expectation-maximisation algorithm.

In this study, the observed data were heterogeneous. Practice sport and emotional state are categorical, and the rest of the variables are assumed to be real-valued. Both MMs and HMMs can deal with missing data without requiring imputation before training via marginalisation. For the Gaussian parameters, the diagonal covariance matrices were considered. Furthermore, both generative models were trained semi-supervised for emotional valence and arousal-valence discrete observations. Namely, the different emotional states' emission probabilities were fixed for some components, whereas others were adjusted during training, such as the other model parameters. For instance, in a 5-component MM with binary label emissions, the emission probability for label 0 of the three components can be set to 1, forcing the components constantly to emit label 0. In contrast, the other two components can always be forced to emit label 1.

3.2.3 Emotion Prediction Models

A series of experiments were conducted for emotional status prediction using both non-temporal and temporal machine learning models. The underlying motivation was to analyse whether there were long-term dependencies in the data concerning patients' daily emotional states.

Probabilistic generative models (MM and HMM) were used to perform the imputation. Note that only the input features were imputed, and the emotion labels were not. When using MMs, first, for each observation, the posterior distribution needs to be inferred to find which component the observation is most likely to belong to; then, the missing attributes are imputed by a sample generated from that component. Information about the emotional state of the current observation was not included in the posterior computation (otherwise, the model would overfit). When using HMMs, all observation sequences were first decoded using the Viterbi algorithm on the trained HMM. This method finds the most likely sequence of components that could have resulted in the given observation sequence. Once the state sequence was determined, the missing data were imputed by the samples generated from the corresponding states for each time step. The state posterior probabilities were computed by applying the forward algorithm [156], leaving out the current emotional observation.

For non-temporal machine learning methods, LR, support vector classifier (SVC), random forest classifier (RFC), and multilayer perceptron (MLP) were considered. These models allow comparison with previous emotional state studies [34], [86], [169], [192]. A grid search was performed for each case for hyper-parameter tuning.

RNNs [165] have recurring inputs to the hidden layer; this allows them to remember input states from previous time steps, which can carry important information for future time-step predictions. There are three common types of RNNs: vanilla RNN, LSTM [81], and gated recurrent units (GRUs) [35]. Vanilla RNNs have short-term memory. If the observation sequence is rather long, these models have difficulty remembering relevant information from earlier time steps. LSTM and GRU cells, which contained gates that regulate the information flow, were designed to solve this problem.

In this experiment, RNNs of each of the three types were tested. A single layer with 64 hidden units was used, whose output was connected to a dense layer. Finally, the softmax activation function provides the predictions. The model was trained using the Adam method and the negative-log-likelihood loss for 50 epochs, using early stopping. One-layer RNNs with vanilla RNN, LSTM, and GRU cells were trained using 64 hidden units for each case. More complex models have also been tried, such as dilated RNN, multilayer RNN, and temporal convolutional networks. However, they did not improve performance, proving that simpler RNNs could explain the data's temporal correlations.

3.2.4 Personalised Models

Hierarchical Bayesian regression models were proposed to improve the above models by accounting for individual differences to predict the emotional state of patients. This technique is proper because it includes information from the population via partial pooling of the data. The proposed model allows intercepts to vary across patients according to a random effect while having a fixed slope for the predictor (i.e., all patients will have the same slope).

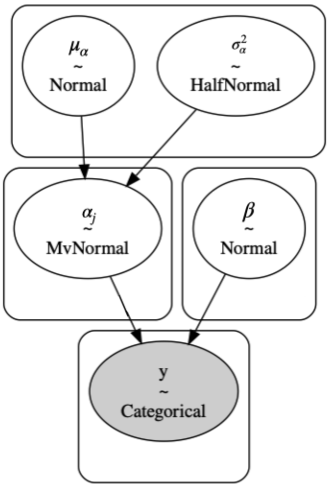


Figure 3.3: The proposed Bayesian network of the hierarchical linear regression model.

In our model, showed in Figure 3.3, for individual j, observation i, target variable y_{ji}, and input features x_{ji}:

$$y_{ji} = Categorical(\alpha_j + \beta x_{ij})$$

where the random intercept effect is drawn from the population distribution:

$$\alpha_j \sim Normal(\mu_\alpha, \sigma^2_\alpha)$$

Moreover, the population mean and SD are independent normal and half-normal priors. By setting a separate bias term for each patient, rather than fitting separate regression models for each patient, multilevel modelling shares strength among patients, allowing for a more reasonable inference in patients with little data. The models were trained with Stein Variational Gradient Descent [109], [110] for 50 epochs using the Adam optimiser.

3.2.5 Evaluation and Interpretability

Accuracy, area under the receiver operating characteristics curve (AUC-ROC), and area under the precision-recall curve (AUC-PRC) were used as the evaluation metrics. AUC-ROC is commonly used for balanced and imbalanced classification problems because it is not biased toward the majority or minority class. However, AUC-PRC scores provide more insight into the minority class when the problem is skewed. As the AUC-ROC and AUC-PRC scores are computed for binary classification problems, different types of averaging can be performed on the data in the case of multiclass targets. The reported results were micro-averaged, meaning that the metrics are global, computed by counting the total number of true positives, false negatives, and false positives.

Based on several model interpretability methods, Lundberg and Lee [114] defined the Shapley Additive Explanations (SHAP) value, a modality to explain any machine learning model's output. The SHAP values can provide global interpretability to the machine learning models by showing how much each feature contributes, positively or negatively, to the target variable. This study used this approach to analyse the model's feature importance. Moreover, this method can be applied to analyse the decisions for individual predictions, which provides better insights into the relationships between passively collected mobile data and self-reported emotions.

3.2.6 Experiments

For MM and HMM training, only those patient sequences with at least partial observations for each of the seven features and emotions were used. Moreover, the maximum sequence length was limited to 365 days, and sequences with more than 30 days of consecutive missing data for all variables were discarded. After this elimination process, 233 sequences were used to train both the MMs and HMMs with different numbers of states. These patient sequences were excluded from the training and test sets of the later models. For the global models, the data set containing the remaining 710 patient sequences were divided into training and test sets using 80% of the sequences for training and 20% for testing. These data sets were kept independent. The train-test split cannot be done for the personalised models by randomly selecting a given percentage of the patients for training while leaving the others for testing, but all 710 patients must be included. Therefore, the patient sequences themselves were split into training and test sections. The first 80% of the labelled observations, in chronological order, were used for training, and the remaining samples were used for testing.

As the LR, SVC, RFC, and MLP cannot directly exploit time-series data, we created the following 2 cases as inputs for these models. First, the input-output pairs consisted of 1 day of labelled observation. Second, three days a week before the entered emotion was considered and concatenated into a single feature vector. In the case of the temporal models, training was performed with 30-day, 3-month, and 6-month long sequences. Before creating the above feature vectors, the missing data in each patient sequence were imputed by the MM or HMM samples. For models trained with mini-batch stochastic gradient descent, every data point is imputed every time it enters the optimiser. The sequences were decoded multiple times, and missing data were imputed by samples generated from the corresponding state.

We designed two types of experiments. The first type is limited to the projection of the recorded emotions to a single axis of the arousal-valence plane, and the second set of experiments considered 2-dimensional projections. A total of 3 different settings were analysed for the classifiers' input features:

- using the imputed raw data
- using the MM or HMM posterior probabilities instead of the raw input features
- using the raw inputs concatenated with the MM or HMM posterior probabilities

3.3 Results

3.3.1 Generative Models

After experimenting with several hidden state setups, seven hidden components captured the data's underlying patterns well, leading to the best results when a classifier was applied to the data later to predict the emotions and provide interpretable states. In this case, the emission probabilities of the five states were fixed such that two pairs of states always emit negative and positive emotions, and one always emits a neutral emotion. The different components turned out to be specialised, capturing contrasting behaviours, as Figure 3.5 shows. In terms of features, "steps total" refers to step count, "distance" refers to the distance travelled, "sleep" refers to the hours of sleep, "app use" refers to the hours spent using different apps, and "home cluster" refers to the time spent at home, "clusters count" refers to the number of visited locations, and "practised sport" is an indicator of whether the patient practised any sports. Of note, the negative mean values were a result of the normalisation of the features.

Focusing on the three components that mainly emit negative emotional valence (1, 2, and 3), it can be seen that the corresponding modelled behaviours are contrasting. Component 1 represents days when the patients are quite active, visit multiple locations, spend a significant amount of time using their phones, and sleep very few hours. Component 2 is characterised by fewer steps and low app use. Component 3, however, captures days with low activity and is mostly spent at home. The corresponding sport-related discrete emissions show that the patients practice some sport (>15 minutes of walking, biking, running, other, or a combination of those) in components 1 and 2, but less likely in component 3. Components 0, 5, and 6 correspond to positive emotional valence. They also seem to capture significantly different behavioural patterns.

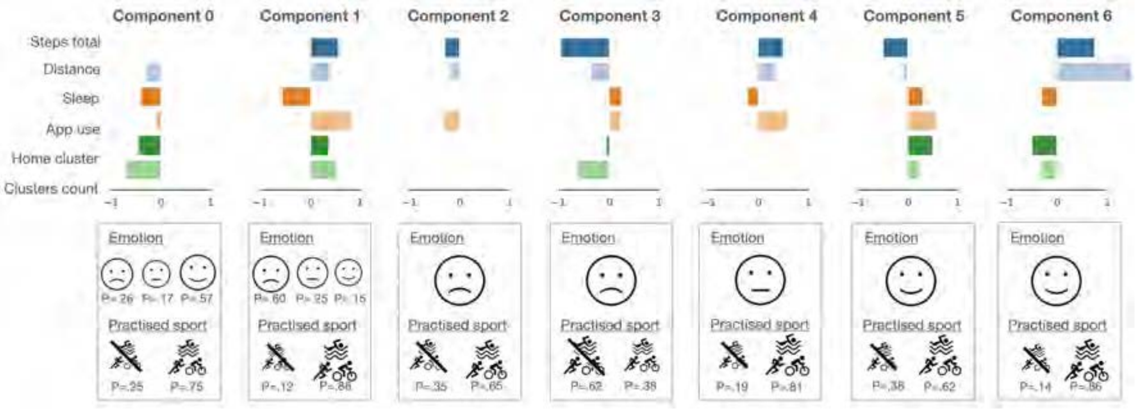


Figure 3.4: The 7-component mixture model structure was used for emotional valence modelling with each Gaussian mean in each component and indicating discrete emission probabilities. The size of the icons indicates the magnitude of the discrete emission probabilities (emotion and sport).

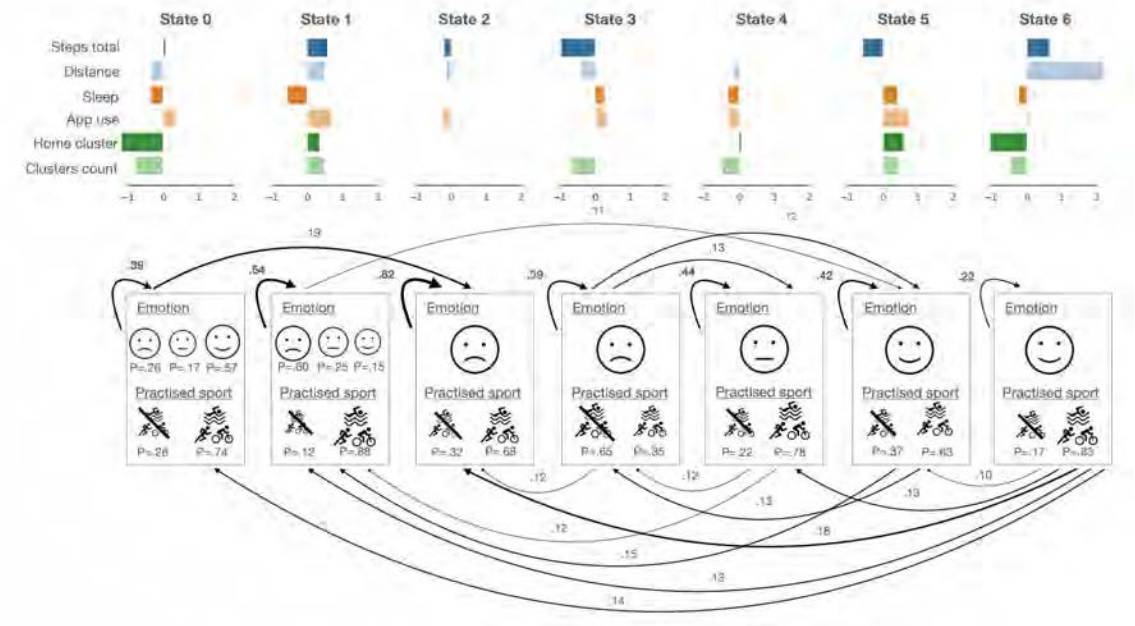


Figure 3.5: The 7-state hidden Markov model structure was used for emotional valence modelling with each Gaussian mean in each component and indicating discrete emission probabilities. The size of the icons indicates the magnitude of the discrete emission probabilities (emotion and sport). Only the transitions with a higher than 0.1 probability are shown in the graph.

In component 0, the patients seemed to sleep less and did not spend much time at home; component 5 captured days with more time spent at home and excessive phone use. Component 6 captures the days of travel. Finally, the component capturing neutral emotions indicates days with medium activity and more app use.

Including the temporal properties of HMMs, the trained generative model with seven hidden states and the same fixed emotional state emissions led to very similar interpretable outcomes as the MM (Figure 3.5). The temporal characteristics were not very strong. States 2 (with fixed negative emotional valence emission) and 1 (with mainly negative emotional valence emission) had the highest self-transition probabilities. If the self-transition probabilities are large, it indicates a stable state. States 0, 3, 4, and 5 have somewhat large self-transition probabilities, suggesting that days with positive and negative but neutral emotions following each other are common in the patient population.

In the arousal-valence case, the 7-state generative models had one state assigned to all the emotional state emissions, and the other two were trained with the rest of the parameters. Similarly, as before, the states appear to capture specific behaviours, such as days of medium activity but mostly spent at home, more active days, days with more travel, and so on (Figures 3.6 and 3.7 provide the sketches of the 7-component MM and HMM, respectively).

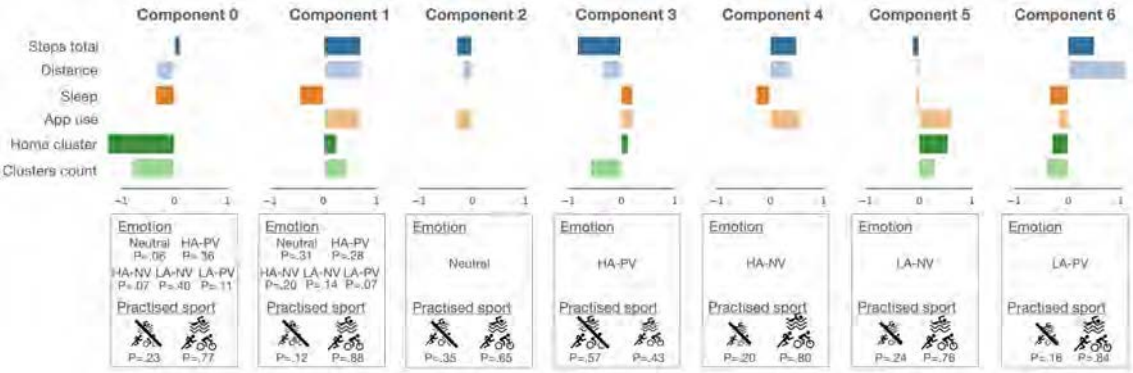


Figure 3.6: The 7-component mixture model structure was used for emotional arousal-valence modelling with each Gaussian mean in each component and indicating discrete emission probabilities. The size of the icons indicates the magnitude of the discrete emission probabilities (emotion and sport).

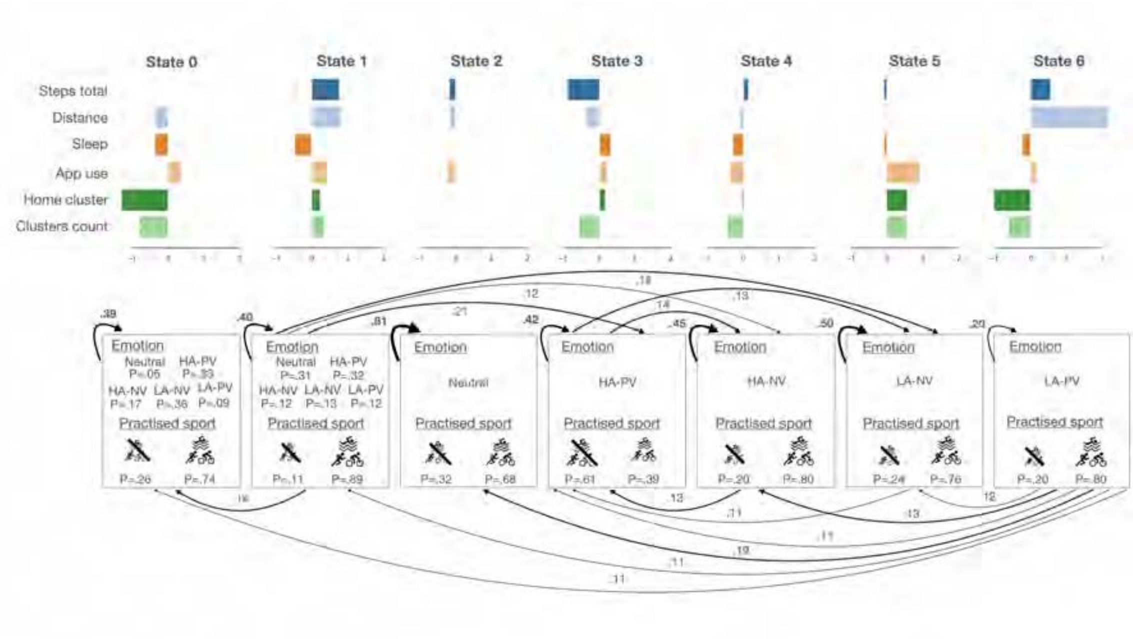


Figure 3.7: The 7-state hidden Markov model structure was used for emotional arousal-valence modelling with each Gaussian mean in each component and indicating discrete emission probabilities. The size of the icons indicates the magnitude of the discrete emission probabilities (emotion and sport). Only the transitions with a higher than 0.1 probability are shown in the graph.

3.3.2 Predicting Emotional Valence

Figure 3.8 compares the accuracy and the micro-average AUC-ROC and AUC-PRC scores for the trained classifiers in the three experimental setups, as described in the Experiments section[1]. Most classifiers achieved significantly higher performance than random guessing (AUC-ROC=0.5). As the figure shows, the models perform the worst on the raw data. Using the HMM or MM posteriors as input features or combining the raw data with the posteriors increases the performance.

Table 3.1 compares the best-performing models using the MM and HMM posteriors. The difference in the results obtained with the MM posteriors and HMM posteriors is minimal. This indicates that the temporal dimension is not very relevant to the problem at hand; hence, a simpler generative model is sufficient for the problem.

The best performing model was the MLP with the posteriors of 7 days of observations as input features. Concatenating the posterior probabilities for three days or seven days of observations significantly improves the performance; however, training RNNs with longer

[1] Model notations: LR/SVC/RFC/MLP - x = LR/SVC/RFC/MLP classifiers trained with input features formed of x-days of observations concatenated to create a single feature vector. RNN/LSTM/GRU - x = RNN/LSTM/GRU - RNNs with different cells using x-months-long input sequences. Input feature notations: w/o posteriors = raw features used as classifier input; only posteriors = the MM component posterior probabilities used as classifier input features; w/ posteriors = raw features concatenated with the MM component posterior probabilities used as classifier input features. Model abbreviations: LR = logistic regression, SVC = support vector classifier, RFC = random forest classifier, MLP = multilayer perceptron, RNN = recurrent neural network, LSTM = long short term memory, GRU = gated recurrent unit.

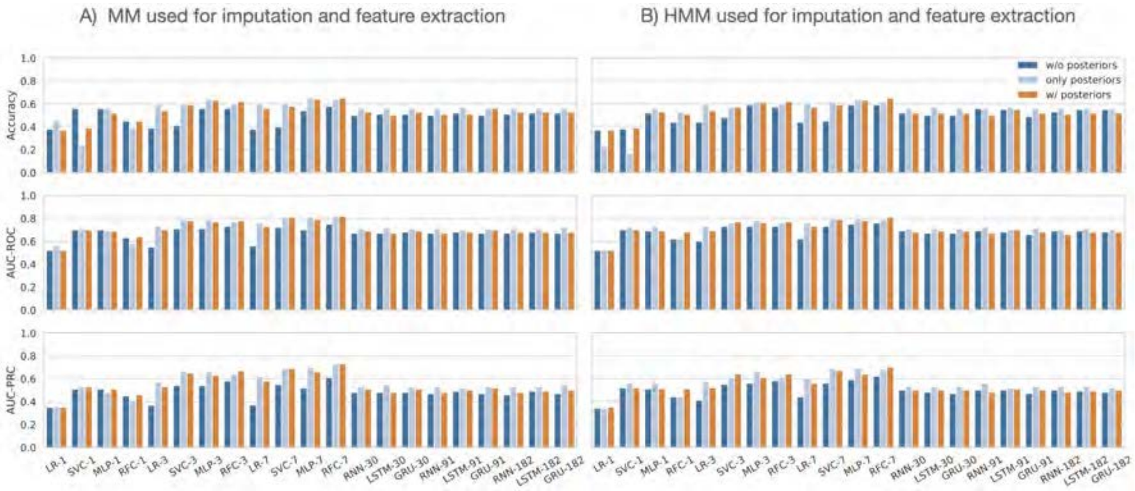


Figure 3.8: Classifier performance comparison plot - Emotional valence case.

Model and classifier input features	Accuracy (%)	AUC-ROC	AUC-PRC
Multilayer perceptron using 7 days of observations as input features			
Mixture model posteriors	65	0.81	0.70
Hidden Markov model posteriors	64	0.80	0.69

Table 3.1: Performance comparison of the best performing models using mixture model and hidden Markov model posteriors as classifier input features.

observation sequences leads to decreased performance. This suggests no substantial seasonality or long-term trend of the self-reported emotions; thus, time-series models are not needed for the emotional state prediction task.

Generally, the most misclassified emotional state is the neutral state (refer to Table A1 in Appendix A for confusion matrices). In most cases, it is confused with a negative emotional state and reasonably often with a positive one. There is some confusion between positive and negative emotional states, but somewhat fewer for negative emotions. This suggests that the models are more sensitive to detecting negative emotions, which can be desirable; for example, if the app's goal is to detect periods when the patient is feeling down.

3.3.3 Predicting Emotional Arousal-Valence

In the second experiment, the target variables were the emotion projections into the 2-dimensional arousal-valence space, based on the categories in the study by Scherer [172]. Hence, the problem becomes a 5-class classification task. Here, we aimed to test the possibility of predicting daily emotions on a finer scale than the 2-class valence analysis presented above. Figure 3.9 provides a comparative overview of the model performance.

The best performance for the emotional arousal-valence prediction, with 48% accuracy (compared with the baseline of 20%), 0.77 AUC-ROC, and 0.50 AUC-PRC, was obtained by the RFC with seven days of data concatenated with the MM posteriors. The

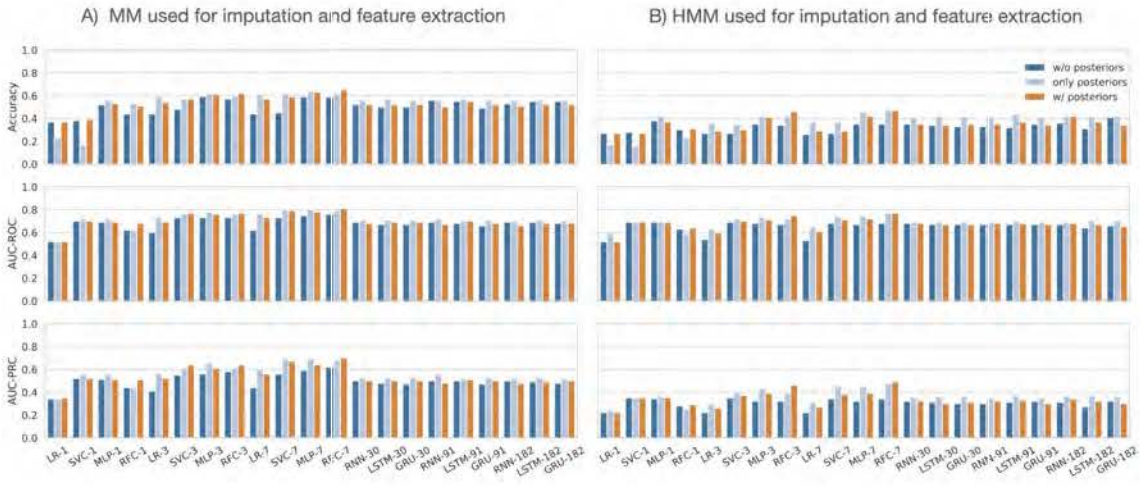


Figure 3.9: Classifier performance comparison plot - Emotional arousal-valence case.

GRU network trained on 30-day sequences reached results closest to those from the temporal models: 42% accuracy, 0.69 AUC-ROC, and 0.36 AUC-PRC. In this setting, the added MM posteriors' effect was more significant than the emotional valence prediction case. Using the posteriors as input features led to a 23% performance increase in some models. Table A2 in Appendix A provides a detailed performance comparison of the models.

Predicting more refined emotional states is a difficult task, as there are more classes to distinguish, and the class imbalance is also more accentuated. The trained models became somewhat biased toward the majority classes, causing the wrong classification of the minority classes (high arousal-positive valence and low arousal-positive valence). Generally, when the predictor variable is well separable, and there are no overlaps between the different classes, this separation can compensate for the imbalance; however, in this data set, that is not the case. Standard techniques to combat the imbalance problem, such as upsampling of minority classes, down-sampling of majority classes, and one-versus-rest training, were applied; however, these only improved slightly. Therefore, these results have not been reported.

3.3.4 Personalised Models

The previously presented models try to explain the variability of the observations by considering the patient population. As shown before, these models do not provide enough diversity when the classifier takes 1-day worth of data as input. Personalised models can provide a more scalable and accurate way to better representations for individual patients.

The posterior probabilities obtained from the MM components were used as input features for the personalised models because they proved to improve the prediction outputs of earlier experiments. In the global models presented previously, features representing one day of data led to insufficient classifier accuracy, especially in the LR models, which only reached a maximum of 43% for the 3-class problem and 16% for the 5-class problem. The proposed hierarchical Bayesian LR method led to a significant increase in performance, reaching 64% accuracy, 0.81 AUC-ROC, and 0.70 AUC-PRC for the 3-class problem and 52% accuracy, 0.82 AUC-ROC, and 0.55 AUC-PRC for the 5-class problem. This demonstrates that accounting for individual differences through a simple hierarchical model can substantially improve emotional state prediction performance without relying on previous days of data.

3.3.5 Feature Importance Analysis

Figure 3.10 provides an overview of which features are most important for the emotional valence MLP models using the raw data and using the raw data and MM posteriors as input features. In terms of features, "steps total" refers to step count, "distance" refers to the distance travelled, "sleep" refers to the hours of sleep, "app use" refers to the hours spent using different apps, "home cluster" refers to the time spent at home, "clusters count" refers to the number of visited locations, "practised sport" is an indicator of whether the patient practised any sports, and $P(s_i|x_t)$ refers to the posterior probability in component i. The mean SHAP values (Evaluation and Interpretability section) of every feature for every sample were computed to obtain an overview of which features are most important for the models. The plot below sorts features by the mean absolute value of the SHAP value magnitudes over all samples.

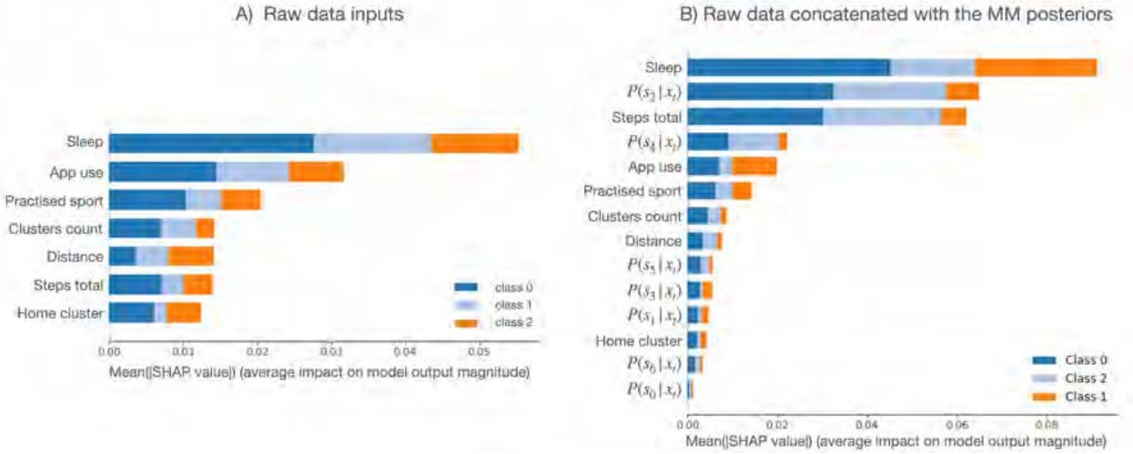


Figure 3.10: Summary plot of feature importance for the multilayer perceptron models for emotional valence prediction, showing raw data and raw data concatenated with mixture model posteriors. The following class labels were used: 0=negative; 1=neutral; and 2=positive emotional valence. MM: mixture model; SHAP: Shapley additive explanations.

The hours of sleep and the time spent using their phone (app use) influenced all classes' outcomes the most. The other features have an almost similar influence on the positive and negative classes. The negative output (class 0) is also strongly influenced by the step count, sports indicator, and time spent at home. If the posterior probabilities are combined with the raw features as inputs to the model, some outweigh the raw features in the decision-making process. For instance, the MLP relies heavily on the hours of sleep, the posterior probability of state 2, and the step count. The other classes seem more involved, requiring several posterior probabilities and raw values to form the prediction. The importance of posterior probabilities underlines the robust feature extraction provided by MM.

Similarly, the arousal-valence classifiers can be analysed. In the raw data case (Figure 5), although the model emphasises the hours of sleep and the step count, the other parameters become slightly less important. In the second case (Figure 3.11), some of the posterior probabilities seem to weigh more in the decision-making process than the raw features, as in the first experiment.

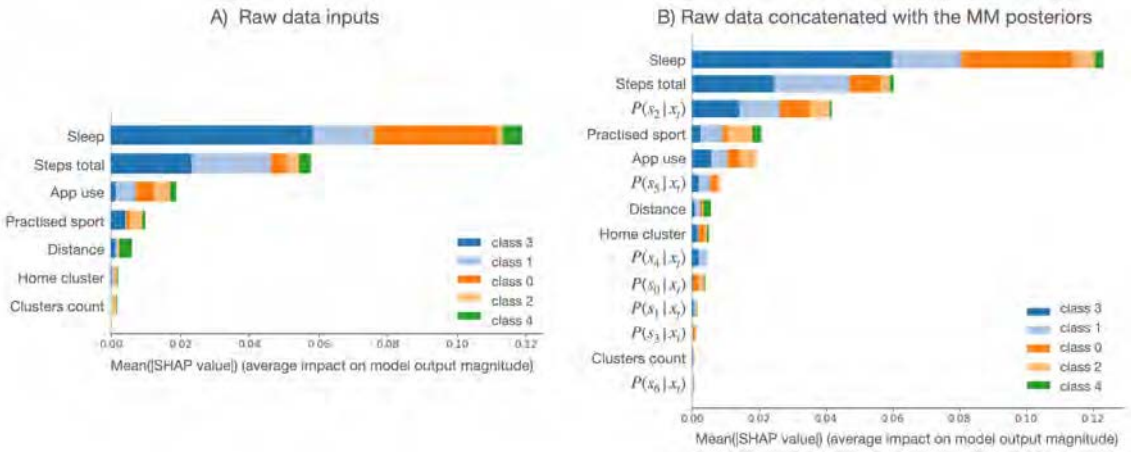


Figure 3.11: Summary plot of feature importance for the random forest models for emotional arousal-valence prediction, showing raw data and raw data concatenated with mixture model posteriors. The following class labels were used: 0=neutral; 1=high arousal-positive valence; 2=high arousal-negative valence; 3=low arousal-negative valence; and 4=low arousal-positive valence. MM: mixture model; SHAP: Shapley additive explanations.

3.4 Discussion

3.4.1 Principal Findings

A variety of different machine learning methods were used to analyse passively sensed behavioural data from 6 sources (step count, distance travelled, hours of sleep, hours of phone use, time spent at home, number of locations visited, and a binary variable indicating whether the patient practised sports during the day). These models were used to predict self-reported emotional state (valence or combination of valence and arousal) in a large, international sample of treatment-seeking patients with clinically significant psychological and emotional symptoms. Preliminary inspection of this data set revealed that the data exhibited significant missingness (approximately 84% of the observations were partial). This represents real-world clinical data sets, which usually contain many missing samples and are sparsely labelled. The fact that this kind of data is both noisy and often non-randomly missing means that developing robust imputation techniques is a nontrivial problem. However, developing such methods is vital if this information is used to support clinical decision-making.

We addressed this problem by training generative models to handle missing data. These models were then used for data imputation and latent state (feature) extraction for emotional state prediction. Predictive models performed significantly better when MM or HMM posterior probabilities were included alongside the raw behavioural input features. This suggests that the latent representation of the passively sensed behavioural variables discovered by the probabilistic generative models contains information relevant to daily emotional experience fluctuations. However, using HMMs over MMs did not improve the classification performance, implying that there are no strong temporal correlations in the daily observations that an HMM can capture. Furthermore, the nonlinear models outperformed the other static models in both experiments. RNNs did not improve daily emotion predictions, suggesting that long-term behaviour does not significantly influence patients' everyday emotional states.

When using raw data alone as input features, the hours of sleep had the most influence on the emotional state predictions. The importance of activity-related features varied between the two experimental setups. When posterior state probabilities were included in the model, some proved to be more important than the raw features. This indicates that the MM provided excellent feature representation and filtering of the observed behavioural signals. Interestingly, an inspection of the confusion matrices for the best performing models revealed that, for the valence prediction analysis, models were more sensitive to detecting negative, compared with positive or neutral emotional states. This is a useful feature as this is the domain of emotional experience most likely to be relevant for clinicians or self-monitoring trends in overall mental health.

Finally, we proposed a hierarchical Bayesian regression with varying intercepts and a common slope to personalise the models. This approach performs personalised predictions while accounting for population-level characteristics. The personalised models using 1-day long feature vectors achieved similar performance to the nonlinear variants using 3-day long feature vectors. Moreover, they performed significantly better than global linear LR models. Personalised models outperforming the generalised models are intuitively reasonable as the mood is very personal, and its perceptions among individuals differ.

3.4.2 Limitations

This study has some limitations. As previously mentioned, the data analysed here contain a large percentage of missing observations: approximately 84% of the observations are partial, only a bit over 5% are complete, and the remaining 10% are entirely missing. Some of the patient sequences had large chunks of consecutively missing observations, possibly because of sensor or software errors or the patients not using their devices for an extensive amount of time. Moreover, information about emotional states was sporadically reported. Therefore, only 10% of the behavioural data were labelled with respect to the outcome of interest.

Recording emotions is a subjective process; the regular reflection of the emotional state may influence how one answers. Most registered emotions were negatively valenced, meaning that the prediction models were somewhat biased toward negative emotional states. As a result, the models were most sensitive in the negative domain, and the overall prediction accuracies were not high in some cases. In addition, this study did not analyse mood variability, another important point in psychiatric disorders. However, it will be important to explore in the future to differentiate better whether it is a pathological mood state or a mood within the normal range.

3.4.3 Conclusions

This work is an initial step toward developing more robust and informed models for predicting emotional states from passively sensed data. It presents a sound basis for further exploration by proposing a solution to missing and sparsely labelled data, allowing the future focus to be directed toward developing more advanced models. Further research options include examining other deep learning models to improve prediction accuracy and analysing effects at a more refined time scale. Another intriguing question is to consider the effect of seasonality (weekdays and weekends, seasonal variation) on patients' emotional states. Moreover, the possibilities of specialised models for different patient groups or individual patients will be further investigated.

CHAPTER 4

PREDICTING CLINICAL ANXIETY FROM SOCIAL MEDIA APP USAGE DURING COVID-19 LOCKDOWN

Parts of the work described in this section were published in *JMIR Mental Health* 2021;8(9): e30833, DOI: 10.2196/30833, and it was presented in the format of a contributed talk at *IBC2022*.

4.1 Introduction

During the early peaks of casualties from the first wave of the COVID-19 pandemic, governmental lockdown measures generated anxiety, and physical isolation among a large portion of the global population [95]. The mental health consequences of these lockdown measures are only beginning to be understood on a larger scale as more population data and clinical outcomes are becoming available. Quarantine and lockdown measures have been linked with short- and long-term adverse psychiatric consequences such as suicidal ideation, depression, and post-traumatic stress disorder (PTSD) in current and previous outbreaks [74], [78], [160]. Physical isolation can also increase the intensity and perception of threat, especially when its uncertain nature is explicit such as a high-mortality novel virus outbreak [61], [203]. Anxiety can cause maladaptive coping behaviour such as substance use, which can, in turn, lead to adverse mental health outcomes in a negative feedback loop [159]. It can also compromise effective health-related social decision-making, as seen in panic buying, hoarding, and excessive online information search during the COVID-19 pandemic [88], [162].

On the other hand, positive public health outcomes are driven by individuals' sound health decisions based on accurate perceptions of the costs and benefits to self and society [206]. Therefore, identifying the severity of short-term anxiety symptoms in the population exposed to lockdown measures is a significant public health agenda, and it may lead to early detection of those at risk for psychiatric sequelae.

When in-person communication is diminished, individuals experiencing anxiety symptoms may turn to the digital world to connect with others [18]. In recent years, passive smartphone sensor data have been utilised in empirical studies to identify various psychiatric presentations and mental health-related behaviours [64], including social anxiety severity, through rich real-time analysis of users' functioning in the digital world within their natural environment [84].

There are conflicting perspectives concerning the role of social media use in developing anxiety during crises. Although much literature on social media has been produced regarding their adverse effects on mood and mental health, the interplay network among social media use, user characteristics, and the manifestation of anxiety is likely more complex and layered [98], [196]. Excessive time searching for news on social media has been linked with higher anxiety during COVID-19, and Ebola outbreaks [58], [62], [136]. Conversely, social media exposure to public health information during the MERS outbreak was positively related to forming appropriate risk perceptions in the population, moderated by users' information processing style and self-efficacy traits [36]. A recent report suggested that increased social media usage predicts increased physical activity, possibly promoting healthy behaviour during COVID-19 [138]. In other words, finding models to describe how users engage in social media, as opposed to whether or not they are using it, appears to be highly relevant to clinical and public health.

4.2 Methods

4.2.1 Data

From February 1 through May 3, 2020, passive smartphone usage data were collected using *eB2 Mindcare* [23], [30], [49], a clinically validated eHealth platform. On March 14, a country-wide state of emergency was declared due to rising mortality rates from the coronavirus pandemic, and the government-mandated a lockdown of all individuals who were not essential workers (i.e., they were restricted to their residences, except when purchasing food and medicines or attending emergencies). On May 4, Madrid entered the first step in de-escalating the lockdown, which allowed the reopening of small businesses and walking outside within set time slots [181].

Daily time (in seconds) automatically logged on communication apps and social networking apps were extracted and analysed during the pre-lockdown (i.e., February 1 through March 13) and the lockdown periods (i.e., March 14 through May 3). Social media app categories—communication and social networking—were based on the labels designated in the Google App store. Communication apps included messaging, chat/IM, dialer, and browser apps such as WhatsApp, Telegram, Facebook Messenger, and Gmail; social networking apps were primarily those for sites such as Instagram, Twitter, and TikTok.

A clinical psychologist collected short-term mental health outcomes, including self-reported intensity of psycho-social stressors during the lockdown and Generalised Anxiety Disorder 7-item scale (GAD-7), by phone follow-up between May 12 and June 3 after the initial lockdown measures had been lifted [137]. *Clinical anxiety* was defined as a GAD-7 score of 10 or greater, given its diagnostic value in screening for severe GAD, panic disorder, and social phobia [97]. COVID-19 exposures, risk perception, and social

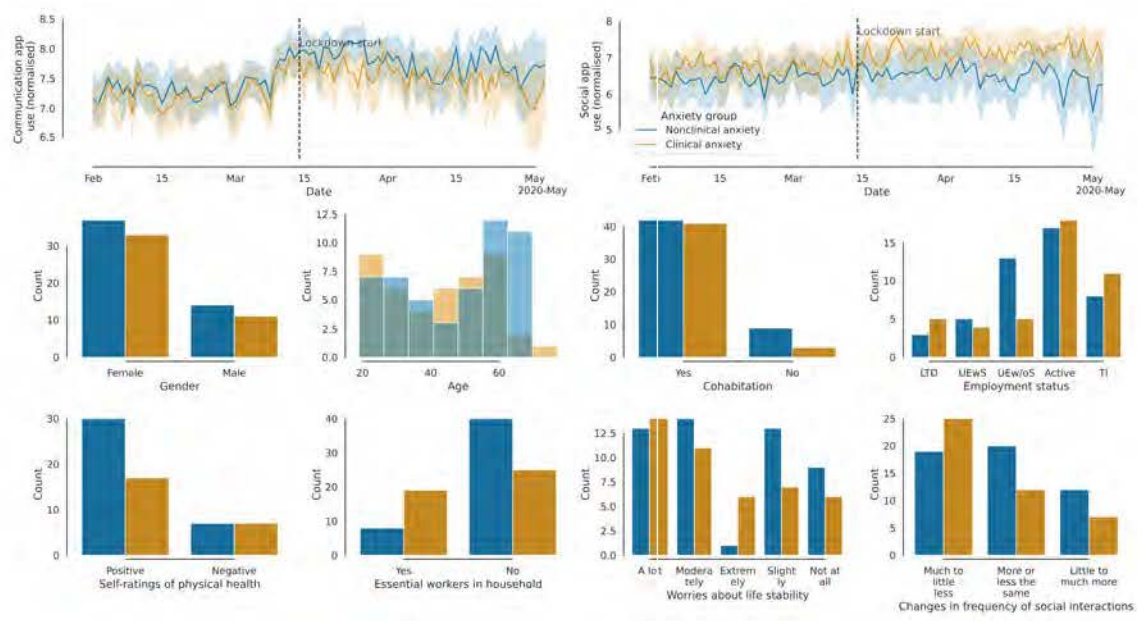


Figure 4.1: Distribution of temporal (mean and 95% confidence interval) and static variables in the data set consisting of the patients (n=95) are grouped by anxiety. Usage data collected in seconds were logarithmically transformed and scaled. Abbreviations: LTD = Long-term disability, UEwS = Unemployed with subsidy, UEwoS = Unemployed without subsidy, TI = Temporarily incapacitated.

behaviours during the lockdown period were also assessed during the phone call.

Figure 4.1 provides an overview of the data distribution in the studied population. A pronounced increase in time logged on communication apps in the nonclinical anxiety group (GAD-7<10) versus clinical anxiety group (GAD-7>10) after March 14, and increased overall time logged on social network apps in the clinical anxiety group.

4.2.2 Machine Learning Pipeline

We designed a 2-step approach that combined a probabilistic generative model, namely a hidden Markov model (HMM) [156] for temporal data processing and aggregation, with logistic regression to predict the binary outcome (clinical anxiety versus nonclinical anxiety) by dichotomised GAD-7 (Figure 4.2).

Nonclinical anxiety outcome (n=51) was encoded as the negative label, and clinical anxiety outcome (n=44) was encoded as the positive label. The class imbalance problem was insignificant. Continuous longitudinal daily communication and social networking app usage in seconds were chosen as independent variables, with anxiety-associated clinical variables as additional predictors.

HMMs are commonly used for time-series analysis. HMMs model generative sequences, which are characterised by a set of observable sequences. A first-order Markov chain process generates the states of the HMM. The following components specify an HMM: $S = s_1, s_2 ... s_N$, a set of N states; $A = a_{11} ... a_{NN}$, a transition probability matrix;

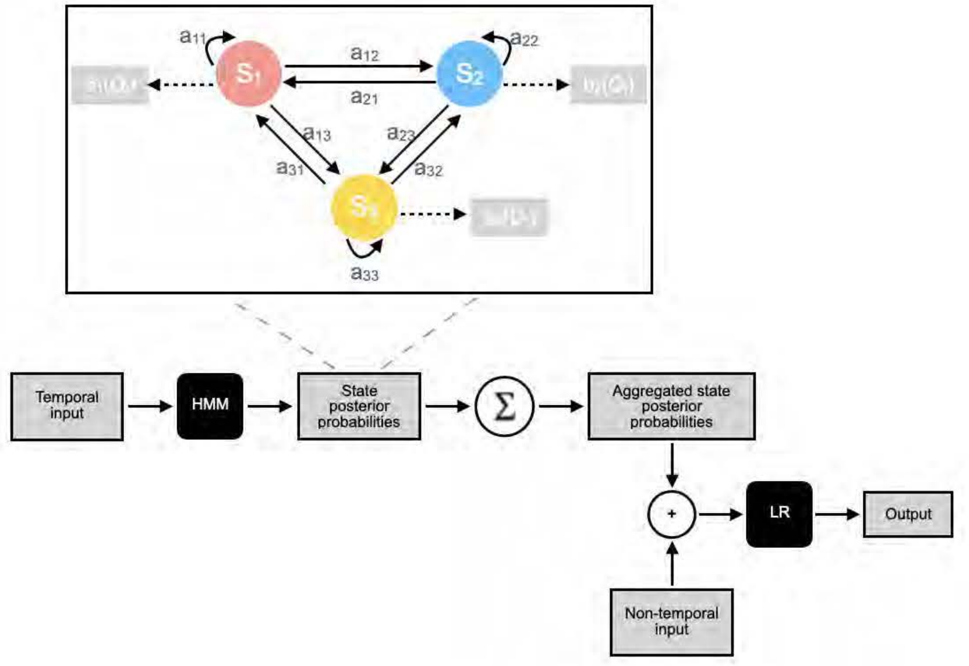


Figure 4.2: The proposed hidden Markov model-based anxiety prediction pipeline. Notations: HMM - hidden Markov model, LR - logistic regression, O_t - observation at time point t, S_i - state i, P_{ij} - transition probability from state i to state j, $f_i(O_t)$ - emission probability of O_t from state i, where $i, j \in \{1, 2, 3\}$.

$O = o_1, o_2...o_T$, a sequence of T observations; $B = b_i(o_t)$, emission probabilities, expressing the probability of an observation o_t being generated from state i; and $\pi = \pi_1, ...\pi_N$, an initial probability distribution over states.

The state-space of the applied HMM is discrete, while the observations can be discrete or continuous. This study treats communication and social networking app usage as continuous variables from a Gaussian distribution. The parameters of an HMM can be trained with the Baum-Welch algorithm, a variation of the expectation-maximisation algorithm. The model can deal with missing data using marginalisation without requiring imputation before training. To select the optimal number of hidden states, we computed the Akaike information criterion and the Bayesian information criterion [151] after training HMMs with 2-19 states.

Once the optimal HMM state sequence was selected for each temporal sequence, we computed the state posterior probabilities $P(s_i = k|x)$ (the probability of being in state k at position i of the sequence x) for each time point and aggregated them by summing over time for each patient. This feature vector of length N was then concatenated with non-temporal clinical features of length $N_{clinical}$ to form the feature vector of length $N + N_{clinical}$. Hence the data set of size for the logistic regression was $N_{patients} \times (N + N_{clinical})$. Age, gender, self-reported worries about life instability during the lockdown, health status, presence of an essential worker in the household, changes in the frequency of social interactions during quarantine, and current employment status were chosen as non-temporal features for the model training. These features were selected because of the differences between the clinical and non-clinical anxiety groups and correlations with GAD-7. Given the clinical association between social isolation and anxiety disorder in the literature, [187] and the impact of the lockdown impacting the social media app usage for those living alone in our sample, we also included cohabitation status.

4.2.3 Performance Evaluation

The evaluation was performed using *k*-fold cross-validation [71], due to the limited data samples. Ten train-test splits were created from the dataset. Similarly, ten logistic regression models were created and trained for evaluation. Since we had 95 patients, this means that in the first five splits, we trained the model on data from 85 patients and tested the model on data from 10 patients, while in the last five splits trained the model on 86 patients and tested the model on nine patients. The results are summarised with a mean and standard deviation of the model accuracy and area under the receiver operating curve (AUC-ROC) scores.

Finally, we also performed feature importance analysis by computing Shapley additive explanations (SHAP) values [114], which provide an overview of important features in the machine learning models by designating the weight of predictability of each feature positively or negatively to the target variable. We averaged the SHAP values over the 10-fold cross-validation for every feature for each patient.

4.3 Results

4.3.1 Predicting Clinical Anxiety

Only the patients with any communication and social network app usage data during both the pre-lockdown (≥1 out of 42 days) and lockdown period (≥1 out of 51 days) were considered for the model training. This resulted in 95 patients in the model with varying individual app usage data sequences. In these sequences, 8.76% (655/7476) of the communication app and 30.26% (2262/7476) of the social network app usage data were missing in the data set.

After experimenting with several set-ups (2-19 hidden states) of hidden Markov models, an HMM with three hidden states proved to capture the underlying patterns in the data the best according to the AIC and BIC analysis, leading to the most interpretable states. Subfigures A and B in Figure 4.3 provide a sketch of the transition probabilities and means of the 3-state HMM. Temporal variables were normalised before model training, providing the negative means. Large state transition probabilities suggest that the states were relatively stable.

State 2 was the most stable (self-transition probability of 0.88) while transitioning between states 1 and 3 was more likely. State 3 captured days with relatively low commu-

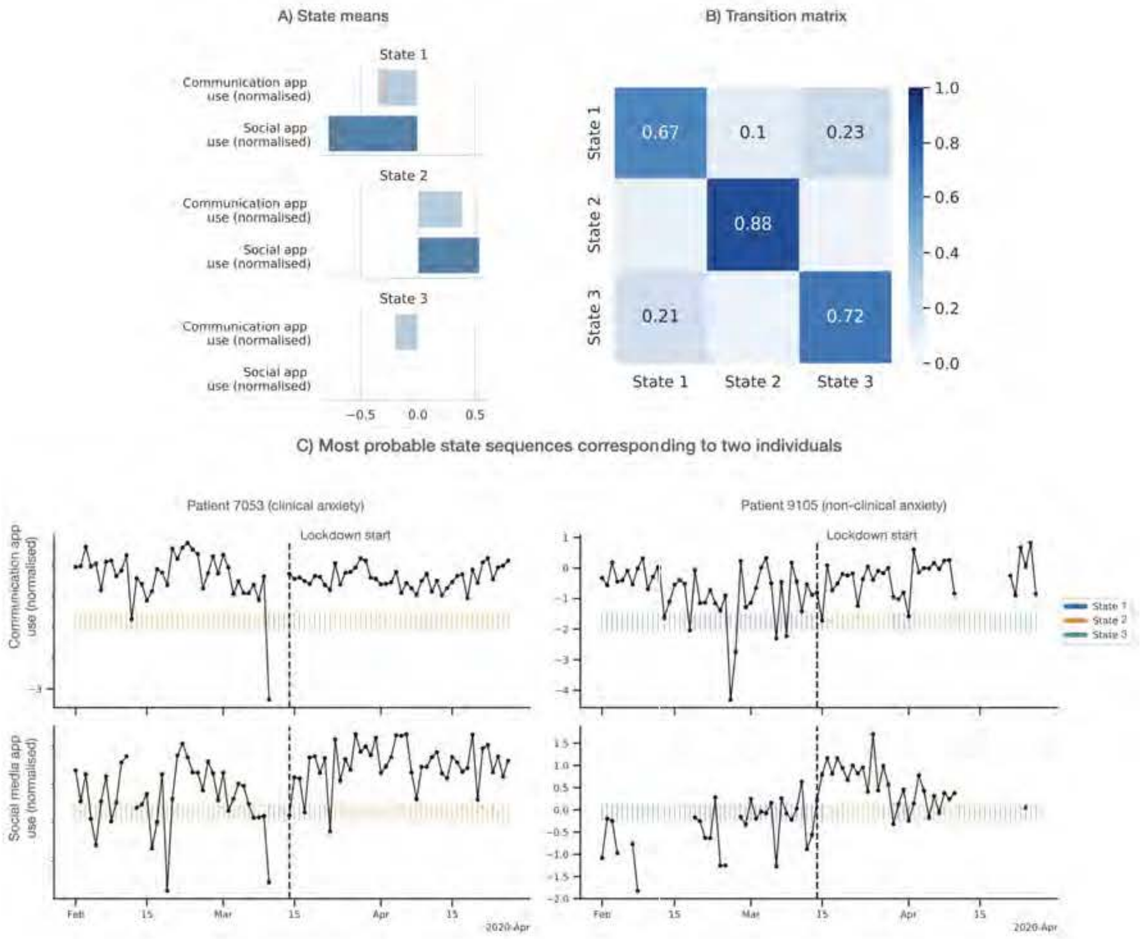


Figure 4.3: The 3-state Hidden Markov model parameters used for temporal data modelling and most probable HMM states applied to daily communication and social media app usage of example individuals with clinical and non-clinical anxiety.

nication app usage and average social networking usage in the sample, while states 1 and 2 captured days with lower and higher app usage, respectively. When applied to individual observation sequences, state 3 preferentially represented the missing observations (i.e., days the apps were not consistently used). State 2 preferentially represented the days of active and consistent social media usage, and state 1 preferentially represented the days of still active (but less so) and volatile usage (Figure 3C). For example, for patient 7053 with clinical anxiety, most days were in state 2, punctuated with three missing/inactive days (state 3), and social networking app usage increased after the lockdown. In the case of patient 9105 with nonclinical anxiety, days after the lockdown were marked with increased communication app usage (state 2), but during the overall period, social networking app usage was less, capturing missing (state 3) and inactive or volatile (state 1) days.

Our model achieved a mean accuracy of 62.30% (SD=16%) and the AUC-ROC score of 0.70 (SD=0.19) on the left-out test sets. Performance metrics in Table 4.1 show that the model performs well on most splits; however, it underperforms on splits 6, 7, and 10. This was partly due to containing non-representative demographic features for the clinical anxiety group (Figure A1 in Appendix B). For example, in case of Split 10, which had the lowest predictability with an AUC-ROC score of 0.40 for the model, we found that clinical anxiety individuals had either atypical risk perception (only one individual reported the presence of essential workers in the household) or self-report patterns (reported clinical anxiety despite having relatively good health and few worries about life instability during the lockdown).

Fold	**Accuracy %**		**AUC-ROC**	
	Training set	**Test set**	**Training set**	**Test set**
1	53.61	87.50	0.67	0.88
2	70.57	60.00	0.80	0.76
3	69.40	60.00	0.79	0.76
4	69.51	80.00	0.77	0.88
5	69.20	70.00	0.77	0.84
6	74.23	45.00	0.81	0.50
7	76.90	47.50	0.80	0.65
8	66.96	77.50	0.77	0.80
9	74.73	55.00	0.80	0.60
10	75.82	40.00	0.81	0.30
Mean (SD)	70.10 (6.70)	62.30 (16.00)	0.78 (0.04)	0.70 (0.19)

Table 4.1: Achieved accuracy and area under the receiver operating curve (AUC-ROC) in the 10-fold cross-validation of the pipeline.

4.3.2 Feature Importance Analysis

To get an overview of which features were most important for the models, we computed the SHAP values of every feature for every sample in each phase of the 10-fold cross-validation and averaged the SHAP values. Figure 4.4 shows the summary plot of feature importance and direction of effects in predicting the clinical anxiety group. The features are ordered downwards by their descending importance and coloured by their values from low to high. Each point is a SHAP value for a feature and an instance, and overlapping points are jittered in the y-axis direction. Positive SHAP values encode the feature's predictability to classify the subject in the clinical anxiety group and negative values in the non-clinical anxiety group.

The majority of non-temporal features, led by the presence of essential workers in the household, outweighed the aggregated representation of the temporal features in importance. Among temporal features, the aggregated posterior probability of state 2 (higher social networking app use) was the most important predictor of the clinical anxiety group. Despite their lower feature importance, states 1 and 3 still provided significant insight into users' longitudinal behaviour, such that inactive and volatile social media usage patterns, specifically in lower communication app usage, predicted the clinical anxiety group. This is also consistent with our finding that the clinical anxiety group's communication app use was significantly lower during the lockdown period.

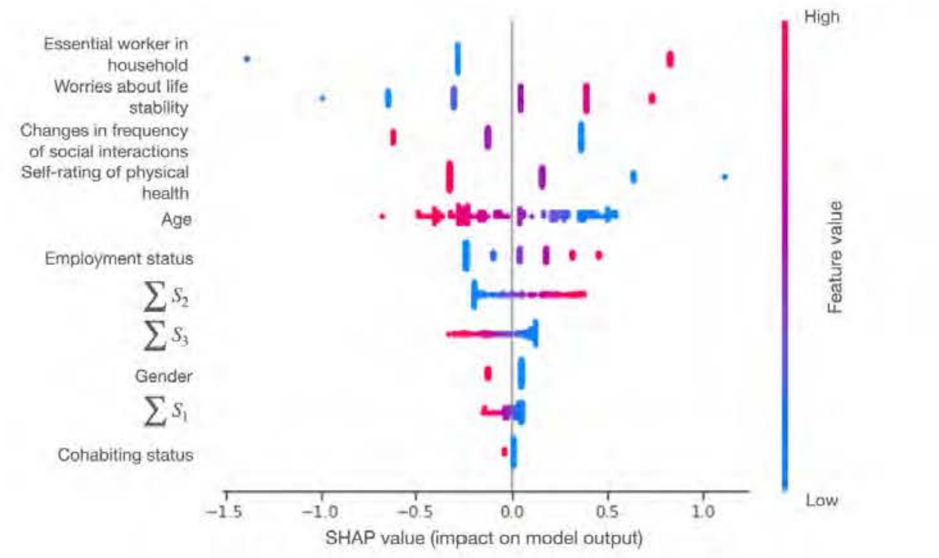


Figure 4.4: Summary plot of feature importance for the logistic regression model trained for the anxiety prediction task. Notations: Covid34 - essential worker in the household, Covid37 - worries about life instability during the lockdown, Covid43 - changes in the frequency of social interactions, Covid25 - self-rating of physical health, $\sum S_i$- the sum of posterior probabilities of the state i over time, $i \in 1, 2, 3$.

4.4 Discussion

Our ML-based model trained on the temporal series of communication and social network app usage and clinically important features of self-report and demographic variables accurately predicted the clinical anxiety group from higher social network app usage and lower communication app usage. Taken together, our ML-based model results suggest that passive tracking of decreased communication app usage and increased social network app usage through the lockdown period can predict users reporting clinical anxiety symptoms, at risk for impaired decision-making, maladaptive coping, and psychiatric sequelae during public health crises and lockdown periods. Early remote detection of at-risk individuals would, in turn, allow allocating limited mental health resources to serve those with the highest need and prevent or ameliorate adverse mental health outcomes.

The analysis was based on observing a small number of patients and should be interpreted with the following limitations. First, the data cannot explain the causal link between app usage and the severity of anxiety. Secondly, besides "general worries about life instability during the lockdown," there were no other independent variables that may reflect the evolution of subjective emotions included in the model to predict the anxiety states at clinical follow-up. Study participants had a daily mood self-reporting option on their smartphones, but such reporting was voluntary, and mood data were largely missing during the lockdown. We acknowledge that our study participants were in an unprecedented and anxiogenic natural circumstance at the time. The lockdown likely increased all users' anxiety and stress levels (mean GAD-7 was high at 9.6, with a clinical cut-off of 10). However, we did not have their baseline GAD-7 collected before the lockdown to make a comparison statement. Therefore, the utility of our model is limited to detecting those whose anxiety symptoms were registering at the clinical severity, i.e., GAD-7≥10.

Our work is the first to suggest that category-based passive sensing of a shift in smartphone usage patterns can be markers of clinical anxiety symptoms. Novel studies to digitally phenotype short-term reports of anxiety using granular behaviours on social media are necessary for public health research when in-person psychiatric evaluations are limited during mandated physical isolation.

CHAPTER 5

ASSESSING WHODAS 2.0 SCORES FROM BEHAVIOURAL BIOMARKERS

The first part of the work presented in this section was submitted to *SAGE Digital Health*. At the same time, the results of the deep learning-based solution are only preliminary and will be ulteriorly published.

5.1 Introduction

Functional limitations are associated with poor clinical outcomes, higher mortality, and disability rates, especially in the elderly [31]. Moreover, they are closely related and used for predicting transitions in daily living/instrumental activities of daily living (ADL/IADL) disability, significantly impacting the quality of life of the elderly and other age groups [127]. COVID-19 has been associated with functional limitations in post-COVID patients, further increasing an already present problem for older adults [52], [161], [190]. Just examining sarcopenia, a progressive loss of muscle due to ageing, the estimated cost of hospitalisations for adults in the United States was USD 40.4 billion [66]. Early detection of an increase in disability is of great importance for clinical practice, as it can still be stabilised or even reversed in the early stages, as in the case of sarcopenia, which can be prevented, treated, and reversed by exercise. Moreover, one of the cornerstones of rehabilitation research is the reduction of disability and restoration of function [55].

There is a great need and much to be gained from defining a way to measure functioning and disability on a relevant scale, ideally daily. However, assessing everyday functioning and disability is complicated due to current measurement modalities (e.g., self-report, proxy-report, clinician ratings) [149]. These reports are time-consuming and tedious to fill in on follow-up visits. In addition, there are disagreements between disciplines about what constitutes a disability and the methods to measure this disability, especially in a clinical setting [122]. Ecological momentary assessment (EMA) allows for more continuous assessment and monitoring of patients without face-to-face appointments and has the crucial advantage of providing data that is more relevant to daily life [126]; however, it still requires active patient input leading to refusal and attrition. Developing adequate passive EMA tools may increase retention and help overcome the limitations of active EMA [153].

Patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) are increasingly recognised as tools providing valuable information about patients' health status and perception of treatment at a particular point in time [80]. Including such tools in the healthcare workflow aims to provide a patient-centred, value-based healthcare system [17]. A commonly used PROM for disability assessment is the second version of the World Health Organisation Disability Assessment Schedule (WHODAS 2.0) [193]. This 36-item questionnaire provides a generic tool to measure health and disability. It assesses difficulties due to health conditions, including diseases or illnesses, short or long-lasting health problems, injuries, mental or emotional problems, and substance-use disorders [92]. WHODAS 2.0 captures the level of functioning in six domains of life: (i) cognition – understanding and communicating, (ii) mobility – moving and getting around, (iii) self-care – attending to one's hygiene, dressing, eating, and staying alone, (iv) getting along – interacting with other people, (v) life activities – domestic responsibilities, leisure, work, and school, (vi) participation – joining in community activities, participating in society. Respondents are asked to reflect over the last 30 days and answer a set of questions, thinking about how much difficulty they had doing the given activities. Using WHODAS 2.0, there is a possible maximum score of five points for all items, indicating a rising level of difficulty in performing the activity in situations experienced over the previous 30 days, and taking into account current health conditions: 1 - none, 2 - mild, 3 - moderate, 4 - severe, and 5 - extreme. A higher final score value, calculated as a total score or score by domain, indicates a higher level of disability [148].

Mobile sensing offers various sources, such as GPS, accelerometer, gyroscope, and light sensor, that can be used to implement behavioural measures [116]. Unlike traditional assessment tools, these technologies enable long-term passive and ecological measurement of patient function that is non-intrusive. While there has been some work in digital mental health and machine learning [82], no studies predict WHODAS 2.0 functionality score changes using smartphone sensor data. These approaches are particularly important since they may enable the analysis of individuals' functioning and disability evolution and provide a clinical tool to monitor the progression and efficacy of treatment. In addition, they provide the opportunity to build targeted just-in-time adaptive interventions in a designated population [133]. Such frameworks deliver interventions within the context of daily life. Including passive data-driven solutions as part of the typical PROM frameworks could enrich existing information and better inform decisions.

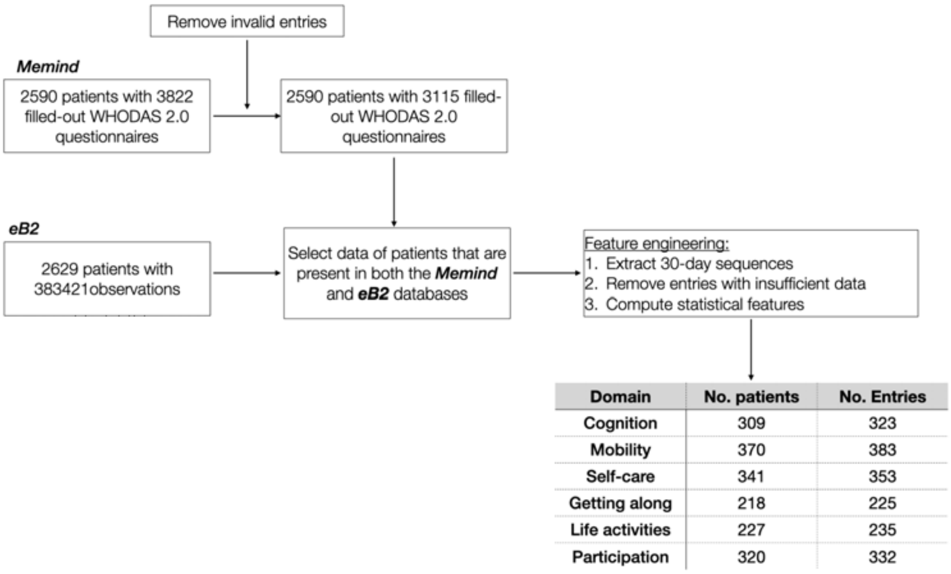
5.2 A Baseline Approach

This work aimed to provide a baseline analysis of the feasibility of using machine learning to predict patients' WHODAS 2.0 functionality scores from passively gathered digital biomarkers. Furthermore, we aimed to determine which behavioural features are the most important for predicting different WHODAS 2.0 domains.

5.2.1 Data Selection and Preprocessing

The *eB2 MindCare* [21], [30] mobile application collects data from different sources (the mobile phone's sensors and wearables) at different intervals. For this work, we focused on the data streams related to patient mobility (daily step count, distance travelled, the number of locations visited by the patient, time spent at home, time spent performing activities such as walking, running, exercising), and time spent asleep. Daily summaries were calculated on the values of these variables, which were then used to extract 64 descriptive, statistical features for characterising the patients' behaviour in a 30-day interval.

Figure 5.1 shows the data selection and feature extraction process. For each domain's score: incomplete answers - if some questions were not answered within a specific domain, incorrect scores for the individual questions - if the registered score was out of the range of the possible scores. After our data filtering, 1526 WHODAS 2.0 domain entries of 396 participants collected between 01/2017 and 04/2021 were selected for our analysis. The cohort of patients had a median age of 44 (IQR: 33, 53) years at baseline, 63.13% (250/396) were female, and 29.04% (115/396) were male. Age and gender information was unavailable for 8.08% (32/396) and 7.83% (31/396) of the participants. Socio-demographic information was not inputted into the model.


Remove invalid entries
Memind
2590 patients with 3822 filled-out WHODAS 2.0 questionnaires
2590 patients with 3115 filled-out WHODAS 2.0 questionnaires
eB2
2629 patients with 383421observations
Select data of patients that are present in both the Memind and eB2 databases
Feature engineering:
1. Extract 30-day sequences
2. Remove entries with insufficient data
3. Compute statistical features

Domain	No. patients	No. Entries
Cognition	309	323
Mobility	370	383
Self-care	341	353
Getting along	218	225
Life activities	227	235
Participation	320	332

Figure 5.1: Data selection and feature engineering flowchart.

The particular questions' scores from 1 to 5 were scaled as suggested by the WHO to either value from 0 to 4 or 0, 1, 1, 2, 2 [148]. Finally, the score by domain was computed as the sum of the scores of the respective questions. These scores will serve as our target for the supervised prediction problem. Figure 5.2 insights about the distribution of the scores for all WHODAS 2.0 domains in the overall population.

To build the input dataset, we cropped a 30-day window of the data sequences for each

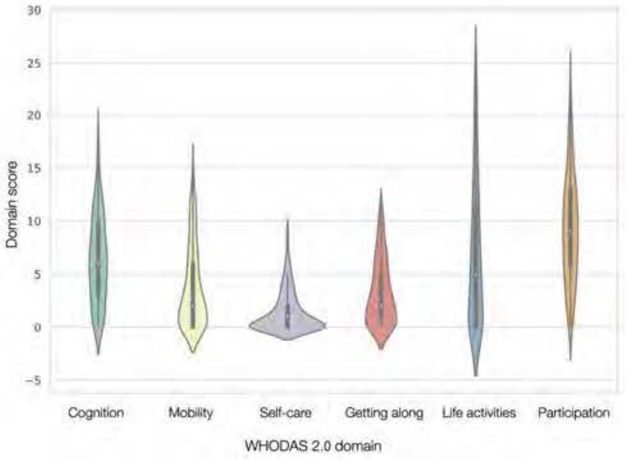


Figure 5.2: The distribution of WHODAS 2.0 functionality scores per domain in the patient cohort.

WHODAS 2.0 entry. For the baseline WHODAS 2.0 score, due to it being registered at enrolling in the study, we consider the next 30-days of observations because no previous mobile sensed data was collected. For follow-up scores, which are usually collected bi-annually, we centred the window on encapsulating 15-days before and after the score. The time-series dataset must first be transformed to be modelled as a supervised learning problem. Therefore, we extracted statistical summary features (count, minimum, maximum, mean, standard deviation, IQR) from the sequences for each variable and obtained a dataset of 64 features. We filtered sequences by requiring every feature to contain at least two counts (days) of data for comprehensive statistics calculation and removing missing values.

We divided the datasets for each domain into two independent subsets based on the patients, ensuring no overlap. The first subset is the training dataset, consisting of 80% of the entries used to fit the feature selection and predictive models. The second dataset was held-out for testing the model performance. The training set was split into four equal folds of 20% for cross-validation. The train-test and cross-validation splits were done by ensuring the grouping of entries of the same user within the same set/fold since the model is user-independent and stratifying by the interquartile range of the WHODAS 2.0 scores. The stratification ensures that the model can train and test low, middle, and high WHODAS 2.0 scores within the population. Then the features were standardised. Moving features to a similar scale helps avoid feature weight problems and provides an interpretable bias in the case of linear regression.

5.2.2 Feature Selection and Predictive Modelling

We used sequential forward selection (SFS) [53], also known as a sequential feature selection or stepwise forward selection, a greedy search algorithm for feature selection, which reduces an initial d-dimensional feature space to a *k*-dimensional feature subspace where

$k < d$. SFS avoids the feature selection stability problems of the lasso with a similar idea as best subset selection but on a reduced set of subsets, which is computationally feasible [72]. In SFS, features are sequentially added to an empty set of features until extra features do not reduce the criterion. To find the best set of features in the case of each domain, we performed a search by iterating from k=1 to 20 with 4-fold cross-validation over the training set and selecting the k with the highest average performance across folds with the same model design as our final model. Finding the best features for predicting each domain provides greater interpretability to our models, which is essential in a clinical setting where clinicians need reliable and straightforward decision rules [145].

Once the best features were found, we trained linear regression models to perform the prediction task. To better suit the ordinal classification problem, we performed a simple modification after the regressor by thresholding the predictions between the minimum (0) and maximum values of the specific domain and rounding. The final models were evaluated on the held-out test set. We computed test mean absolute error (MAE) and test mean absolute percentage error (MAPE) as performance evaluation metrics. MAPE provides a metric to compare the different domains with a different number of questions and different total scores. We applied this approach separately for the different WHODAS 2.0 domains, using all the extracted features and the best subset of features.

5.2.3 Results

We performed the feature selection using SFS, followed by training unregularised linear regressors with the best feature subset for each domain. We then compared the performance to linear regressors trained on the entire feature set.

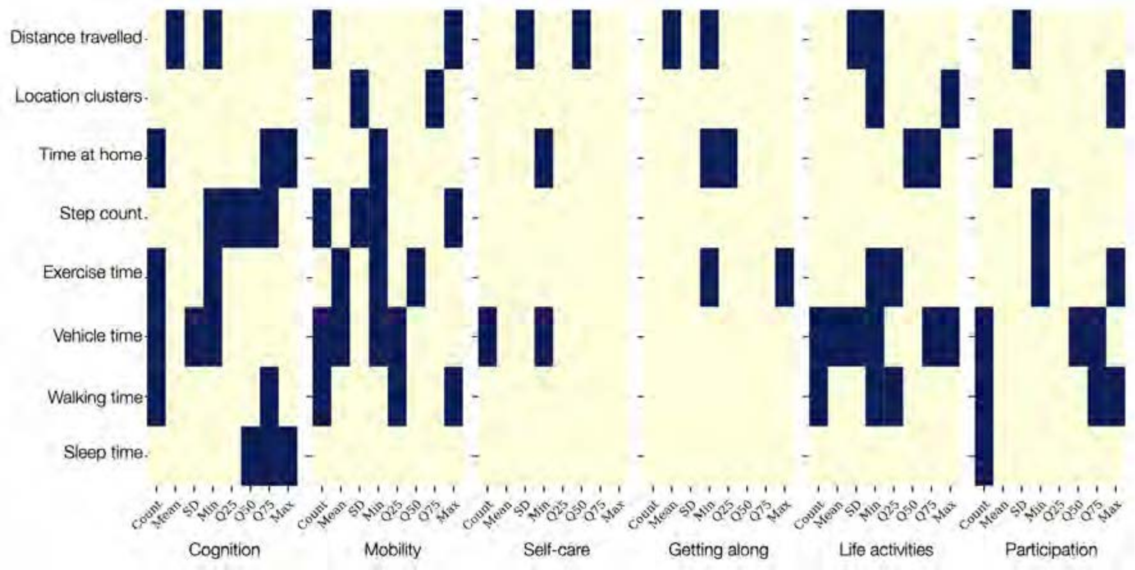


Figure 5.3: Selected features per WHODAS 2.0 domain.

Figure 5.3 provides a graphical summary of the best feature subsets selected across all six domains (see Table A3 in Appendix C for a detailed overview). In each case, the

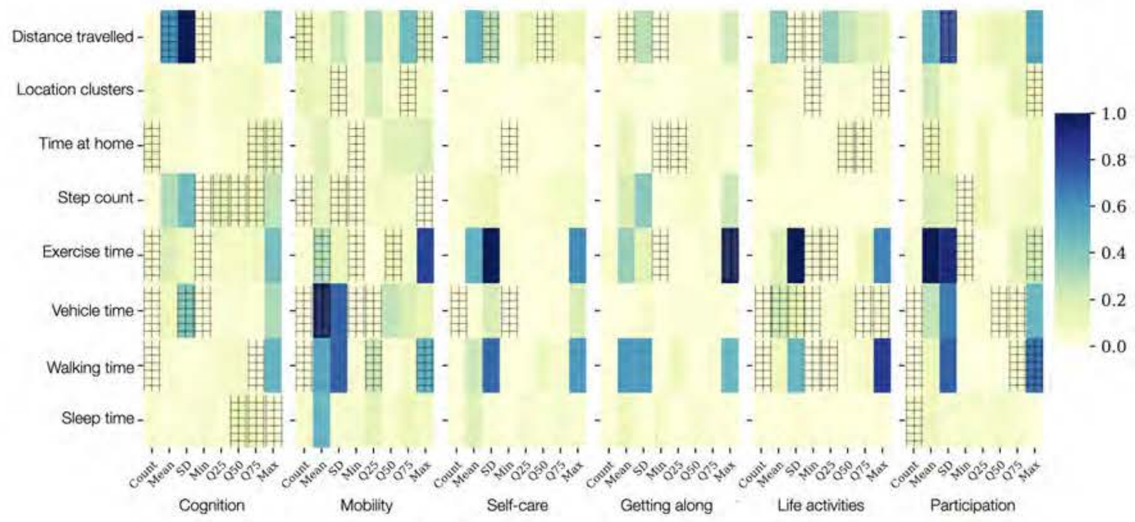


Figure 5.4: The feature importance of the linear regressor with all features per WHODAS 2.0 domain overlaid with the selected ones.

model discarded most of the input features, reducing the feature space to 19, 19, 5, 6, 17, and 13 features from the total of 64 respective to the above domains. Figure 5.4 shows the feature weights of the linear regressor trained with all features per domain overlaid with the best subset features denoted by a grid. Note that the absolute value of regression coefficients per domain model was normalised (0,1) to compare feature importance. While each feature statistic is not shared, it can be seen that the models coincide in the feature groups important to the all feature model. Both models capture the relevant data from each feature group but use different statistics. All the statistics (count, mean, min, max, quartiles) are interrelated, so a selection of a few could be sufficient to summarise the relevant information for the regressor.

Across all domains, at least one statistic of distance travelled and time spent at home was selected. The time spent at home and distance travelled features impart information on daily movement patterns. These movement patterns may indicate many elements of an individual's lifestyle, including work (or lack of work), socialisation (in and out of the home), and isolation, among many others. This focus on movement patterns was further reinforced by including vehicle time, step count, and walking time statistics in multiple domains. Step count and walking time also double physical activity descriptors with exercise time. Physical activity biomarkers proved important in the cognition, mobility, life activities, and participation domains, while sleep-related biomarkers were only selected for the cognition and participation domains. Physical well-being in both exercise and sleep is reasonably related to these domains. The self-care domain and getting along domain were described with the lowest amount of important features compared to the four other domains.

Table 5.1 shows the domain prediction errors as MAE and MAPE for both experimental set-ups. Note that these are negatively-oriented scores, which means lower values are better.

WHODAS 2.0 domain	Score range	Predicting with all 80 features		Predicting with selected features		
		MAE (SD)	MAPE (SD) [%]	No. of features selected	MAE (SD)	MAPE (SD) [%]
Cognition	0-20	3.76 (3.21)	18.84 (16.09)	19	3.55 (2.90)	17.76 (14.54)
Mobility	0-16	3.50 (2.44)	21.91 (15.26)	19	3.40 (1.98)	21.26 (12.42)
Self-care	0-10	1.56 (1.75)	15.69 (17.54)	5	1.48 (1.50)	14.86 (15.09)
Getting along	0-12	3.13 (2.72)	26.11 (22.74)	6	2.44 (1.79)	20.37 (14.96)
Life activities	0-24	7.57 (6.40)	31.56 (26.66)	17	6.53 (5.40)	27.21 (22.52)
Participation	0-24	3.88 (3.14)	16.16 (13.10)	13	3.73 (2.69)	15.54 (11.23)

Table 5.1: Regression evaluation metrics. Notation: MAE=mean absolute error, MAPE=mean absolute percentage error, SD=standard deviation.

The regression models trained on the reduced feature space outperformed, even if only by a small margin, those that were trained on all features. Regression models estimate parameters for every term in the model; therefore, non-informative variables may add uncertainty to the predictions, reducing the overall performance. However, this small margin implies that relevant information is being captured for both models even if the model trained on all features is forced to regress with a larger number of features with no regularisation.

The overall model performance can be explained by the distribution of the outcome variable in the respective data sets: mid-range values dominate in each domain; therefore, it becomes harder for the models to predict the more extreme scores. The performance of a regression model may suffer from the fact that the distribution of the target variable is not normally distributed and skewed. Moreover, the above-elaborated problem is particularly challenging due to missing values and a target better suited for ordinal classification. While linear regression has downsides in this setting, it does have the advantage of not overfitting, in general, and particularly to the noise in real-world data. Using multiple linear regression and a large feature space also allows more flexibility to the linear regression concerning linearity. Non-linear models may perform better, but they are more prone to overfitting and loss in interpretability and ease of explainability, which is important when applying machine learning to a clinical setting [73], [103].

5.2.4 Discussion

Principal Findings

In this work, we addressed the problem of predicting WHODAS 2.0 functionality scores per domain from solely passively collected digital biomarkers. Statistical feature engineering followed by a simple machine learning approach of selecting features through SFS for linear regression showed the feasibility of predicting functionality from passively sensed data. Moreover, using a simple linear regression model for prediction ensured the interpretability of the model's decisions.

Extracting statistical measures of the time series sequences allowed dealing with missing data without applying imputation techniques; however, several entries had to be discarded due to their limited information content (e.g., sequences with completely missing observations per feature). We searched for the most relevant features from an ample feature space, removing non-informative or redundant predictors from the model for each domain. We found that 5-19 features were sufficient for each domain, the most relevant ones being the distance travelled, time spent at home, time spent walking, exercise time, and vehicle time. These features that were most informative for linear regression are biomarkers for daily movement patterns and individuals' physical activity.

Our machine learning–based models, trained on the best feature subsets per domain, outperformed the ones trained on the entire feature space and predicted patients' WHODAS 2.0 functionality scores per domain with a maximum MAPE of 27.21% on the life activities domain and a minimum MAPE of 14.86% on the self-care domain. These are reasonable errors for a linear regression performing a complicated ordinal classification task. The lowest and highest MAPE was seen in the domains with the least features selected and the smallest range of possible scores. Compared to the all-feature model, feature selection did not cause a change in the error ranking of the domains.

Limitations

Although this approach showed promising results, it also has limitations. The automatically-generated wearable device data was passively collected in a real-world setting. This is a strength in ecological validity but has the downside of considerable missing data and noise commonly present in real-world passive data acquisition. A few data quality problems are missing data due to users not wearing the device or incorrect data due to malfunctioning. This may have lowered the predictive performance and biased the variable importance. Missing data also posed a problem in the features used from the eB2 database, as other helpful features, such as app usage and phone unlocks, were filtered out, causing a severely reduced dataset. These features would provide information directly related to social domains. Even with removing these features, our sample size is still quite limited. Overall, the dataset presents many challenges because it is an inclusive combination of two real-world databases with missing values, erroneous entries, and noise.

In many cases, individuals have only a single score, which does not allow for training personalised models that could better account for the intra-individual variations. Although step data should be relevant to determining mobility, an individual's lifestyle, work conditions, and other factors greatly influence step count. We hypothesise that a model that learns individual variability and patterns and then examines the population would be better suited, but having longitudinal large population datasets combining clinical data and wearable data would be a challenge.

Conclusion

This work is the first to suggest a machine learning-based approach for assessing WHODAS 2.0 functionality from passively sensed data. The findings indicate the feasibility of designing a pipeline to monitor patients' functionality over time passively. However, the different results between different WHODAS 2.0 domains show that it is difficult to predict each domain's scores equally well. Nevertheless, the feature selection approach provides an insight into relevant behavioural measures for yielding the predictions, leading to better interpretability of the results, which is important for real-world and clinical application.

5.3 A Deep Learning Approach

Another possible line of work addressing the problem is using deep learning-based temporal methods with the data at a more refined time scale. Instead of statistical feature engineering and selection, these models can learn a representation from a raw input sequence that is most relevant for the prediction problem. As wearable or mobile sensor data collected in the wild are noisy and frequently missing, it is also necessary to apply adequate imputation methods and time-series models to capture underlying patterns in the data. Moreover, including socio-demographic information about the patients can improve the predictions since these features strongly correlate to the individuals' functionality, and groups of patients can share typical behavioural patterns based on these factors.

5.3.1 Data Selection and Preprocessing

We considered 48-half-hour daily summaries of 4 passively collected observations: step count, distance travelled, time spent at home, and exercise time. The mobile sensed data was collected between January 2016 - April 2022 from 2,348 individuals, yielding 516,604 entries (31.5% collected in 2019). The final data set contained many missing observations, as illustrated in Figure 5.5. The overall missingness percentage was over 60% for each variable: distance - 72.13%, step count - 79.04%, time spent at home - 73.64%, and time spent exercising - 60.55%.

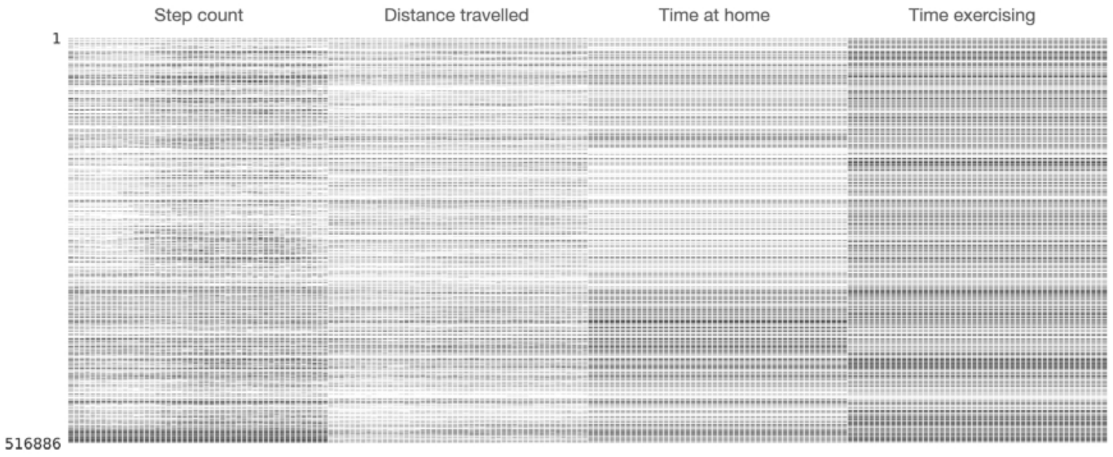


Figure 5.5: Missingness pattern in the mobile sensed data. The shaded areas correspond to the presence of the observation. Each column represents a 30 minutes time slot.

A subgroup of 2011 patients from the two studies had clinical evaluations for the outcomes of interest. Table 5.2 provides an overview of the distribution of socio-demographic information at baseline and the mental health outcome scores in the two study groups. The two health outcomes that we focus on here are the World Health Organisation Disability Assessment Schedule 2.0 (WHODAS 2.0) [122] and the Generalised Anxiety Disorder Assessment (GAD-7) [181] scores.

Table 5.2: The study cohorts.

Variable	Value	Study group	
		A	B
		N = 283	N = 1728
Socio-demographic information at baseline			
Age (years), mean (SD)		42 (14)	43 (15)
Gender, n (%)	Male	102 (36.04%)	526 (30.44%)
	Female	179 (63.25%)	1184 (68.52%)
	Not known	2 (0.71%)	18 (1.04%)
Cohabitating, n (%)	No	50 (17.66%)	177 (10.24%)
	Yes	216 (76.34%)	1517 (87.79%)
	Not known	17 (6.00%)	34 (1.97%)
Family status, n (%)	Single	115 (40.64%)	620 (35.88%)
	Separated	55 (19.43%)	231 (13.37%)
	Widowed	7 (2.47%)	42 (2.43%)
	Married or cohabiting for >6 months	104 (36.75%)	822 (47.57%)
	Not known	2 (0.71%)	13 (0.75%)
Employment status, n (%)	Employed, student or homemaker	122 (43.11%)	811 (46.94%)
	Unemployed without subsidy	45 (15.90%)	272 (15.74%)
	Unemployed with subsidy	14 (4.95%)	149 (8.62%)
	Permanently incapacitated	26 (9.19%)	106 (6.14%)
	Temporarily incapacitated	55 (19.43%)	286 (16.55%)
	Retired	15 (5.40%)	92 (5.32%)
	Not known	6 (2.12%)	12 (0.69%)
Clinical information, median (IQR)			
WHODAS 2.0 mobility score [%]		13 (0, 38)	19 (6, 44)
GAD-7 score		9 (6, 12)	-
Entry statistics, median (min, max)[2]			
No. entries per patient		1 (1, 1)	1 (1, 4)
No. score changes per patient	WHODAS 2.0 mobility	0	0 (0, 2)
	GAD-7	0	-

In **Study group A**, 417 patients have two or more entries, 161 have one change in the score, and 2 have two changes over the study period. In the dichotomised case, this translates to 54 patients having a single change.

We dichotomise the WHODAS 2.0 mobility and GAD-7 scores to create the target outcomes. For the WHODAS 2.0 mobility scores, the cut-off for the negative label is set at 25% of the overall domain score. In contrast, for the GAD-7 score, a cut-off at 10 is considered. In both cases, there is an imbalanced distribution between the two categories, as shown in Figure 5.6.

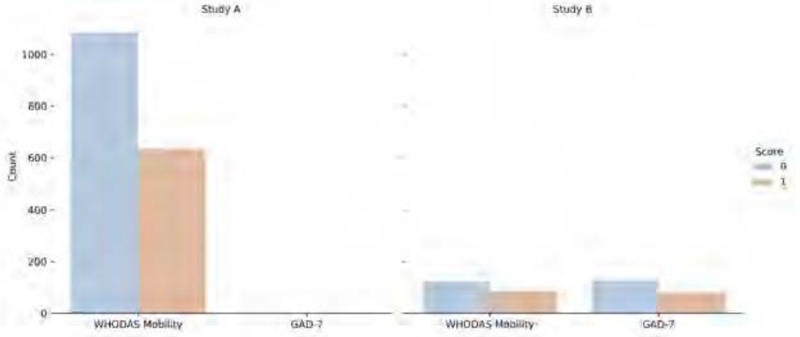


Figure 5.6: The distribution of dichotomised target outcomes.

To build the input data set for the classification task, we cropped a 30-day window of the data sequences for each target label entry. For the baseline score, due to it being registered at enrolling in the study, we consider the next 30-days of observations because no previous mobile sensed data was collected. For follow-up scores, usually collected bi-annually, we centred a 30-day window on encapsulating the most complete observation sequence around the score.

In the case of the socio-demographic covariates, the categorical data were one hot encoded. At the same time, the patient age was binned into ten categories, then one-hot encoded. We introduced an additional category to indicate missingness for covariates that were not reported.

5.3.2 The Proposed Pipeline

Figure 5.7 shows the framework of our approach, consisting of an HMM for data imputation, the LSTM- and self-attention-based temporal encoder, coupled with a dense layer acting as a logistic regressor on the temporal embeddings concatenated with the static covariates.

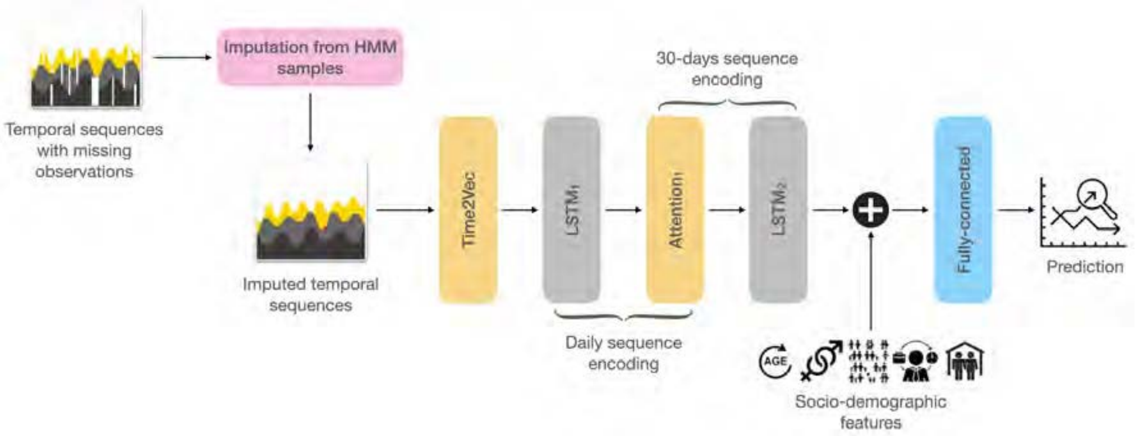

Temporal sequences with missing observations
Imputation from HMM samples
Imputed temporal sequences
Time2Vec
LSTM₁
Attention₁
LSTM₂
30-days sequence encoding
Daily sequence encoding
Fully-connected
Prediction
AGE
Socio-demographic features

Figure 5.7: The overall structure of the designed pipeline.

Due to the high percentage of missing data, imputing statistical measures such as the mean, median, or even interpolation fails. These approaches do not generalise to wearable characteristics or participant behaviour, can reduce variability in the data set, and introduce bias. Probabilistic generative models, such as hidden Markov models (HMMs) [156] can learn the underlying distributions in a data set by adjusting the model parameters to best account for the data to maximise the evidence, even in the presence of missing data.

Only those 48-slot daily patient sequences with at least 80% of observations were considered for HMM training. After this elimination process, 91047 sequences were used to train the HMMs with different numbers of states, $n = \{2, 3, ...23\}$. The best model was selected using the Bayesian and Akaike information criteria [48] on a randomly selected subset of 10000 sequences with varying missingness. Given this model, we imputed the missing observations repeatedly during the the mini-batch stochastic gradient descent. Every time a new batch of data was generated,the sequences were decoded using the

Viterbi algorithm [54], and the missing observations were imputed by samples generated from the corresponding most probable state.

Our proposed pipeline performs feature encoding for the daily information by applying Time2Vec [91], followed by two LSTM [81] encoders with self-attention [197] for the 30-day input sequence. A feed-forward layer on top of the second attention layer's outputs concatenated with a simple embedding of the socio-demographic data is then used to get the predictions.

Time2Vec gives a model-agnostic vector representation for time. Consisting of a periodic activation function and a linear term, it can capture the periodicity of time series signals and the non-periodic patterns that depend on time. Mathematically, for a given scalar notion of time τ, Time2Vec of τ, denoted as $\mathbf{t2v}(\tau)$, is a vector of size $k+1$ defined as:

$$\mathbf{t2v}(\tau)[i] = \begin{cases} \omega_i \tau + \phi_i, & \text{if } i = 0 \\ \mathcal{F}(\omega_i \tau + \phi_i), & \text{if } 1 \leq i \leq k \end{cases} \tag{5.1}$$

where $\mathbf{t2v}(\tau)[i]$ is the i^{th} element of $\mathbf{t2v}$, $\mathcal{F}$ is a periodic activation function, and the ω_i and ϕ_i parameters are learnable.

The LSTM layers encode the input sequences into a fixed-length internal representation. In contrast, the attention layers learn to pay selective attention to the inputs and relate them to items in the output. While this increases the computational burden of the model, it results in a more targeted and better-performing model. In addition, the model can also show how attention is paid to the input sequence when predicting the output.

Understanding the relationship between input and output, namely, which within-day and within-month temporal patterns contribute to correct predictions in a model like we proposed here, is complicated since massive non-linear operations are involved. Therefore, we used the self-attention weights to interpret the importance of the input signals in the functionality assessment task. We visualised self-attention as heat maps to understand the overall importance of features and time. Besides understanding which temporal patterns contribute to the outcome, these self-attention weights can provide insights into relevant changes over time, which is paramount to determining worsening of patient state.

5.3.3 Experiments

The data from *Study B* was used for cross-validation in all experiments, except in the task transfer learning set-up. We kept the data from *Study A* as a held-out test set and in the cross-validation of the mentioned experiment. All models were trained for 35 epochs, using an Adam optimiser with a learning rate of $1e-3$ and batch size of 64.

We evaluated prediction performance using the area under the receiver operating characteristic curve (AUC-ROC) and area under the precision-recall curve (AUC-PRC) scores [168] to gain valuable insights into the classification performance on the imbalanced problems. We report the average score and the corresponding standard deviation from the cross-validation for all evaluations unless mentioned otherwise. Furthermore, we report the performance on the unseen data set, except for the task transfer learning experiment.

Defining the Baseline

We started by re-using the pipeline defined in [184] as a baseline for prediction performance. We applied sequential forward selection (SFS) and logistic regression with L2-regularisation on the manually extracted statistical summary features (count, minimum, maximum, mean, standard deviation, IQR) from the sequences for each variable and combined them with the socio-demographic information. After the feature extraction, there were 20% missing values in the dataset (24.49% in the step count, 11.13% in the distance travelled, 52.58% in the time at home and 6.51% in the time at exercising feature), which we imputed using simple mean imputation. These values occur because we do not discard sequences with a single measurement in the features in order to be able to directly compare the results.

Nested Cross-validation

We first performed model hyper-parameter optimisation and model selection using a nested cross-validation approach [99]. As such, a k-fold cross-validation procedure for model hyper-parameter optimisation is nested inside a k-fold cross-validation procedure for model selection. This way, the risk of the search procedure overfitting the original data set is reduced, and we gain insight into the average model performance. By randomly sampling possible model architecture candidates from a pre-defined search space of possible hyper-parameter values [19], we try to discover a set of hyper-parameters that perform well on the data set.

Table 5.3: The search space for the model hyper-parameters.

Hyper-parameter	**Search space**
Time2Vec	
Embedding dimension	$\{4, 6, 8, 10, 12\}$
Activation function	$\{\sin, \cos\}$
LSTM	
Hidden dimension - Block 1	$\{x + 8 \mid x \in \mathbb{N} \cap [32, 128]\}$
Hidden dimension - Block 2	$\{x + 8 \mid x \in \mathbb{N} \cap [64, 256]\}$
Bidirectional	{True, False}
Number of layers	$\{1, 2, 3\}$
Dropout	$\{0.1, 0.2, 0.3\}$

We used $k = 5$ for the hyper-parameter search and tested 10 random combinations of model hyper-parameters from a pre-defined search space (see Table 5.3). Nested cross-validation with $k = 5$ folds in the outer loop would fit and evaluate 250 models. The final model is configured by applying the outer loop to the entire data set. This model is then used to make predictions on the unseen data.

Ablation Study

When creating a complex machine learning model, it is helpful to understand the impact of each of its components separately [125]. Therefore, we defined an ablation study, systematically eliminating parts of the model, and analysed its effect on overall model performance. We used 3-fold cross-validation to estimate how the models are expected to perform when used to make predictions on data not used during training and to find the optimal number of epochs to train the model to avoid overfitting. The models were then trained on the entire data set for the found number of epochs and evaluated on the held-out test set in each case.

Temporal Encoder Pre-training

Given the limited labelled sample size, we propose using a transfer learning approach for the temporal encoder. First, we pre-train the temporal encoder weights to perform a generic task, such as predicting the average mobility biomarkers of the next day based on the previous 30-days. Then we use the model fit on this auto-regressive task as the starting point for a model in the functionality prediction setting, such that it would lead to better general embedding of the time series sequences regardless of the target label of interest. We extracted 20,272 30-day sequences with 7-day overlap, for which observations were collected for all the features. The pre-training was run for five epochs.

We compare two transfer learning approaches: feature extraction and fine-tuning. In the first setting, we freeze the weights of the temporal encoder part; hence we solely use it for temporal feature extraction, and we train the classification layer of the network. The second approach consists of training the whole model on the task-specific dataset and adjusting the weights of the temporal encoder. By slightly changing the temporal encoder weights, we expect the network to be better adjusted to the specific 30-day periods around the evaluation.

Task Transfer Learning

The core symptom of general anxiety disorder is chronic, excessive, and uncontrolled worry [163], which is reflected in individuals' behavioural patterns. Therefore, it is reasonable to expect that we can apply the above-defined pipeline to predict GAD-7 outcomes from the same behavioural biomarkers. However, in this case, we are facing a significantly lower labelled sample size, which makes it difficult for such complex models to learn to generalise well to unseen data instead of simply overfitting the training set. Therefore, we propose fine-tuning the model trained on the WHODAS 2.0 outcome prediction task to predict the GAD-7 scores. This way, the new task can be learned by transferring knowledge from a related task that has already been learned [140].

5.3.4 Results

Finding the Model Architectures

After analysing the elbow points of both the AIC and BIC information criteria, we found that five hidden components best captured the data's underlying patterns. Therefore, we used that HMM in the following experiments to infer the most probable state sequence for each daily data sequence and impute the missing observations from samples generated from the corresponding state each time a mini-batch is loaded.

The hyper-parameter tuning for the resulted in the following architecture:

- Time2Vec with embedding dimension 4 and sine activation
- 2-layer uni-directional LSTM blocks with 64 recurrent units each, incorporating 0.1 recurrent dropout rate in each block

Baseline, Ablation and Temporal Pre-training

In Table 5.4, we summarise the model performance results of the baseline approach along with the ablation and transfer learning experiments. The DL pipeline outperformed the baseline approach in the AUC-ROC score, but achieved slightly worse performance in AUC-PRC scores in the cross-validation. On the held-out test set the DL model outperformed the baseline in the AUC-PRC sense.

We will now examine the results of the ablation study in reference to the full pipeline. Removing the attention layer, but keeping the Time2Vec layer led to a significant performance decrease in cross-validation and hold-out test. Removing the Time2Vec block led to lower AUC-ROC and slightly higher AUC-PRC in cross-validation, while in test provided the best performance. The model without the Time2Vec and self-attention layer performed similarly to the model without Time2Vec in cross validation; however, while in test it did improve upon the full model it performed worse than the former model. It can be seen that removal of Time2Vec is overall helpful to the DL pipeline, but retaining the self-attention layer provides the greatest performance while also providing the advantage of greater interpretability.

With respect to transfer learning compared to without, pre-training the temporal encoder block of the entire model using all the available data sequences led to a slight improvement in the model performances. With the fine-tuning approach, the average cross-validation model performance increased to 0.558 AUC-PRC, as opposed to the 0.532 AUC-PRC achieved without pre-training. In contrast, the model only reaches 0.545 average AUC-PRC after training with the feature extraction approach. Nonetheless, the feature extraction approach reaches a slightly higher AUC-PRC on the held-out test set.

As Figure 5.8 shows, more attention is paid on average to the activity in the evening hours (slots 36-47), very low attention weights are associated with the night activity, and

Table 5.4: Model performance comparison for the binary WHODAS 2.0 mobility impairment prediction task.

Experiment	Model	Cross-validation performance score - mean (SD)		Performance on held-out test set	
		AUC-ROC	AUC-PRC	AUC-ROC	AUC-PRC
Baseline	Random	0.500 (0.000)	0.369 (0.000)	0.500	0.411
	SFS + LR [184]	0.684 (0.028)	**0.553 (0.032)**	**0.603**	0.536
	DL pipeline	**0.693 (0.023)**	0.532 (0.051)	0.586	**0.542**
Ablation study	DL pipeline	**0.693 (0.023)**	0.532 (0.051)	0.586	0.542
	No self-attention	0.666 (0.035)	0.528 (0.062)	0.596	0.541
	No Time2Vec	0.677 (0.021)	0.538 (0.037)	**0.605**	**0.570**
	No Time2Vec & self-attention	0.681 (0.040)	**0.539 (0.066)**	0.591	0.552
Transfer learning	DL pipeline				
- feature extraction approach		0.674 (0.029)	0.545 (0.055)	0.575	**0.538**
- fine-tuning approach		**0.675 (0.027)**	**0.558 (0.075)**	**0.579**	0.533

varying patterns during the day in both cohorts. As for the monthly sequences, the attention weights are pretty uniform over the 30-day interval in both groups, with occasionally more attention being assigned to the last days of the period.

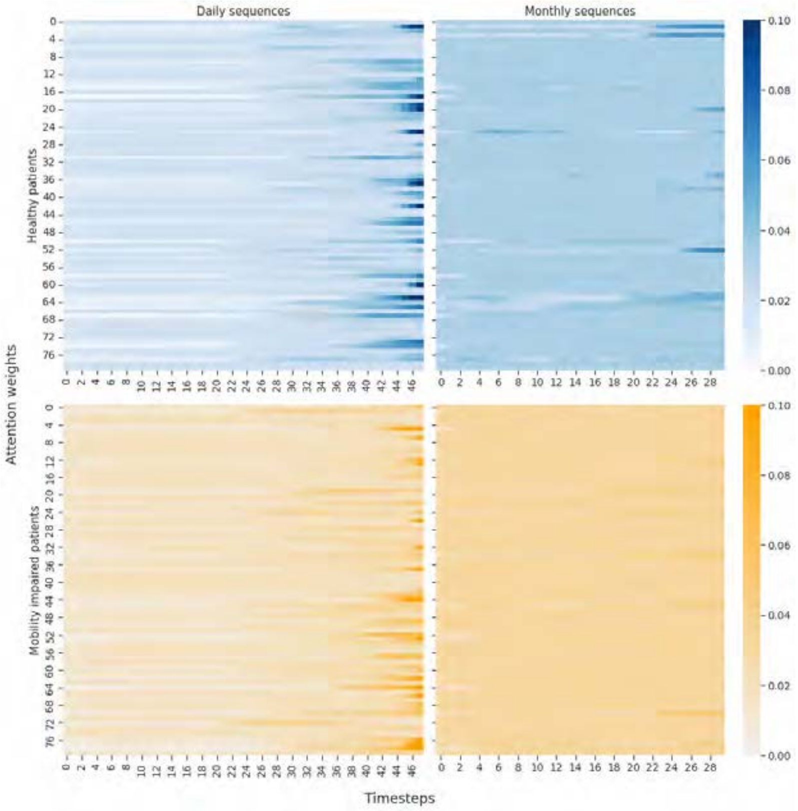


Figure 5.8: Daily and monthly average attention weights for 160 randomly selected patients grouped by their mobility difficulty levels.

When analysing the attention weights at the patient level (Figure 5.9), we can see different patterns arise based on mobility impairment and possibly individual-level differences. In the case of the healthy patient, the larger daily attention weights consistently appear in the second half of the day. In contrast, some days, more attention is paid to the night hours for the patient with mobility difficulty. Finally, we also analysed but did not find a clear correlation between the data missingness and the attention weights, which indicates that the weights are assigned in the function of the observation values rather than driven by the missingness factor.

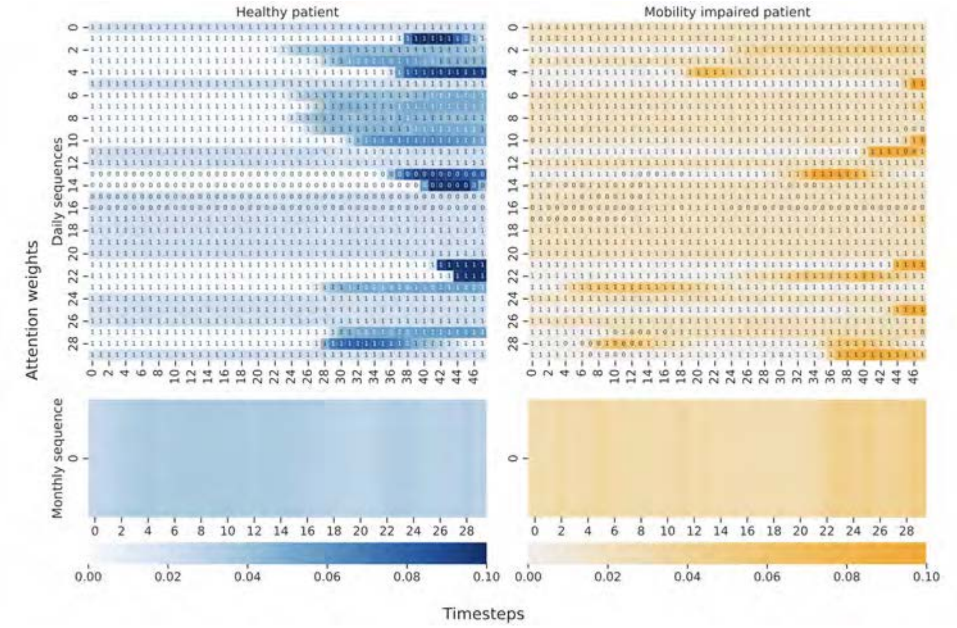


Figure 5.9: Daily and monthly attention weights for 2 randomly selected of patients with different mobility difficulty levels. We indicate with 1 the presence of a sample, while with 0 its missingness.

Task Transfer Learning

Table 5.5 shows the dichotomised GAD-7 classification performance scores with and without transfer learning, respectively. The performance achieved by the simple baseline model is almost equivalent to random guessing. The DL model overfitted the training data when we tried learning the GAD-7 prediction task from scratch since the sample size was relatively small. The achieved performance is slightly better than random, but the variance between splits is relatively large.

Table 5.5: Performance results of predicting the dichotomised GAD-7 scores with and without transfer learning.

Experiment	**Model**	**Cross-validation performance score - mean (SD)**	
		AUC-ROC	**AUC-PRC**
Baseline	Random	0.500 (0.000)	0.392 (0.011)
	SFS + LR [184]	0.505 (0.077)	0.450 (0.068)
	DL pipeline	**0.518 (0.143)**	**0.463 (0.148)**
Task transfer learning	DL pipeline		
- feature extraction approach		0.530 (0.107)	0.504 (0.148)
- fine-tuning approach		**0.603 (0.121)**	**0.556 (0.148)**

When we fine-tuned the model trained on the mobility impairment classification task, a significant performance increase was reached. This process works because the features are suitable for both base and target tasks and because the other data set is more extensive and covers a broader range of covariates. Hence the starting weights of the network are more representative. However, the overall model performance still leaves room for improvement.

5.3.5 Discussion

Principal Findings

This work tackled common problems in modelling mental health outcomes from passively sensed digital biomarkers. One of the main difficulties we faced was dealing with the large amounts of missing data in a meaningful way. We used HMMs trained on the 48-half-hour time slot sequences describing patients' daily activities, which we then used for imputation.

Then we aimed to leverage deep learning techniques to automatically learn the underlying patterns in the monthly patient sequences to predict mobility difficulty and generalised anxiety outcomes. We showed that the proposed transfer learning methods could improve the performance of target outcome estimation, especially in the case of data sets with few samples. Our results showed that even though the binary classification performance varies on each split, which is expected partly due to the non-uniform representation of certain socio-demographic features in the data set, the variance was not especially significant; hence the models are quite robust to the data shifts.

Applying a pre-training step for the temporal encoder block of the model helped with a more meaningful initialisation of the model weights for the classification task at hand. Moreover, as the data set from Study B covered a more comprehensive range of socio-demographic representations, that might have helped to avoid covariate shifts between training and test sets in the task transfer set-up on the much smaller data set of Study A.

The self-attention heatmaps showed different general within-day and within-month patterns emerging in the healthy versus mobility-impaired cohorts. In addition, we could analyse the different emerging patterns over time of patients' who manifested a change in their mobility difficulty level between visits. These simple visualisations provide a helpful tool for clinicians to gain insights into the individuals' activity patterns and what led to the decline. They can trigger proper interventions to help slow down or reverse the decline.

Limitations

Although our approach showed promising results, it faces additional challenges and leaves room for improvement. One of the limitations of our work comes from the large percentage of missing observations in the mobile data and the sparsity of labels. Another limitation commonly occurring in health applications is the relatively small labelled sample size. A larger patient cohort and multiple labels per patient could help train more robust models and even allow for a data-driven personalisation, thus accounting for inter- and intra-individual differences in behavioural patterns.

Moreover, we interpreted the temporal patterns found significant by the model via visualising the self-attentions. However, such interpretation only explains the variation of the behavioural patterns regarding the outcome of interest. This work could further be extended to bring more interpretability to the decision-making, providing better insights for clinicians.

Conclusion

Previous work on the topic used manually extracted features from the mobile sensed sequences, avoiding the missingness and intra-day variations. In this work, we investigated using a deep learning model with multimodal inputs, complemented by a hidden Markov model for missing value imputation, for the prediction tasks. Applying this pipeline results in accurate predictions and interpretability of intra-day and intra-month variations concerning the outcome of interest, thanks to the self-attention layers. Moreover, our transfer learning approach shows promising results in efficiently diversifying the prediction tasks, even to smaller data sets.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

6.1 From Digital Phenotyping to Healthcare Solutions

Digital phenotyping and machine learning provide an unprecedented opportunity for mental health specialists to benefit from in-situ and continuous patient information. In the digital era that we live in, evidence-based psychiatry tailored to each individual based on objectively measured behavioural patterns could allow for personalised predictions of early diagnosis, treatment selection, and dose adjustment in order to reduce the burden of disease. Machine learning is predisposed to address these issues in psychiatry's era of personalised medicine.

Machine learning uses quantitative models to learn general patterns from a series of observations without explicit instructions [25], [67]. These methods make few a priori assumptions, allowing the data to "speak for themselves" and can process vast amounts of data. We can differentiate between supervised and unsupervised methods. The former is specialised in the best possible outcome prediction given a labelled dataset, while the latter effectively discovers statistical configurations in an unlabelled dataset. The increasing data availability, computing power and cheaper data storage have encouraged a continuous surge in research for new algorithms and applications in various fields [118], and the field of mental health is no exception.

This work aimed to contribute to the state-of-the-art with three main ML applications in the mental health domain while also tackling the missing data problem. We addressed the problems of mood, general anxiety and functionality prediction, proving feasibility and drawing baselines. We tried to answer relevant questions, such as how to process appropriately and impute these mobile sensed data streams to predict mental health outcomes and what data types are necessary and relevant for the different predictive tasks.

The first project focused on emotional state prediction from behavioural markers derived from passively sensed data. These regularly sampled but frequently missing and heterogeneous time series were analysed using probabilistic latent variable models for data averaging and feature extraction: mixture model (MM) and hidden Markov model (HMM). The extracted features were combined with a classifier to predict the emotional state. Finally, a personalised Bayesian model was proposed to improve the performance by considering the individual-specific differences in the data by applying a different classifier bias term for each patient.

Probabilistic generative models proved to be good preprocessing and feature extractor tools for data with large percentages of missing observations. Models that considered the posterior probabilities of the MM and HMM latent states outperformed those that did not by more than 20%, suggesting that the underlying behavioural patterns identified were meaningful for individuals' overall emotional state. The best performing generalised models achieved a 0.81 area under the receiver operating characteristic curve and a 0.71 area under the precision-recall curve when predicting self-reported emotional valence from behaviour in held-out test data. Moreover, the proposed personalised models demonstrated that accounting for individual differences through a simple hierarchical model can substantially improve emotional state prediction performance without relying on previous days' data.

Next, the aim was to predict clinical anxiety based on the Generalised Anxiety Disorder 7-item scale from time-series data of communication and social networking app usage during the COVID-19 lockdown in Madrid, and anxiety-associated clinical survey variables, including cohabitation with essential workers, worries about life instability, changes in social interactions, and health status. We designed a 2-step approach that combined a probabilistic generative model, namely a hidden Markov model for temporal data processing and aggregation, with logistic regression to predict the binary outcome (clinical anxiety versus nonclinical anxiety) by dichotomised GAD-7.

The pipeline achieved 62.30% (SD=16%) mean accuracy and 0.70 (SD=0.19) area under the receiver operating curve in the 10-fold cross-validation in predicting the clinical anxiety group. Patients who reported severe anxiety symptoms were less active in communication apps after the mandated lockdown and more engaged in social network apps overall, suggesting a different pattern of digital social behaviour in adapting to the crisis. Passive-sensing of a shift in category-based social media app usage during the lockdown can predictively model digital biomarkers of individuals at risk for psychiatric sequelae.

The subsequent project focused on developing a machine learning-based model to passively follow up on patients' functionality over time from mobility descriptor digital biomarkers and socio-demographic data. The WHODAS 2.0 Questionnaire was used as a functionality measurement tool, which queries whether the individual had difficulty performing a set of tasks over the past 30 days.

To start, we applied a sequential feature selection to each WHODAS 2.0 domain (cognition, mobility, self-care, getting along, life activities, participation) on statistical measures (minimum, maximum, mean, median, standard deviation, IQR) extracted from the one-month long time-series data statistical measures (minimum, maximum, mean, median, standard deviation, IQR). Finally, we predicted the WHODAS 2.0 functional domain scores using linear regression using the best feature subsets. Our machine learning-based models for predicting patients' WHODAS functionality scores per domain achieved a mean absolute percentage error varying between 14.86% and 27.21% among the domains with a set of interpretable features for each domain. Our findings show the feasibility of using machine learning-based methods to assess functional health solely from passively sensed mobile data. The feature selection step provides a set of interpretable features for each domain, ensuring better explainability of the models' decisions.

Then, we proposed a pipeline that performs temporal sequence embedding using LSTM and attention layers and then merges the socio-demographic data to perform the predictions. Given the limited labelled sample size, we proposed a transfer learning approach. First, we pre-trained the temporal encoder weights to perform a generic task, such as predicting the average mobility biomarkers for the next 30-days. After that, we used the model fit on this auto-regressive task as the starting point for a model in the mobility difficulty prediction setting, leading to the better general embedding of the time series sequences. In addition, we suggested a simple task transfer learning approach to fine-tune the model for predicting anxiety outcomes. HMMs are used again to deal with the missing data, similar to the emotional state prediction case.

Applying this pipeline results in more accurate predictions and interpretability of intra-day and intra-month variations concerning the outcome of interest, thanks to the self-attention layers. Moreover, our transfer learning approach shows promising results in efficiently diversifying the prediction tasks, even to smaller data sets. However, these results are preliminary, and further improvements are needed for the model performance to become clinically relevant.

As such, these works contribute to this relatively new and heterogeneous field of digital phenotyping for mental health, attempting to make a step towards developing such tools that could improve clinical workflow and passive patient monitoring. We acknowledge that psychiatry is very complex, intertwined with difficult questions and dilemmas that are not easily solvable. Digital phenotyping can bring us closer to finding the answers and analysing the problems from a different perspective.

6.2 Challenges

Several challenges must be considered when using ML techniques in mental health applications. As such, rather than replacing other research or analytic approaches, ML has the potential to add value to mental health research.

The quality and availability of training data inevitably limit such models' performance. The size of cohorts in most studied datasets is considerably small, and their phenotypic descriptions (medical history, comorbidities, progression in symptoms, questionnaire evaluations, treatment and response) are insufficient. Within the medical field, the mental health domain was claimed to capture the most extensive amounts of data [50]. However, the sample sizes are still significantly smaller than the millions of samples in non-medical domains where ML methods achieve state-of-the-art performance. Besides the limited sample size, insufficient specificity and granularity of the patients' behavioural information also hinder exploiting ML technologies.

Retrospective collection of data across sites faces the problems of differences in data quality, acquisition parameters, preprocessing procedures, assessment methods and questionnaires, missing data and socio-demographic aspects of the cohorts. This heterogeneity often leads to reduced prediction performance as the amount of available data increases [202]. On the other hand, prospective data collection, which standardises data acquisition, ensures higher data comparability.

Notably, in the case of digital phenotyping, massive longitudinal data can be accumulated; however, in addition to privacy concerns, the digital sensors are constantly changing and improving, which increases heterogeneity in the collected data. Moreover, missing observations are frequent due to intended or unintended user behaviour, loss of communication, sensor failure, or insufficient energy [11]. These factors significantly influence ML models' trustworthiness and efficiency; hence, robust methods need to be developed to account for these issues.

Accumulating observational data without identifying and accounting for influences of the possible confounding factors can inflate the prediction performance if the training and test data are mutually dependent, regardless of how subtle this dependence is. Further bias might arise due to recruitment specifications, for example, if the study restricts recruitment to subjects exposed to mental health institutions rather than also including individuals without diagnosed mental problems.

Collaboration between researchers and clinicians to share and harmonise data is therefore a crucial bottleneck in accessing the training data. Moreover, minimal research demonstrates the efficacy of the developed models in real-world settings. Thus further research is needed to ensure that models that appear promising in lab settings will still be efficient when deployed, mainly if applied across different contexts (across data acquisition means, geographic locations, populations, and different clinical settings) [202].

Finally, we cannot ignore that ML algorithms are prone to biases and other limitations, which can negatively impact the study results' validity, objectivity and reproducibility. These errors and biases mainly result from the data or the sample selection. The data used for model training reflects the real world, which leads to concerns about how societal structural inequalities appear in this data and how the models will potentially pick these up. Moreover, ML algorithms and most currently defined and investigated tasks strongly rely on self-reported measures and questionnaire-based clinical assessment for ground truth, i.e. classification labels that the ML algorithm can learn from and evaluate. This introduces a certain level of undesired subjectivity to the digital phenotyping problem. Further, performing analysis on small study groups does not imply that the results will generalise well to the entire population because these samples might not accurately reflect the diversity of the population. Hence, such biases should always be considered and accounted for to train fair models.

6.3 Directions for Future Work

Future studies should build upon the current thesis to continue exploring the potential for digital personalised medicine by integrating digital phenotyping and digital interventions in the mental health field. Predicting the specific questionnaire outcomes could allow for early intervention and relapse prevention, possibly the two most promising early warning services that could be offered to individuals who would otherwise find it hard to sustain self-monitoring.

Additional research should address the problem of representation learning from unlabelled data, allowing a better analysis of patient-specific patterns. Moreover, speech and language content aspects can inform mental disorders' diagnosis and outcome prediction. Hence, combining the passively sensed data with audio speech samples and text-based information (e.g. diary entries, doctors' notes) could provide a more descriptive and discriminative representation of patients to make predictions.

Lastly, as more data becomes available, shifting the focus towards personalised models becomes more relevant. These approaches can learn an explicit parameter set for each individual instead of focusing on global patterns over the whole population. As such, the predominant patterns of variation among patients could be better captured. Like in the emotional state prediction case, significant improvements could be achieved using personalised approaches.

When it comes to predictive modelling, especially in the clinical setting, interpretability, i.e. understanding why the model made a particular decision, becomes crucial — knowing 'why' can provide insights into the problem, the data and also the reason why a model might fail. Unlike linear models, which are well studied and understood, DL models with millions of parameters are less well understood, and more research should be focused on this problem. Collaboration with psychologists also plays a crucial role here.

6.4 Ethical Considerations

All of the challenges discussed above raise critical ethical issues, including the ethics of collecting, storing and sharing mental health data, as well as the level of autonomy and privacy afforded to ML systems. Therefore, there is a need to consider these issues, especially in such vulnerable groups as people with mental health conditions, when conducting research using digital phenotyping.

Besides the generic guidelines for ethical conduct, such as the World Medical Association's Declaration of Helsinki [8] or the United States Belmont Report [155], specific guidelines for research involving mobile sensing in mental health have recently been investigated [26], [119], [120].

When designing a mobile sensing-based study, an important issue is what device to use. It is essential to balance the risk of using commercial devices, the inherent lack of control over the data, and the device's obtrusiveness for participants. The Digital Health Decision-Making Checklist [134] calls upon researchers to review the privacy policies and terms of service of the chosen commercial devices carefully and to consider how "the technology can be tailored to the end user" and how "technology is accessible to diverse populations".

Study participants should always be informed about what information is collected and with whom this information could be shared during the consent process. Participants should have as much control over the data collection process as possible, and the option to temporarily stop data collection on the refusal to enter/continue participating.

A significant body of research focuses on ethical problems around adequate data anonymisation [44], [57], [60]; however, anonymisation does not exclude the risk of sensitive personal information leaks. Participants may not be aware of the detail and scope of information that can be inferred from the mobile sensed or even missing data [57]. Discussing such risks must be an essential part of the informed consent process.

Lastly, one cannot ignore concerns regarding the potential impact of digital phenotyping - based solutions on the clinical workflow, and the patient experience of continuous monitoring. These topics remain relevant and need further investigation as digital phenotyping increasingly infiltrates clinical practice.

6.5 Conclusions

In clinical care, digital phenotyping, which captures a variety of objective data streams in patients' everyday lived experiences, is a constantly surging topic expected to improve mental health early prevention, diagnosis and treatment. With the help of machine learning, behavioural fluctuations can be captured and used to provide additional insights into the disease evolution at the individual level. This additional knowledge could provide solid ground for interventions, early detection of symptom worsening, or predicting relapses.

Many studies have emerged in this context. This work aimed to enrich the state-of-the-art by developing pipelines that can serve passive patient monitoring. The provided solutions, as such, are proof of concept, which require further clinical validation to be deployable in the clinical workflow. Still, the results are promising and lay some foundations for future research and collaboration among clinicians, patients, and computer scientists. Nevertheless, as with any promising new approach, the risks and unintended consequences must be considered to ensure the safe and trusted development of digital phenotyping-based clinical solutions.

BIBLIOGRAPHY

[1] L. P. A/S, *Digital phenotyping - turning our smartphones inward*, Apr. 2019. [Online]. Available: https://leoinnovationlab.com/2019/04/11/digital-phenotyping-turning-our-smartphones-inward/.

[2] S. Abdullah *et al.*, 'Sensing technologies for monitoring serious mental illnesses,' *IEEE MultiMedia*, vol. 25, no. 1, pp. 61–75, 2018.

[3] S. A. Adams *et al.*, 'The effect of social desirability and social approval on self-reports of physical activity,' *American journal of epidemiology*, vol. 161, no. 4, pp. 389–398, 2005.

[4] H. Akaike, 'A new look at the statistical model identification,' *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[5] A. Althubaiti, 'Information bias in health research: Definition, pitfalls, and adjustment methods,' *Journal of multidisciplinary healthcare*, vol. 9, p. 211, 2016.

[6] J. Alvarez-Lozano *et al.*, 'Tell me your apps and i will tell you your mood: Correlation of apps usage with bipolar disorder state,' in *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, 2014, pp. 1–7.

[7] M. D. Anestis *et al.*, '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*, vol. 48, no. 7, pp. 607–613, 2010.

[8] W. M. Association *et al.*, 'Declaration of helsinki. ethical principles for medical research involving human subjects,' *Jahrbuch Für Wissenschaft Und Ethik*, vol. 14, no. 1, pp. 233–238, 2009.

[9] A. Baccarelli *et al.*, 'Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the seveso chloracne study,' *Chemosphere*, vol. 60, no. 7, pp. 898–906, 2005.

[10] D. Bahdanau *et al.*, 'Neural machine translation by jointly learning to align and translate,' *arXiv preprint arXiv:1409.0473*, 2014.

[11] S. Bähr *et al.*, 'Missing data and other measurement quality issues in mobile geolocation sensor data,' *Social Science Computer Review*, vol. 40, no. 1, pp. 212–235, 2022.

[12] S. Bai *et al.*, 'An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,' *arXiv preprint arXiv:1803.01271*, 2018.

[13] I. Barnett *et al.*, 'Relapse prediction in schizophrenia through digital phenotyping: A pilot study,' *Neuropsychopharmacology*, vol. 43, no. 8, pp. 1660–1666, 2018.

[14] I. Barnett *et al.*, 'Inferring mobility measures from gps traces with missing data,' *Biostatistics*, vol. 21, no. 2, e98–e112, 2020.

[15] M. L. Barrigón *et al.*, 'User profiles of an electronic mental health tool for ecological momentary assessment: Memind,' *International journal of methods in psychiatric research*, vol. 26, no. 1, e1554, 2017.

[16] G. E. Batista *et al.*, 'An analysis of four missing data treatment methods for supervised learning,' *Applied artificial intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[17] J. F. Baumhauer *et al.*, 'Value-based healthcare: Patient-reported outcomes in clinical decision making,' *Clinical Orthopaedics and Related Research®*, vol. 474, no. 6, pp. 1375–1378, 2016.

[18] J. J. V. Bavel *et al.*, 'Using social and behavioural science to support covid-19 pandemic response,' *Nature human behaviour*, vol. 4, no. 5, pp. 460–471, 2020.

[19] J. Bergstra *et al.*, 'Random search for hyper-parameter optimization.,' *Journal of machine learning research*, vol. 13, no. 2, 2012.

[20] J. Berrevoets *et al.*, 'To impute or not to impute?–missing data in treatment effect estimation,' *arXiv preprint arXiv:2202.02096*, 2022.

[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 and uHealth*, vol. 6, no. 12, e9472, 2018.

[22] C. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Springer, 2006, ISBN: 9780387310732. [Online]. Available: https://books.google.es/books?id=qWPwnQEACAAJ.

[23] P. Bonilla-Escribano *et al.*, 'Assessment of e-social activity in psychiatric patients,' *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2247–2256, 2019.

[24] G. E. Box *et al.*, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[25] L. Breiman, 'Statistical modeling: The two cultures (with comments and a rejoinder by the author),' *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[26] S. Breslin *et al.*, 'Research ethics for mobile sensing device use by vulnerable populations,' *Social Science & Medicine*, vol. 232, pp. 50–57, 2019.

[27] L. E. Burke *et al.*, 'Ecological momentary assessment in behavioral research: Addressing technological and human participant challenges,' *Journal of medical Internet research*, vol. 19, no. 3, e7138, 2017.

[28] J. Busk *et al.*, ‘Forecasting mood in bipolar disorder from smartphone self-assessments: Hierarchical bayesian approach,’ *JMIR mHealth and uHealth*, vol. 8, no. 4, e15028, 2020.

[29] L. Canzian *et al.*, ‘Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis,’ in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1293–1304.

[30] P. Carretero *et al.*, ‘Ecological momentary assessment for monitoring risk of suicide behavior,’ in *Behavioral Neurobiology of Suicide and Self Harm*, Springer, 2020, pp. 229–245.

[31] A. M. Chamberlain *et al.*, ‘Multimorbidity, functional limitations, and outcomes: Interactions in a population-based cohort of older adults,’ *Journal of comorbidity*, vol. 9, 2019.

[32] C. Che *et al.*, ‘An rnn architecture with dynamic temporal matching for personalized predictions of parkinson’s disease,’ in *Proceedings of the 2017 SIAM international conference on data mining*, SIAM, 2017, pp. 198–206.

[33] K.-O. Cheng *et al.*, ‘Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data,’ *Pattern recognition*, vol. 45, no. 4, pp. 1281–1289, 2012.

[34] C.-H. Cho *et al.*, ‘Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study,’ *Journal of medical Internet research*, vol. 21, no. 4, e11029, 2019.

[35] K. Cho *et al.*, ‘Learning phrase representations using rnn encoder-decoder for statistical machine translation,’ *arXiv preprint arXiv:1406.1078*, 2014.

[36] D.-H. Choi *et al.*, ‘The impact of social media on risk perceptions during the mers outbreak in south korea,’ *Computers in Human Behavior*, vol. 72, pp. 422–431, 2017.

[37] R.-G. Cirstea *et al.*, ‘Correlated time series forecasting using multi-task deep neural networks,’ in *Proceedings of the 27th acm international conference on information and knowledge management*, 2018, pp. 1527–1530.

[38] V. P. Cornet *et al.*, ‘Systematic review of smartphone-based passive sensing for health and wellbeing,’ *Journal of biomedical informatics*, vol. 77, pp. 120–132, 2018.

[39] N. Council *et al.*, *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press, 2010, ISBN: 9780309186513. [Online]. Available: https://books.google.es/books?id=%5C_CSF1v2c8jQC.

[40] D. S. Courvoisier *et al.*, 'Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics.,' *Psychological assessment*, vol. 24, no. 3, p. 713, 2012.

[41] V.-A. Darvariu *et al.*, 'Quantifying the relationships between everyday objects and emotional states through deep learning based image analysis using smartphones,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–21, 2020.

[42] R. Dawkins *et al.*, *The Extended Phenotype: The Gene as the Unit of Selection*. Freeman, 1982, ISBN: 9780716713586. [Online]. Available: `https://books.google.es/books?id=5DNmQgAACAAJ`.

[43] M. De Choudhury *et al.*, 'Predicting postpartum changes in emotion and behavior via social media,' in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 3267–3276.

[44] Y.-A. De Montjoye *et al.*, 'Unique in the crowd: The privacy bounds of human mobility,' *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[45] A. P. Dempster *et al.*, 'Maximum likelihood from incomplete data via the em algorithm,' *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[46] S. DeSarbo *et al.*, 'An alternating least-squares procedure for estimating missing preference data in product-concept testing,' *Decision Sciences*, vol. 17, no. 2, pp. 163–185, 1986.

[47] E. Dogan *et al.*, 'Smartphone-based monitoring of objective and subjective data in affective disorders: Where are we and where are we going? systematic review,' *Journal of medical Internet research*, vol. 19, no. 7, e262, 2017.

[48] N. Dridi *et al.*, 'Akaike and bayesian information criteria for hidden markov models,' *IEEE Signal processing letters*, vol. 26, no. 2, pp. 302–306, 2018.

[49] *Eb2 evidence-based behavior*, 2022. [Online]. Available: `https://eb2.tech/?lang=en`.

[50] H. A. Eyre *et al.*, 'Tech giants enter mental health,' *World Psychiatry*, vol. 15, no. 1, p. 21, 2016.

[51] X. Feng *et al.*, 'Automatic instance selection via locality constrained sparse representation for missing value estimation,' *Knowledge-Based Systems*, vol. 85, pp. 210–223, 2015.

[52] C. Fernández-de-Las-Peñas *et al.*, 'Post-covid functional limitations on daily living activities are associated with symptoms experienced at the acute phase of sars-cov-2 infection and internal care unit admission: A multicenter study,' *Journal of Infection*, vol. 84, no. 2, pp. 248–288, 2022.

[53] F. J. Ferri *et al.*, 'Comparative study of techniques for large-scale feature selection,' in *Machine Intelligence and Pattern Recognition*, vol. 16, Elsevier, 1994, pp. 403–413.

[54] G. D. Forney, 'The viterbi algorithm,' *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[55] W. R. Frontera *et al.*, *DeLisa's physical medicine and rehabilitation: principles and practice*. Lippincott Williams & Wilkins, 2019.

[56] B. D. Fulcher, 'Feature-based time-series analysis,' *arXiv preprint arXiv:1709.08055*, 2017.

[57] D. Fuller *et al.*, 'Ethical implications of location and accelerometer measurement in health research studies with mobile sensing devices,' *Social Science & Medicine*, vol. 191, pp. 84–88, 2017.

[58] I. C.-H. Fung *et al.*, 'Ebola and the social media.,' *The Lancet*, 2014. DOI: `https://doi.org/10.1016/S0140-6736(14)62418-1`.

[59] A. Gaggioli *et al.*, 'From mobile mental health to mobile wellbeing: Opportunities and challenges,' *Medicine Meets Virtual Reality 20*, pp. 141–147, 2013.

[60] S. Gambs *et al.*, 'Show me how you move and i will tell you who you are,' in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, 2010, pp. 34–41.

[61] A. Gangemi *et al.*, 'Two reasoning strategies in patients with psychological illnesses,' *Frontiers in psychology*, p. 2335, 2019.

[62] J. Gao *et al.*, 'Mental health problems and social media exposure during covid-19 outbreak,' *Plos one*, vol. 15, no. 4, e0231924, 2020.

[63] A. Gershon *et al.*, 'Electronic ecological momentary assessment (ema) in youth with bipolar disorder: Demographic and clinical predictors of electronic ema adherence,' *Journal of Psychiatric Research*, vol. 116, pp. 14–18, 2019.

[64] C. M. Gillan *et al.*, 'Smartphones and the neuroscience of mental health,' *Annual Review of Neuroscience*, vol. 44, p. 129, 2021.

[65] D. A. A. Gnana *et al.*, 'Literature review on feature selection methods for high-dimensional data,' *International Journal of Computer Applications*, vol. 136, no. 1, pp. 9–17, 2016.

[66] S. Goates *et al.*, 'Economic impact of hospitalizations in us adults with sarcopenia,' *The Journal of frailty & aging*, vol. 8, no. 2, pp. 93–99, 2019.

[67] I. Goodfellow *et al.*, *Deep learning*. MIT press, 2016.

[68] W. R. Gove *et al.*, 'Response bias in surveys of mental health: An empirical investigation,' *American journal of Sociology*, vol. 82, no. 6, pp. 1289–1317, 1977.

[69] J. W. Graham, 'Missing data analysis: Making it work in the real world,' *Annual review of psychology*, vol. 60, pp. 549–576, 2009.

[70] B. T. Griffeth *et al.*, 'A psychiatric-specific entrustable professional activity for the evaluation of prospective psychiatric residents: Towards a national standard,' *MedEdPORTAL*, vol. 13, p. 10 584, 2017.

[71] T. Hastie *et al.*, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[72] T. Hastie *et al.*, 'Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons,' *Statistical Science*, vol. 35, no. 4, pp. 579–592, 2020.

[73] D. M. Hawkins, 'The problem of overfitting,' *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[74] L. Hawryluck *et al.*, 'Sars control and psychological effects of quarantine, toronto, canada,' *Emerging infectious diseases*, vol. 10, no. 7, p. 1206, 2004.

[75] K. He *et al.*, 'Deep residual learning for image recognition,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[76] W. He, 'Load forecasting via deep neural networks,' *Procedia Computer Science*, vol. 122, pp. 308–314, 2017.

[77] I. for Health Metrics *et al.*, *Global health data exchange (ghdx)*, 2022. [Online]. Available: https://vizhub.healthdata.org/gbd-results/.

[78] J. Henley, *Lockdown living: How europeans are avoiding going stir crazy*, Mar. 2020. [Online]. Available: http://www.theguardian.com/world/2020/mar/28/lockdown-living-europe-activities-coronavirus-isolation.

[79] H. Hewamalage *et al.*, 'Recurrent neural networks for time series forecasting: Current status and future directions,' *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021.

[80] J. P. Higgins *et al.*, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

[81] S. Hochreiter *et al.*, 'Long short-term memory,' *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[82] S. Hornstein *et al.*, 'Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach,' *Digital Health*, vol. 7, 2021.

[83] T. Huybrechts *et al.*, 'How to estimate moments and quantiles of environmental data sets with non-detected observations? a case study on volatile organic compounds in marine water samples,' *Journal of Chromatography A*, vol. 975, no. 1, pp. 123–133, 2002.

[84] N. C. Jacobson *et al.*, ‘Digital biomarkers of social anxiety severity: Digital phenotyping using passive smartphone sensors,’ *Journal of medical Internet research*, vol. 22, no. 5, e16875, 2020.

[85] S. Jäger *et al.*, ‘A benchmark for data imputation methods,’ *Frontiers in big Data*, p. 48, 2021.

[86] N. Jaques *et al.*, ‘Predicting students’ happiness from physiology, phone, mobility, and behavioral data,’ in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2015, pp. 222–228.

[87] C. S. Jensen *et al.*, ‘Temporal data models,’ in *Encyclopedia of Database Systems*. Boston, MA: Springer US, 2009, pp. 2952–2957, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_394. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_394.

[88] S. M. Jungmann *et al.*, ‘Health anxiety, cyberchondria, and coping in the current covid-19 pandemic: Which factors are related to coronavirus anxiety?’ *Journal of anxiety disorders*, vol. 73, p. 102 239, 2020.

[89] H. Junninen *et al.*, ‘Methods for imputation of missing values in air quality data sets,’ *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, 2004.

[90] N. Kasabov, ‘Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach,’ *Pattern Recognition Letters*, vol. 28, no. 6, pp. 673–685, 2007.

[91] S. M. Kazemi *et al.*, ‘Time2vec: Learning a vector representation of time,’ *arXiv preprint arXiv:1907.05321*, 2019.

[92] T. A. Kelley, ‘International consortium for health outcomes measurement (ichom),’ *Trials*, vol. 16, no. 3, pp. 1–1, 2015.

[93] J.-O. Kim *et al.*, ‘The treatment of missing data in multivariate analysis,’ *Sociological Methods & Research*, vol. 6, no. 2, pp. 215–240, 1977.

[94] D. Kimhy *et al.*, ‘Mobile assessment guide for research in schizophrenia and severe mental disorders,’ *Schizophrenia bulletin*, vol. 38, no. 3, pp. 386–395, 2012.

[95] E. Koeze *et al.*, *The virus changed the way we internet*, Apr. 2020. [Online]. Available: https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html.

[96] T. Kolenik *et al.*, ‘Persuasive technology for mental health: One step closer to (mental health care) equality?’ *IEEE Technology and Society Magazine*, vol. 40, no. 1, pp. 80–86, 2021.

[97] K. Kroenke *et al.*, ‘Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection,’ *Annals of internal medicine*, vol. 146, no. 5, pp. 317–325, 2007.

[98] E. Kross *et al.*, 'Social media and well-being: Pitfalls, progress, and next steps,' *Trends in Cognitive Sciences*, vol. 25, no. 1, pp. 55–66, 2021.

[99] D. Krstajic *et al.*, 'Cross-validation pitfalls when selecting and assessing regression and classification models,' *Journal of cheminformatics*, vol. 6, no. 1, pp. 1–15, 2014.

[100] V. Kumar *et al.*, 'Feature selection: A literature review,' *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[101] J. Lee *et al.*, 'Personalized mortality prediction driven by electronic medical data and a patient similarity metric,' *PloS one*, vol. 10, no. 5, e0127428, 2015.

[102] K. J. Lee *et al.*, 'Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework,' *Journal of Clinical Epidemiology*, vol. 134, pp. 79–88, 2021, ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2021.01.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089543562100010X.

[103] P. Li *et al.*, 'A survey on implicit bias of gradient descent,' in *2022 14th International Conference on Computer Research and Development (ICCRD)*, IEEE, 2022, pp. 108–114.

[104] Y. Liang *et al.*, 'A survey on big data-driven digital phenotyping of mental health,' *Information Fusion*, vol. 52, pp. 290–307, 2019.

[105] R. LiKamWa *et al.*, 'Moodscope: Building a mood sensor from smartphone usage patterns,' in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 389–402.

[106] Z. C. Lipton *et al.*, 'A critical review of recurrent neural networks for sequence learning,' *arXiv preprint arXiv:1506.00019*, 2015.

[107] R. J. Little *et al.*, 'The prevention and treatment of missing data in clinical trials,' *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[108] R. J. Little *et al.*, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[109] Q. Liu *et al.*, 'Stein variational gradient descent: A general purpose bayesian inference algorithm,' *Advances in neural information processing systems*, vol. 29, 2016.

[110] Y. Liu *et al.*, 'Stein variational policy gradient,' *arXiv preprint arXiv:1704.02399*, 2017.

[111] Z.-g. Liu *et al.*, 'Classification of incomplete data based on belief functions and k-nearest neighbors,' *Knowledge-Based Systems*, vol. 89, pp. 113–125, 2015.

[112] H. Lu *et al.*, 'Stresssense: Detecting stress in unconstrained acoustic environments using smartphones,' in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 351–360.

[113] J. H. Lubin *et al.*, 'Epidemiologic evaluation of measurement data in the presence of detection limits,' *Environmental health perspectives*, vol. 112, no. 17, pp. 1691–1696, 2004.

[114] S. M. Lundberg *et al.*, 'A unified approach to interpreting model predictions,' *Advances in neural information processing systems*, vol. 30, 2017.

[115] Y. Ma *et al.*, 'Daily mood assessment based on mobile phone sensing,' in *2012 ninth international conference on wearable and implantable body sensor networks*, IEEE, 2012, pp. 142–147.

[116] S. Majumder *et al.*, 'Smartphone sensors for health monitoring and diagnosis,' *Sensors*, vol. 19, no. 9, p. 2164, 2019.

[117] N. K. Malhotra, 'Analyzing marketing research data with incomplete information on the dependent variable,' *Journal of marketing research*, vol. 24, no. 1, pp. 74–84, 1987.

[118] J. Manyika *et al.*, *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.

[119] N. Martinez-Martin *et al.*, 'Data mining for health: Staking out the ethical territory of digital phenotyping,' *NPJ digital medicine*, vol. 1, no. 1, pp. 1–5, 2018.

[120] N. Martinez-Martin *et al.*, 'Ethical development of digital phenotyping tools for mental health applications: Delphi study,' *JMIR mHealth and uHealth*, vol. 9, no. 7, e27343, 2021.

[121] S. M. Mattingly *et al.*, 'The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers,' in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.

[122] C. McKibbin *et al.*, 'Assessing disability in older patients with schizophrenia: Results from the whodas-ii,' *The Journal of nervous and mental disease*, vol. 192, no. 6, pp. 405–413, 2004.

[123] A. Mehrotra *et al.*, 'Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, 2017.

[124] A. Mehrotra *et al.*, 'Using autoencoders to automatically extract mobility features for predicting depressive states,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–20, 2018.

[125] R. Meyes *et al.*, *Ablation studies in artificial neural networks*, 2019. eprint: arXiv:1901.08644.

[126] C. Miguelez-Fernandez *et al.*, 'Evaluating attention-deficit/hyperactivity disorder using ecological momentary assessment: A systematic review,' *ADHD Attention Deficit and Hyperactivity Disorders*, vol. 10, no. 4, pp. 247–265, 2018.

[127] M. E. Miller *et al.*, 'Physical activity, functional limitations, and disability in older adults,' *Journal of the American Geriatrics Society*, vol. 48, no. 10, pp. 1264–1272, 2000.

[128] D. C. Mohr *et al.*, 'Personal sensing: Understanding mental health using ubiquitous sensors and machine learning,' *Annual review of clinical psychology*, vol. 13, pp. 23–47, 2017.

[129] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[130] M. Morrison-Valfre, *Foundations of Mental Health Care-E-Book*. Elsevier Health Sciences, 2016.

[131] M. B. Morshed *et al.*, 'Prediction of mood instability with passive sensing,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–21, 2019.

[132] R. Muñoz-Navarro *et al.*, 'Screening for generalized anxiety disorder in spanish primary care centers with the gad-7,' *Psychiatry research*, vol. 256, pp. 312–317, 2017.

[133] I. Nahum-Shani *et al.*, 'Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support,' *Annals of Behavioral Medicine*, vol. 52, no. 6, pp. 446–462, 2018.

[134] C. Nebeker *et al.*, 'Development of a decision-making checklist tool to support technology selection in digital health research,' *Translational Behavioral Medicine*, vol. 10, no. 4, pp. 1004–1015, 2020.

[135] K. Ng *et al.*, 'Personalized predictive modeling and risk factor identification using patient similarity,' *AMIA Summits on Translational Science Proceedings*, vol. 2015, p. 132, 2015.

[136] M. Y. Ni *et al.*, 'Mental health, risk factors, and social media use during the covid-19 epidemic and cordon sanitaire among the community and health professionals in wuhan, china: Cross-sectional survey,' *JMIR mental health*, vol. 7, no. 5, e19009, 2020.

[137] Nigel *et al.*, *Lifting of lockdown in spain - full details of all phases for all regions*, Feb. 2021. [Online]. Available: https://www.spainenglish.com/2020/06/18/lifting-lockdown-spain-full-details-phases/.

[138] A. Norbury *et al.*, 'Social media and smartphone app use predicts maintenance of physical activity during covid-19 enforced isolation in psychiatric outpatients,' *Molecular psychiatry*, vol. 26, no. 8, pp. 3920–3930, 2021.

[139] R. O'neill *et al.*, 'The prevention and treatment of missing data in clinical trials: An fda perspective on the importance of dealing with it,' *Clinical Pharmacology & Therapeutics*, vol. 91, no. 3, pp. 550–554, 2012.

[140] E. S. Olivas *et al.*, *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global, 2009.

[141] J.-P. Onnela, 'Opportunities and challenges in the collection and analysis of digital phenotyping data,' *Neuropsychopharmacology*, vol. 46, no. 1, pp. 45–54, 2021.

[142] W. H. Organization, *Adolescent mental health*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health%5C#:%5C~:text=Key%5C%20facts,illness%5C%20and%5C%20disability%5C%20among%5C%20adolescents..

[143] W. H. Organization, *Mental health*, 2022. [Online]. Available: https://www.who.int/health-topics/mental-health.

[144] W. H. Organization, *Mental health and covid-19: Early evidence of the pandemic's impact: Scientific brief, 2 march 2022*, 2022. [Online]. Available: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1.

[145] G. Orrù *et al.*, 'Machine learning in psychometrics and psychological research,' *Frontiers in psychology*, vol. 10, p. 2970, 2020.

[146] A. Ortiz *et al.*, 'Electronic monitoring of self-reported mood: The return of the subjective?' *International journal of bipolar disorders*, vol. 4, no. 1, pp. 1–8, 2016.

[147] L. V. Panlilio *et al.*, 'Stress, craving and mood as predictors of early dropout from opioid agonist therapy,' *Drug and alcohol dependence*, vol. 202, pp. 200–208, 2019.

[148] M. Paton *et al.*, 'Clinimetrics: World health organization disability assessment schedule 2.0.,' *Journal of Physiotherapy*, vol. 66, no. 3, pp. 199–199, 2020.

[149] T. L. Patterson *et al.*, 'Ucsd performance-based skills assessment: Development of a new measure of everyday functioning for severely mentally ill adults,' *Schizophrenia bulletin*, vol. 27, no. 2, pp. 235–245, 2001.

[150] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in python,' *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[151] J. Pohle *et al.*, 'Selecting the number of states in hidden markov models: Pragmatic solutions illustrated using animal movement,' *Journal of Agricultural, Biological and Environmental Statistics*, vol. 22, no. 3, pp. 270–293, 2017.

[152] J. P. Pollak *et al.*, ‘Pam: A photographic affect meter for frequent, in situ measurement of affect,’ in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2011, pp. 725–734.

[153] A. Porras-Segovia *et al.*, ‘Real-world feasibility and acceptability of real-time suicide risk monitoring via smartphones: A 6-month follow-up cohort,’ *Journal of Psychiatric Research*, vol. 149, pp. 145–154, 2022.

[154] S. A. Prince *et al.*, ‘A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review,’ *International journal of behavioral nutrition and physical activity*, vol. 5, no. 1, pp. 1–24, 2008.

[155] U. S. N. C. for the Protection of Human Subjects of Biomedical *et al.*, *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*. Department of Health, Education, and Welfare, National Commission for the . . ., 1978, vol. 2.

[156] L. R. Rabiner, ‘A tutorial on hidden markov models and selected applications in speech recognition,’ *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[157] M. Rahman *et al.*, ‘Kdmi: A novel method for missing values imputation using two levels of horizontal partitioning in a data set,’ in *International Conference on Advanced Data Mining and Applications*, Springer, 2013, pp. 250–263.

[158] R. Razavi-Far *et al.*, ‘Similarity-learning information-fusion schemes for missing data imputation,’ *Knowledge-Based Systems*, vol. 187, p. 104 805, 2020.

[159] L. Reuman *et al.*, ‘Uncertainty as an anxiety cue at high and low levels of threat,’ *Journal of behavior therapy and experimental psychiatry*, vol. 47, pp. 111–119, 2015.

[160] D. L. Reynolds *et al.*, ‘Understanding, compliance and psychological impact of the sars quarantine experience,’ *Epidemiology & Infection*, vol. 136, no. 7, pp. 997–1007, 2008.

[161] P. Roberts *et al.*, ‘Identification of functional limitations and discharge destination in patients with covid-19,’ *Archives of physical medicine and rehabilitation*, vol. 102, no. 3, pp. 351–358, 2021.

[162] A. H. Rogers *et al.*, ‘Psychological factors associated with substance use initiation during the covid-19 pandemic,’ *Psychiatry Research*, vol. 293, p. 113 407, 2020.

[163] K. Rowa *et al.*, ‘Generalized anxiety disorder.,’ *Psychopathology*, 2008.

[164] D. B. Rubin, ‘Inference and missing data,’ *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[165] D. E. Rumelhart *et al.*, ‘Learning internal representations by error propagation,’ California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[166] J. A. Russell, ‘A circumplex model of affect.,’ *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[167] J. A. Russell, 'Core affect and the psychological construction of emotion.,' *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[168] T. Saito *et al.*, 'The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,' *PloS one*, vol. 10, no. 3, e0118432, 2015.

[169] A. Sano *et al.*, 'Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study,' *Journal of medical Internet research*, vol. 20, no. 6, e9410, 2018.

[170] D. L. Schacter, 'The seven sins of memory: Insights from psychology and cognitive neuroscience.,' *American psychologist*, vol. 54, no. 3, p. 182, 1999.

[171] D. O. Scharfstein *et al.*, 'Randomized trials in orthopaedic surgery: Advances and future directions: On the prevention and analysis of missing data in randomized clinical trials: The state of the art,' *The Journal of Bone and Joint Surgery. American volume*, vol. 94, no. Suppl 1, p. 80, 2012.

[172] K. R. Scherer, 'What are emotions? and how can they be measured?' *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[173] T. Schneider, 'Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values,' *Journal of climate*, vol. 14, no. 5, pp. 853–871, 2001.

[174] G. Schwarz, 'Estimating the dimension of a model,' *The annals of statistics*, pp. 461–464, 1978.

[175] J. Seppälä *et al.*, 'Mobile phone and wearable sensor-based mhealth approaches for psychiatric disorders and symptoms: Systematic review,' *JMIR mental health*, vol. 6, no. 2, e9819, 2019.

[176] S. Servia-Rodríguez *et al.*, 'Mobile sensing at the service of mental well-being: A large-scale longitudinal study,' in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 103–112.

[177] Z. Shen *et al.*, 'A novel time series forecasting model with deep learning,' *Neurocomputing*, vol. 396, pp. 302–313, 2020.

[178] S. Sinharay *et al.*, 'The use of multiple imputation for the analysis of missing data.,' *Psychological methods*, vol. 6, no. 4, p. 317, 2001.

[179] Q. Song *et al.*, 'A short note on safest default missingness mechanism assumptions,' *Empirical Software Engineering*, vol. 10, no. 2, pp. 235–243, 2005.

[180] M. Speekenbrink *et al.*, 'Ignorable and non-ignorable missing data in hidden markov models,' *arXiv preprint arXiv:2109.02770*, 2021.

[181] R. L. Spitzer *et al.*, 'A brief measure for assessing generalized anxiety disorder: The gad-7,' *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[182] A. A. Stone *et al.*, 'Ecological momentary assessment (ema) in behavorial medicine.,' *Annals of behavioral medicine*, 1994.

[183] Y. Suhara *et al.*, 'Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks,' in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 715–724.

[184] E. Sükei *et al.*, 'Assessing whodas 2.0 scores from behavioral biomarkers: A data-driven approach (preprint),' 2022.

[185] S. Taylor *et al.*, 'Personalized multitask learning for predicting tomorrow's mood, stress, and health,' *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2017.

[186] A. Teles *et al.*, 'Internet of things applied to mental health: Concepts, applications, and perspectives,' in *IoT and ICT for Healthcare Applications*, Springer, 2020, pp. 33–58.

[187] A. R. Teo *et al.*, 'The role of social isolation in social anxiety disorder: A systematic review and meta-analysis,' *Journal of Anxiety Disorders*, vol. 27, no. 4, pp. 353–364, 2013.

[188] C. Tian *et al.*, 'A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network,' *Energies*, vol. 11, no. 12, p. 3493, 2018.

[189] J. Torous *et al.*, 'Realizing the potential of mobile mental health: New methods for new data in psychiatry,' *Current psychiatry reports*, vol. 17, no. 8, pp. 1–7, 2015.

[190] R. Torres-Castro *et al.*, 'Functional limitations post-covid-19: A comprehensive assessment strategy,' *Archivos de bronconeumologia*, vol. 57, p. 7, 2021.

[191] G. J. Treisman *et al.*, 'Perspectives on the use of ehealth in the management of patients with schizophrenia.,' *Journal of Nervous and Mental Disease*, 2016.

[192] T. Umematsu *et al.*, 'Improving students' daily life stress forecasting using lstm neural networks,' in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, 2019, pp. 1–4.

[193] T. B. Üstün *et al.*, *Measuring health and disability: Manual for WHO disability assessment schedule WHODAS 2.0.* World Health Organization, 2010.

[194] S. Van Buuren *et al.*, 'Mice: Multivariate imputation by chained equations in r,' *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[195] T. F. Van de Mortel, 'Faking it: Social desirability response bias in self-report research,' *Australian Journal of Advanced Nursing, The*, vol. 25, no. 4, pp. 40–48, 2008.

[196] A. Vannucci *et al.*, 'Social media use and anxiety in emerging adults,' *Journal of affective disorders*, vol. 207, pp. 163–166, 2017.

[197] A. Vaswani *et al.*, ‘Attention is all you need,’ *Advances in neural information processing systems*, vol. 30, 2017.

[198] R. Wang *et al.*, ‘Studentlife: Using smartphones to assess mental health and academic performance of college students,’ in *Mobile health*, Springer, 2017, pp. 7–33.

[199] C. K. F. Wen *et al.*, ‘Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis,’ *Journal of medical Internet research*, vol. 19, no. 4, e6641, 2017.

[200] Q. Wen *et al.*, ‘Transformers in time series: A survey,’ *arXiv preprint arXiv:2202.07125*, 2022.

[201] S. R. Wisniewski *et al.*, ‘Prevention of missing data in clinical research studies,’ *Biological psychiatry*, vol. 59, no. 11, pp. 997–1000, 2006.

[202] C.-W. Woo *et al.*, ‘Building better biomarkers: Brain models in translational neuroimaging,’ *Nature neuroscience*, vol. 20, no. 3, pp. 365–377, 2017.

[203] M. Xin *et al.*, ‘Negative cognitive and psychological correlates of mandatory quarantine during the initial covid-19 outbreak in china.,’ *American Psychologist*, vol. 75, no. 5, p. 607, 2020.

[204] G. Xuan *et al.*, ‘Em algorithms of gaussian mixture model and hidden markov model,’ in *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, IEEE, vol. 1, 2001, pp. 145–148.

[205] W. Young *et al.*, ‘A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits,’ *Theoretical Issues in Ergonomics Science*, vol. 12, no. 1, pp. 15–43, 2011.

[206] K. F. Yuen *et al.*, ‘The psychological causes of panic buying following a health crisis,’ *International journal of environmental research and public health*, vol. 17, no. 10, p. 3513, 2020.

APPENDIX A

Model notations: LR/SVC/RFC/MLP - x = LR/SVC/RFC/MLP classifiers trained with input features formed of x-days of observations concatenated to create a single feature vector. RNN/LSTM/GRU - x = RNN/LSTM/GRU - RNNs with different cells using x-months long input sequences. Input feature notations: w/o posteriors = raw features used as classifier input; only posteriors = the MM component posterior probabilities used as classifier input features; w/ posteriors = raw features concatenated with the MM component posterior probabilities used as classifier input features. Model abbreviations: LR = logistic regression, SVC = support vector classifier, RFC = random forest classifier, MLP = multilayer perceptron, RNN = recurrent neural network, LSTM = long short term memory, GRU = gated reccurent unit.

Table A1: Classifier performance overview - Emotional valence case. Class labels: 0 = negative, 1 = neutral, 2 = positive emotional valence.

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix			
LR-1	w/o posteriors	0.38	0.52	0.35		0	1	2
					0	690	596	557
					1	201	177	176
					2	340	176	367
	w/ posteriors	0.37	0.52	0.35		0	1	2
					0	692	597	554
					1	201	175	178
					2	338	185	360
	only posteriors	0.46	0.56	0.36		0	1	2
					0	1038	0	805
					1	296	0	258
					2	406	0	477
LR-3	w/o posteriors	0.39	0.55	0.37		0	1	2
					0	622	687	500
					1	188	170	175
					2	238	170	450
	w/ posteriors	0.54	0.7	0.53		0	1	2
					0	1050	467	292
					1	173	196	164
					2	197	181	480
	only posteriors	0.59	0.73	0.57		0	1	2
					0	1235	297	277
					1	213	147	173
					2	234	131	493

Table A1 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
LR-7	w/o posteriors	0.38	0.56	0.37	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 592 & 699 & 458 \\ 1 & 185 & 149 & 168 \\ 2 & 233 & 155 & 430 \end{array}$
	w/ posteriors	0.56	0.73	0.58	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1074 & 403 & 272 \\ 1 & 158 & 193 & 151 \\ 2 & 180 & 176 & 462 \end{array}$
	only posteriors	0.6	0.76	0.62	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1226 & 275 & 248 \\ 1 & 192 & 159 & 151 \\ 2 & 211 & 149 & 458 \end{array}$
MLP-1	w/o posteriors	0.56	0.7	0.51	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1837 & 4 & 2 \\ 1 & 554 & 0 & 0 \\ 2 & 878 & 2 & 3 \end{array}$
	w/ posteriors	0.52	0.69	0.51	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1572 & 110 & 161 \\ 1 & 478 & 30 & 46 \\ 2 & 741 & 27 & 115 \end{array}$
	only posteriors	0.56	0.69	0.48	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1843 & 0 & 0 \\ 1 & 554 & 0 & 0 \\ 2 & 883 & 0 & 0 \end{array}$
MLP-3	w/o posteriors	0.56	0.71	0.54	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1474 & 188 & 147 \\ 1 & 414 & 43 & 76 \\ 2 & 557 & 38 & 263 \end{array}$
	w/ posteriors	0.63	0.77	0.63	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1515 & 102 & 192 \\ 1 & 333 & 82 & 118 \\ 2 & 403 & 51 & 404 \end{array}$
	only posteriors	0.64	0.79	0.66	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1595 & 33 & 181 \\ 1 & 358 & 44 & 131 \\ 2 & 439 & 14 & 405 \end{array}$
MLP-7	w/o posteriors	0.54	0.7	0.52	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1418 & 97 & 234 \\ 1 & 397 & 27 & 78 \\ 2 & 564 & 35 & 219 \end{array}$
	w/ posteriors	0.64	0.79	0.66	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1474 & 96 & 179 \\ 1 & 310 & 88 & 104 \\ 2 & 369 & 61 & 388 \end{array}$
	only posteriors	0.65	0.81	0.7	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1533 & 51 & 165 \\ 1 & 307 & 66 & 129 \\ 2 & 368 & 45 & 405 \end{array}$

Table A1 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
RFC-1	w/o posteriors	0.45	0.63	0.45	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1072 & 350 & 421 \\ 1 & 306 & 99 & 149 \\ 2 & 461 & 128 & 294 \end{array}$
	w/ posteriors	0.45	0.64	0.46	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1097 & 314 & 432 \\ 1 & 320 & 68 & 166 \\ 2 & 472 & 98 & 313 \end{array}$
	only posteriors	0.39	0.58	0.41	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 856 & 421 & 566 \\ 1 & 251 & 119 & 184 \\ 2 & 409 & 170 & 304 \end{array}$
RFC-3	w/o posteriors	0.56	0.73	0.58	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1240 & 257 & 312 \\ 1 & 287 & 85 & 161 \\ 2 & 307 & 70 & 481 \end{array}$
	w/ posteriors	0.62	0.78	0.67	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1355 & 160 & 294 \\ 1 & 246 & 101 & 186 \\ 2 & 259 & 67 & 532 \end{array}$
	only posteriors	0.6	0.77	0.64	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1258 & 213 & 338 \\ 1 & 190 & 137 & 206 \\ 2 & 221 & 113 & 524 \end{array}$
RFC-7	w/o posteriors	0.58	0.75	0.61	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1256 & 211 & 282 \\ 1 & 283 & 55 & 164 \\ 2 & 290 & 50 & 478 \end{array}$
	w/ posteriors	0.65	0.82	0.73	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1369 & 76 & 304 \\ 1 & 228 & 89 & 185 \\ 2 & 238 & 50 & 530 \end{array}$
	only posteriors	0.64	0.82	0.73	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1327 & 107 & 315 \\ 1 & 194 & 99 & 209 \\ 2 & 196 & 73 & 549 \end{array}$
SVC-1	w/o posteriors	0.56	0.7	0.51	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1837 & 4 & 2 \\ 1 & 554 & 0 & 0 \\ 2 & 878 & 2 & 3 \end{array}$
	w/ posteriors	0.39	0.7	0.53	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 775 & 563 & 505 \\ 1 & 221 & 149 & 184 \\ 2 & 378 & 156 & 349 \end{array}$
	only posteriors	0.24	0.71	0.53	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 0 & 1048 & 795 \\ 1 & 0 & 297 & 257 \\ 2 & 0 & 409 & 474 \end{array}$

Table A1 continued from previous page

<table>
<tr><th>Model</th><th>Input Features</th><th>Accuracy</th><th>AUC-ROC</th><th>AUC-PRC</th><th>Confusion Matrix</th></tr>
<tr><td>SVC-3</td><td>w/o posteriors</td><td>0.41</td><td>0.71</td><td>0.54</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 714 & 616 & 479 \\ 1 & 222 & 156 & 155 \\ 2 & 267 & 162 & 429 \end{array}$</td></tr>
<tr><td></td><td>w/ posteriors</td><td>0.59</td><td>0.78</td><td>0.65</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1271 & 286 & 252 \\ 1 & 231 & 157 & 145 \\ 2 & 276 & 132 & 450 \end{array}$</td></tr>
<tr><td></td><td>only posteriors</td><td>0.59</td><td>0.79</td><td>0.66</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1259 & 285 & 265 \\ 1 & 225 & 146 & 162 \\ 2 & 253 & 124 & 481 \end{array}$</td></tr>
<tr><td>SVC-7</td><td>w/o posteriors</td><td>0.4</td><td>0.72</td><td>0.55</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 679 & 634 & 436 \\ 1 & 198 & 126 & 178 \\ 2 & 237 & 150 & 431 \end{array}$</td></tr>
<tr><td></td><td>w/ posteriors</td><td>0.58</td><td>0.81</td><td>0.69</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1170 & 321 & 258 \\ 1 & 176 & 169 & 157 \\ 2 & 196 & 166 & 456 \end{array}$</td></tr>
<tr><td></td><td>only posteriors</td><td>0.61</td><td>0.81</td><td>0.69</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1274 & 239 & 236 \\ 1 & 203 & 144 & 155 \\ 2 & 235 & 122 & 461 \end{array}$</td></tr>
<tr><td>RNN-30</td><td>w/o posteriors</td><td>0.5</td><td>0.67</td><td>0.48</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1283 & 142 & 263 \\ 1 & 387 & 31 & 80 \\ 2 & 597 & 32 & 185 \end{array}$</td></tr>
<tr><td></td><td>w/ posteriors</td><td>0.53</td><td>0.69</td><td>0.51</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1468 & 57 & 163 \\ 1 & 432 & 12 & 54 \\ 2 & 687 & 10 & 117 \end{array}$</td></tr>
<tr><td></td><td>only posteriors</td><td>0.56</td><td>0.71</td><td>0.53</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1661 & 0 & 27 \\ 1 & 490 & 0 & 8 \\ 2 & 789 & 0 & 25 \end{array}$</td></tr>
<tr><td>RNN-91</td><td>w/o posteriors</td><td>0.5</td><td>0.67</td><td>0.47</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1411 & 150 & 188 \\ 1 & 449 & 31 & 54 \\ 2 & 684 & 25 & 132 \end{array}$</td></tr>
<tr><td></td><td>w/ posteriors</td><td>0.51</td><td>0.67</td><td>0.48</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1440 & 170 & 139 \\ 1 & 459 & 33 & 42 \\ 2 & 688 & 25 & 128 \end{array}$</td></tr>
<tr><td></td><td>only posteriors</td><td>0.56</td><td>0.71</td><td>0.53</td><td>$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1749 & 0 & 0 \\ 1 & 534 & 0 & 0 \\ 2 & 841 & 0 & 0 \end{array}$</td></tr>
</table>

Table A1 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
RNN-182	w/o posteriors	0.51	0.67	0.46	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1382 & 115 & 279 \\ 1 & 419 & 33 & 84 \\ 2 & 637 & 27 & 187 \end{array}$
	w/ posteriors	0.53	0.68	0.48	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1508 & 139 & 129 \\ 1 & 459 & 32 & 45 \\ 2 & 707 & 23 & 121 \end{array}$
	only posteriors	0.56	0.71	0.53	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1705 & 0 & 71 \\ 1 & 512 & 0 & 24 \\ 2 & 770 & 0 & 81 \end{array}$
LSTM-30	w/o posteriors	0.51	0.67	0.48	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1336 & 139 & 213 \\ 1 & 401 & 30 & 67 \\ 2 & 626 & 27 & 161 \end{array}$
	w/ posteriors	0.5	0.67	0.48	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1303 & 146 & 239 \\ 1 & 408 & 17 & 73 \\ 2 & 585 & 44 & 185 \end{array}$
	only posteriors	0.56	0.72	0.55	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1688 & 0 & 0 \\ 1 & 498 & 0 & 0 \\ 2 & 814 & 0 & 0 \end{array}$
LSTM-91	w/o posteriors	0.52	0.68	0.49	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1483 & 82 & 184 \\ 1 & 444 & 28 & 62 \\ 2 & 709 & 18 & 114 \end{array}$
	w/ posteriors	0.51	0.68	0.5	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1418 & 134 & 197 \\ 1 & 441 & 31 & 62 \\ 2 & 681 & 18 & 142 \end{array}$
	only posteriors	0.57	0.7	0.52	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1669 & 0 & 80 \\ 1 & 492 & 0 & 42 \\ 2 & 729 & 0 & 112 \end{array}$
LSTM-182	w/o posteriors	0.52	0.68	0.49	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1520 & 100 & 156 \\ 1 & 470 & 31 & 35 \\ 2 & 750 & 22 & 79 \end{array}$
	w/ posteriors	0.53	0.68	0.49	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1583 & 71 & 122 \\ 1 & 487 & 25 & 24 \\ 2 & 756 & 21 & 74 \end{array}$
	only posteriors	0.56	0.71	0.53	$\begin{array}{cccc} & 0 & 1 & 2 \\ \hline 0 & 1776 & 0 & 0 \\ 1 & 536 & 0 & 0 \\ 2 & 851 & 0 & 0 \end{array}$

Table A1 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
GRU-30	w/o posteriors	0.51	0.68	0.48	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1363 & 114 & 211 \\ 1 & 401 & 32 & 65 \\ 2 & 649 & 24 & 141 \end{array}$
	w/ posteriors	0.53	0.69	0.51	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1379 & 77 & 232 \\ 1 & 418 & 17 & 63 \\ 2 & 613 & 18 & 183 \end{array}$
	only posteriors	0.56	0.71	0.53	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1663 & 0 & 25 \\ 1 & 490 & 0 & 8 \\ 2 & 791 & 0 & 23 \end{array}$
GRU-91	w/o posteriors	0.5	0.67	0.47	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1414 & 144 & 191 \\ 1 & 441 & 35 & 58 \\ 2 & 689 & 36 & 116 \end{array}$
	w/ posteriors	0.56	0.7	0.52	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1742 & 3 & 4 \\ 1 & 531 & 1 & 2 \\ 2 & 826 & 0 & 15 \end{array}$
	only posteriors	0.56	0.71	0.53	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1743 & 1 & 5 \\ 1 & 533 & 0 & 1 \\ 2 & 840 & 0 & 1 \end{array}$
GRU-182	w/o posteriors	0.52	0.67	0.47	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1518 & 93 & 165 \\ 1 & 465 & 28 & 43 \\ 2 & 726 & 37 & 88 \end{array}$
	w/ posteriors	0.53	0.68	0.5	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1527 & 94 & 155 \\ 1 & 467 & 30 & 39 \\ 2 & 707 & 34 & 110 \end{array}$
	only posteriors	0.56	0.72	0.55	$\begin{array}{c\|ccc} & 0 & 1 & 2 \\ \hline 0 & 1774 & 0 & 2 \\ 1 & 535 & 0 & 1 \\ 2 & 851 & 0 & 0 \end{array}$

Table A2: Classifier performance overview - Emotional arousal-valence case. Class labels: 0 - neutral, 1 - high arousal - positive valence, 2 - high arousal - negative valence, 3 - low arousal - negative valence, 4 - low arousal - positive valence.

Model	**Input Features**	**Accuracy**	**AUC-ROC**	**AUC-PRC**	**Confusion Matrix**					
LR-1	w/o posteriors	0.27	0.52	0.22		0	1	2	3	4
					0	155	165	6	181	47
					1	124	271	5	206	47
					2	150	138	0	168	24
					3	431	397	2	438	95
					4	48	74	0	76	32
	w/ posteriors	0.27	0.52	0.22		0	1	2	3	4
					0	155	165	6	181	47
					1	124	271	5	206	47
					2	150	138	0	168	24
					3	431	397	2	438	95
					4	48	74	0	76	32
	only posteriors	0.16	0.54	0.21		0	1	2	3	4
					0	314	123	0	13	104
					1	308	139	0	26	180
					2	292	81	0	19	88
					3	798	260	0	41	264
					4	114	67	0	10	39
LR-3	w/o posteriors	0.27	0.53	0.22		0	1	2	3	4
					0	177	134	27	87	108
					1	111	240	22	134	123
					2	158	129	19	98	66
					3	422	337	43	349	188
					4	45	58	9	50	66
	w/ posteriors	0.29	0.56	0.24		0	1	2	3	4
					0	206	122	68	80	57
					1	129	230	50	140	81
					2	117	108	128	82	35
					3	338	334	218	311	138
					4	61	65	16	36	50
	only posteriors	0.34	0.63	0.29		0	1	2	3	4
					0	174	69	81	96	113
					1	99	142	63	154	172
					2	90	55	155	108	62
					3	225	132	283	540	159
					4	49	28	19	55	77
LR-7	w/o posteriors	0.26	0.53	0.22		0	1	2	3	4
					0	151	120	15	121	95
					1	99	237	19	148	92
					2	153	126	19	104	53
					3	427	343	50	321	153
					4	51	59	5	51	57

Table A2 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix					
	w/ posteriors	0.28	0.56	0.24		0	1	2	3	4
					0	183	98	55	78	88
					1	109	211	34	142	99
					2	115	85	133	66	56
					3	340	288	218	268	180
					4	52	59	16	40	56
	only posteriors	0.36	0.65	0.31		0	1	2	3	4
					0	169	78	72	78	105
					1	96	185	55	108	151
					2	82	54	163	102	54
					3	210	107	291	504	182
					4	46	27	16	45	89
MLP-1	w/o posteriors	0.37	0.69	0.34		0	1	2	3	4
					0	41	47	0	465	1
					1	23	81	0	549	0
					2	51	46	0	383	0
					3	156	101	0	1105	1
					4	20	14	0	196	0
	w/ posteriors	0.38	0.69	0.35		0	1	2	3	4
					0	29	85	0	440	0
					1	15	148	0	490	0
					2	37	75	0	368	0
					3	99	191	0	1073	0
					4	15	26	0	189	0
	only posteriors	0.42	0.69	0.35		0	1	2	3	4
					0	0	28	0	526	0
					1	0	56	0	597	0
					2	0	21	0	459	0
					3	0	45	0	1318	0
					4	0	17	0	213	0
MLP-3	w/o posteriors	0.36	0.68	0.33		0	1	2	3	4
					0	77	89	2	359	6
					1	41	169	9	402	9
					2	63	99	2	304	2
					3	211	216	6	887	19
					4	29	28	1	169	1
	w/ posteriors	0.4	0.72	0.4		0	1	2	3	4
					0	91	108	2	332	0
					1	54	200	2	372	2
					2	50	88	19	313	0
					3	148	192	14	984	1
					4	29	50	2	147	0
	only posteriors	0.44	0.73	0.42		0	1	2	3	4
					0	130	81	12	310	0
					1	69	159	12	390	0
					2	55	42	30	343	0
					3	134	93	38	1074	0
					4	39	35	0	154	0

Table A2 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
MLP-7	w/o posteriors	0.34	0.66	0.3	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 94 & 75 & 0 & 332 & 1 \\ 1 & 37 & 142 & 0 & 412 & 4 \\ 2 & 88 & 74 & 0 & 293 & 0 \\ 3 & 317 & 175 & 0 & 798 & 4 \\ 4 & 40 & 22 & 0 & 161 & 0 \end{array}$
	w/ posteriors	0.41	0.72	0.4	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 74 & 122 & 19 & 278 & 9 \\ 1 & 46 & 258 & 6 & 284 & 1 \\ 2 & 48 & 80 & 52 & 275 & 0 \\ 3 & 158 & 163 & 87 & 870 & 16 \\ 4 & 32 & 59 & 1 & 128 & 3 \end{array}$
	only posteriors	0.46	0.74	0.45	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 74 & 70 & 3 & 355 & 0 \\ 1 & 43 & 160 & 0 & 392 & 0 \\ 2 & 26 & 35 & 0 & 394 & 0 \\ 3 & 56 & 54 & 1 & 1183 & 0 \\ 4 & 12 & 27 & 0 & 184 & 0 \end{array}$
RFC-1	w/o posteriors	0.29	0.62	0.27	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 127 & 149 & 41 & 227 & 10 \\ 1 & 119 & 220 & 53 & 236 & 25 \\ 2 & 113 & 123 & 35 & 195 & 14 \\ 3 & 374 & 290 & 96 & 543 & 60 \\ 4 & 53 & 55 & 10 & 100 & 12 \end{array}$
	w/ posteriors	0.3	0.63	0.28	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 126 & 145 & 31 & 245 & 7 \\ 1 & 116 & 209 & 51 & 261 & 16 \\ 2 & 108 & 112 & 32 & 213 & 15 \\ 3 & 364 & 282 & 81 & 592 & 44 \\ 4 & 52 & 54 & 11 & 98 & 15 \end{array}$
	only posteriors	0.19	0.57	0.23	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 313 & 98 & 41 & 74 & 28 \\ 1 & 331 & 119 & 65 & 85 & 53 \\ 2 & 291 & 62 & 24 & 71 & 32 \\ 3 & 791 & 244 & 89 & 172 & 67 \\ 4 & 120 & 43 & 10 & 46 & 11 \end{array}$
RFC-3	w/o posteriors	0.33	0.66	0.3	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 156 & 154 & 43 & 158 & 22 \\ 1 & 73 & 269 & 35 & 224 & 29 \\ 2 & 89 & 126 & 58 & 185 & 12 \\ 3 & 304 & 327 & 110 & 554 & 44 \\ 4 & 48 & 82 & 7 & 74 & 17 \end{array}$
	w/ posteriors	0.44	0.74	0.44	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 126 & 133 & 50 & 140 & 84 \\ 1 & 72 & 394 & 33 & 86 & 45 \\ 2 & 61 & 92 & 86 & 214 & 17 \\ 3 & 169 & 212 & 146 & 735 & 77 \\ 4 & 37 & 55 & 17 & 52 & 67 \end{array}$

Table A2 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix					
RFC-7	only posteriors	0.41	0.71	0.4		0	1	2	3	4
					0	140	127	63	88	115
					1	90	351	52	47	90
					2	67	89	129	147	38
					3	179	200	242	600	118
					4	48	38	19	39	84
	w/o posteriors	0.35	0.67	0.33		0	1	2	3	4
					0	131	134	13	214	10
					1	58	228	8	288	13
					2	85	95	27	243	5
					3	305	250	46	666	27
					4	55	56	1	101	10
	w/ posteriors	0.48	0.77	0.5		0	1	2	3	4
					0	116	156	18	148	64
					1	56	402	17	95	25
					2	50	89	67	232	17
					3	128	186	91	820	69
					4	33	59	5	64	62
SVC-1	only posteriors	0.46	0.77	0.49		0	1	2	3	4
					0	107	143	38	125	89
					1	60	382	26	73	54
					2	42	91	90	209	23
					3	116	168	160	756	94
					4	28	46	6	56	87
	w/o posteriors	0.28	0.69	0.35		0	1	2	3	4
					0	179	102	7	251	15
					1	174	138	23	295	23
					2	177	76	6	212	9
					3	505	203	18	600	37
					4	66	38	10	106	10
	w/ posteriors	0.28	0.69	0.35		0	1	2	3	4
					0	175	98	20	238	23
					1	187	138	14	274	40
					2	176	74	6	205	19
					3	499	198	23	582	61
					4	68	39	6	105	12
SVC-3	only posteriors	0.14	0.69	0.35		0	1	2	3	4
					0	314	76	0	0	164
					1	308	55	0	0	290
					2	292	53	0	0	135
					3	798	168	0	0	397
					4	114	36	0	0	80
	w/o posteriors	0.28	0.69	0.35		0	1	2	3	4
					0	148	125	78	144	38
					1	96	212	62	194	66
					2	135	117	49	137	32
					3	384	297	127	448	83
					4	44	53	27	73	31

Table A2 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
	w/ posteriors	0.31	0.71	0.38	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 200 & 103 & 54 & 110 & 66 \\ 1 & 123 & 207 & 33 & 176 & 91 \\ 2 & 104 & 92 & 102 & 123 & 49 \\ 3 & 317 & 243 & 175 & 443 & 161 \\ 4 & 58 & 52 & 23 & 55 & 40 \end{array}$
	only posteriors	0.35	0.72	0.42	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 192 & 45 & 80 & 125 & 91 \\ 1 & 112 & 88 & 60 & 209 & 161 \\ 2 & 99 & 31 & 144 & 145 & 51 \\ 3 & 237 & 56 & 267 & 632 & 147 \\ 4 & 58 & 21 & 17 & 72 & 60 \end{array}$
SVC-7	w/o posteriors	0.27	0.68	0.34	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 156 & 95 & 45 & 149 & 57 \\ 1 & 111 & 187 & 50 & 187 & 60 \\ 2 & 161 & 93 & 47 & 124 & 30 \\ 3 & 428 & 264 & 105 & 411 & 86 \\ 4 & 60 & 47 & 23 & 64 & 29 \end{array}$
	w/ posteriors	0.3	0.71	0.39	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 176 & 112 & 39 & 100 & 75 \\ 1 & 112 & 202 & 22 & 148 & 111 \\ 2 & 103 & 86 & 97 & 106 & 63 \\ 3 & 280 & 246 & 165 & 392 & 211 \\ 4 & 62 & 53 & 11 & 55 & 42 \end{array}$
	only posteriors	0.36	0.75	0.45	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 151 & 102 & 71 & 78 & 100 \\ 1 & 83 & 195 & 54 & 131 & 132 \\ 2 & 78 & 51 & 158 & 111 & 57 \\ 3 & 189 & 129 & 287 & 503 & 186 \\ 4 & 37 & 35 & 14 & 52 & 85 \end{array}$
RNN-30	w/o posteriors	0.4	0.69	0.35	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 5 & 53 & 0 & 440 & 0 \\ 1 & 12 & 77 & 0 & 530 & 0 \\ 2 & 3 & 50 & 0 & 395 & 0 \\ 3 & 4 & 114 & 0 & 1122 & 0 \\ 4 & 1 & 17 & 0 & 177 & 0 \end{array}$
	w/ posteriors	0.36	0.67	0.31	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 50 & 93 & 0 & 355 & 0 \\ 1 & 11 & 177 & 0 & 431 & 0 \\ 2 & 68 & 75 & 0 & 305 & 0 \\ 3 & 200 & 195 & 0 & 840 & 5 \\ 4 & 28 & 33 & 0 & 131 & 3 \end{array}$
	only posteriors	0.41	0.69	0.36	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 0 & 0 & 498 & 0 \\ 1 & 0 & 0 & 0 & 619 & 0 \\ 2 & 0 & 0 & 0 & 448 & 0 \\ 3 & 0 & 0 & 0 & 1240 & 0 \\ 4 & 0 & 0 & 0 & 195 & 0 \end{array}$

Table A2 continued from previous page

Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix					
						0	1	2	3	4
RNN-91	w/o posteriors	0.35	0.67	0.31	0	35	87	0	397	15
					1	14	173	0	434	8
					2	86	66	0	311	2
					3	197	171	0	857	59
					4	14	29	0	156	13
						0	1	2	3	4
	w/ posteriors	0.36	0.68	0.32	0	38	85	0	403	8
					1	19	172	1	435	2
					2	60	64	0	340	1
					3	163	181	0	894	46
					4	14	24	0	164	10
						0	1	2	3	4
	only posteriors	0.42	0.7	0.35	0	5	22	0	507	0
					1	1	52	0	576	0
					2	3	10	0	452	0
					3	3	24	0	1257	0
					4	0	5	0	207	0
						0	1	2	3	4
RNN-182	w/o posteriors	0.36	0.67	0.32	0	43	42	0	451	0
					1	36	74	0	518	1
					2	33	48	0	385	1
					3	154	120	0	1035	0
					4	23	17	0	182	0
						0	1	2	3	4
	w/ posteriors	0.39	0.67	0.33	0	35	54	3	444	0
					1	17	114	8	489	1
					2	32	53	0	382	0
					3	111	124	0	1074	0
					4	17	16	0	189	0
						0	1	2	3	4
	only posteriors	0.42	0.69	0.35	0	0	22	0	514	0
					1	0	53	0	576	0
					2	0	10	0	457	0
					3	0	25	0	1284	0
					4	0	3	0	219	0
						0	1	2	3	4
LSTM-30	w/o posteriors	0.33	0.67	0.3	0	59	75	0	361	3
					1	55	136	0	426	2
					2	82	69	0	296	1
					3	257	184	0	780	19
					4	30	32	0	131	2
						0	1	2	3	4
	w/ posteriors	0.36	0.67	0.32	0	50	84	0	362	2
					1	39	154	0	424	2
					2	69	62	0	317	0
					3	211	162	0	861	6
					4	28	32	0	134	1

Table A2 continued from previous page

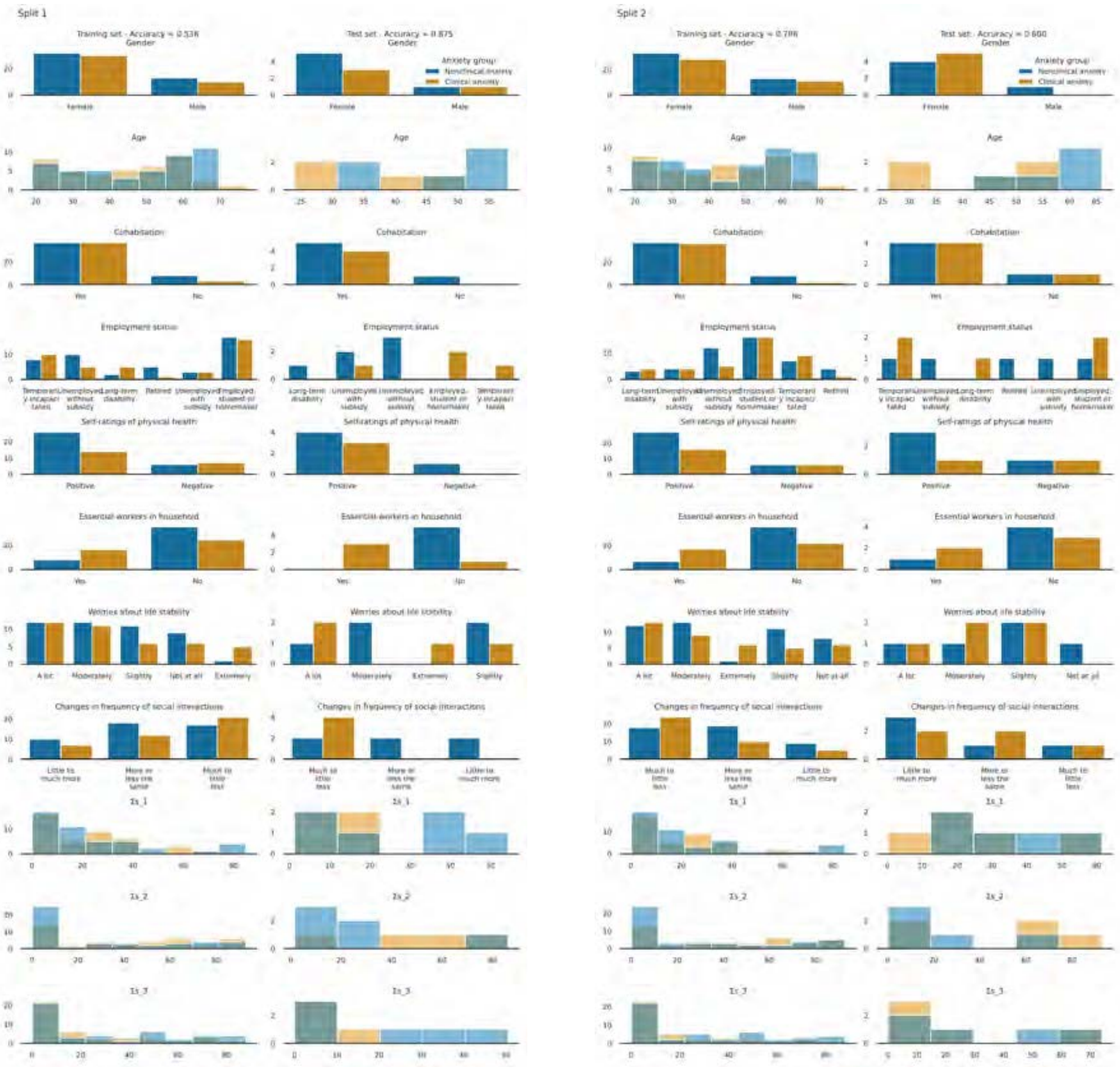
Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
LSTM-91	only posteriors	0.42	0.69	0.37	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 4 & 0 & 494 & 0 \\ 1 & 0 & 31 & 0 & 588 & 0 \\ 2 & 0 & 1 & 0 & 447 & 0 \\ 3 & 0 & 10 & 0 & 1230 & 0 \\ 4 & 0 & 1 & 0 & 194 & 0 \end{array}$
	w/o posteriors	0.33	0.68	0.32	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 50 & 67 & 0 & 415 & 2 \\ 1 & 38 & 126 & 0 & 459 & 6 \\ 2 & 71 & 52 & 0 & 341 & 1 \\ 3 & 242 & 171 & 0 & 868 & 3 \\ 4 & 29 & 27 & 0 & 156 & 0 \end{array}$
	w/ posteriors	0.33	0.67	0.32	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 62 & 80 & 0 & 389 & 3 \\ 1 & 51 & 174 & 0 & 400 & 4 \\ 2 & 85 & 67 & 0 & 312 & 1 \\ 3 & 265 & 217 & 0 & 790 & 12 \\ 4 & 28 & 27 & 0 & 156 & 1 \end{array}$
LSTM-182	only posteriors	0.41	0.69	0.35	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 5 & 0 & 529 & 0 \\ 1 & 0 & 9 & 0 & 620 & 0 \\ 2 & 0 & 0 & 0 & 465 & 0 \\ 3 & 0 & 2 & 0 & 1282 & 0 \\ 4 & 0 & 0 & 0 & 212 & 0 \end{array}$
	w/o posteriors	0.35	0.65	0.3	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 50 & 80 & 0 & 399 & 7 \\ 1 & 23 & 149 & 3 & 443 & 11 \\ 2 & 46 & 70 & 0 & 347 & 4 \\ 3 & 177 & 197 & 0 & 905 & 30 \\ 4 & 27 & 26 & 1 & 163 & 5 \end{array}$
	w/ posteriors	0.36	0.68	0.33	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 50 & 19 & 0 & 467 & 0 \\ 1 & 43 & 59 & 0 & 527 & 0 \\ 2 & 36 & 26 & 0 & 405 & 0 \\ 3 & 182 & 90 & 0 & 1037 & 0 \\ 4 & 28 & 6 & 0 & 188 & 0 \end{array}$
GRU-30	only posteriors	0.42	0.69	0.37	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 22 & 0 & 514 & 0 \\ 1 & 0 & 51 & 0 & 578 & 0 \\ 2 & 0 & 12 & 0 & 455 & 0 \\ 3 & 0 & 19 & 0 & 1290 & 0 \\ 4 & 0 & 8 & 0 & 214 & 0 \end{array}$
	w/o posteriors	0.34	0.67	0.31	$\begin{array}{c\|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 54 & 92 & 0 & 349 & 3 \\ 1 & 32 & 156 & 0 & 428 & 3 \\ 2 & 70 & 79 & 0 & 297 & 2 \\ 3 & 211 & 213 & 0 & 793 & 23 \\ 4 & 29 & 39 & 0 & 123 & 4 \end{array}$

Table A2 continued from previous page

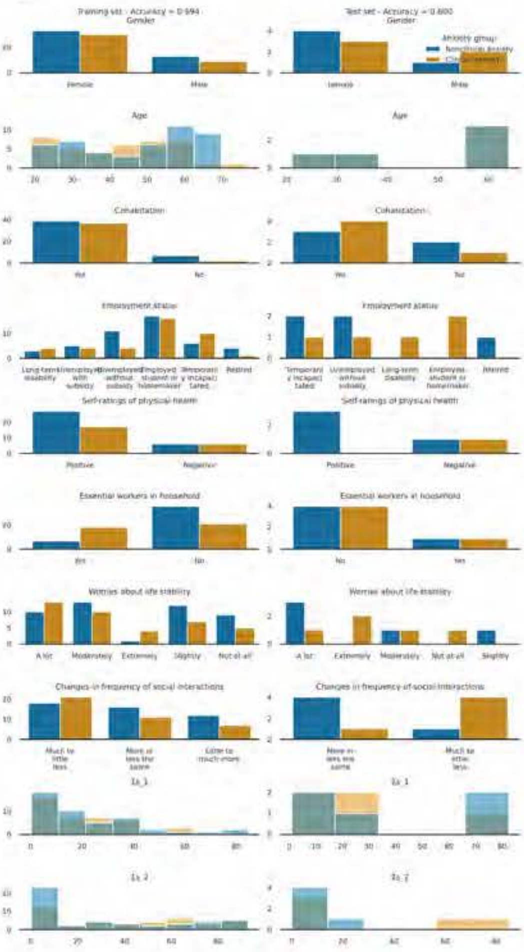
Model	Input Features	Accuracy	AUC-ROC	AUC-PRC	Confusion Matrix
	w/ posteriors	0.34	0.67	0.31	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 44 & 93 & 4 & 350 & 7 \\ 1 & 18 & 175 & 1 & 418 & 7 \\ 2 & 74 & 69 & 0 & 300 & 5 \\ 3 & 201 & 187 & 1 & 810 & 41 \\ 4 & 29 & 37 & 0 & 124 & 5 \end{array}$
	only posteriors	0.42	0.69	0.36	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 22 & 0 & 476 & 0 \\ 1 & 0 & 59 & 0 & 560 & 0 \\ 2 & 0 & 13 & 0 & 434 & 1 \\ 3 & 0 & 29 & 0 & 1211 & 0 \\ 4 & 0 & 8 & 0 & 187 & 0 \end{array}$
GRU-91	w/o posteriors	0.33	0.67	0.32	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 74 & 63 & 0 & 396 & 1 \\ 1 & 60 & 107 & 0 & 459 & 3 \\ 2 & 87 & 54 & 0 & 324 & 0 \\ 3 & 279 & 152 & 0 & 851 & 2 \\ 4 & 31 & 23 & 0 & 158 & 0 \end{array}$
	w/ posteriors	0.34	0.67	0.32	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 56 & 73 & 0 & 399 & 6 \\ 1 & 35 & 184 & 0 & 408 & 2 \\ 2 & 75 & 55 & 0 & 330 & 5 \\ 3 & 253 & 182 & 0 & 819 & 30 \\ 4 & 29 & 26 & 0 & 152 & 5 \end{array}$
	only posteriors	0.41	0.69	0.35	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 6 & 1 & 0 & 527 & 0 \\ 1 & 9 & 2 & 0 & 618 & 0 \\ 2 & 5 & 0 & 0 & 460 & 0 \\ 3 & 8 & 1 & 0 & 1275 & 0 \\ 4 & 1 & 0 & 0 & 211 & 0 \end{array}$
GRU-182	w/o posteriors	0.33	0.65	0.29	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 57 & 70 & 0 & 406 & 3 \\ 1 & 40 & 138 & 0 & 451 & 0 \\ 2 & 77 & 60 & 0 & 326 & 4 \\ 3 & 243 & 174 & 0 & 856 & 36 \\ 4 & 24 & 25 & 0 & 172 & 1 \end{array}$
	w/ posteriors	0.36	0.67	0.33	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 56 & 36 & 0 & 444 & 0 \\ 1 & 50 & 96 & 0 & 483 & 0 \\ 2 & 41 & 42 & 0 & 384 & 0 \\ 3 & 211 & 117 & 0 & 975 & 6 \\ 4 & 28 & 8 & 0 & 186 & 0 \end{array}$
	only posteriors	0.41	0.68	0.35	$\begin{array}{cccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 0 & 0 & 536 & 0 \\ 1 & 0 & 0 & 0 & 629 & 0 \\ 2 & 0 & 0 & 0 & 467 & 0 \\ 3 & 0 & 0 & 0 & 1309 & 0 \\ 4 & 0 & 0 & 0 & 222 & 0 \end{array}$

APPENDIX B

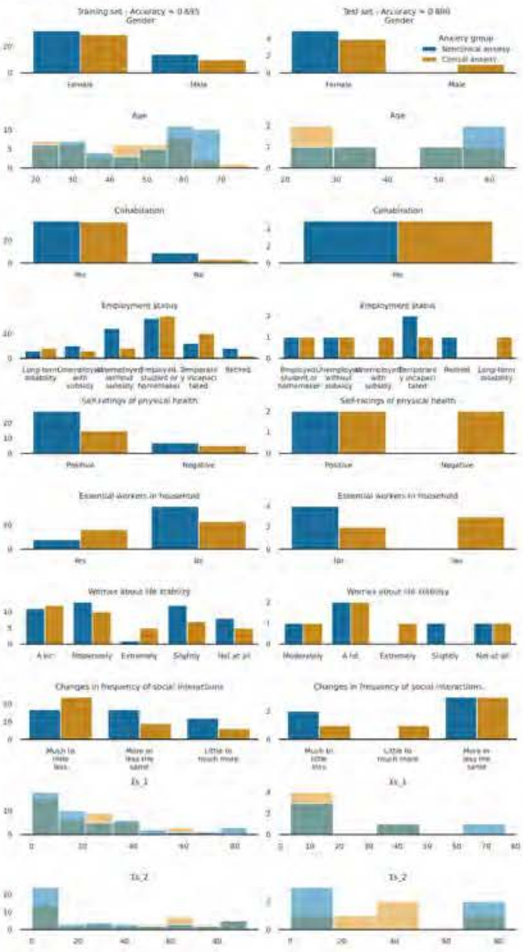
Figure A1: Data distribution in the different splits of the 10-fold cross-validation of our HMM+LR model.

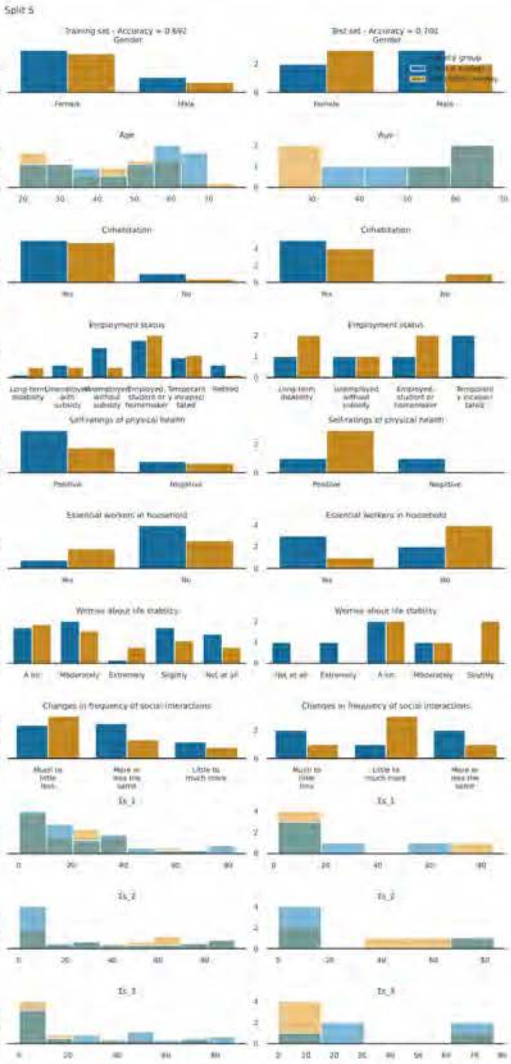


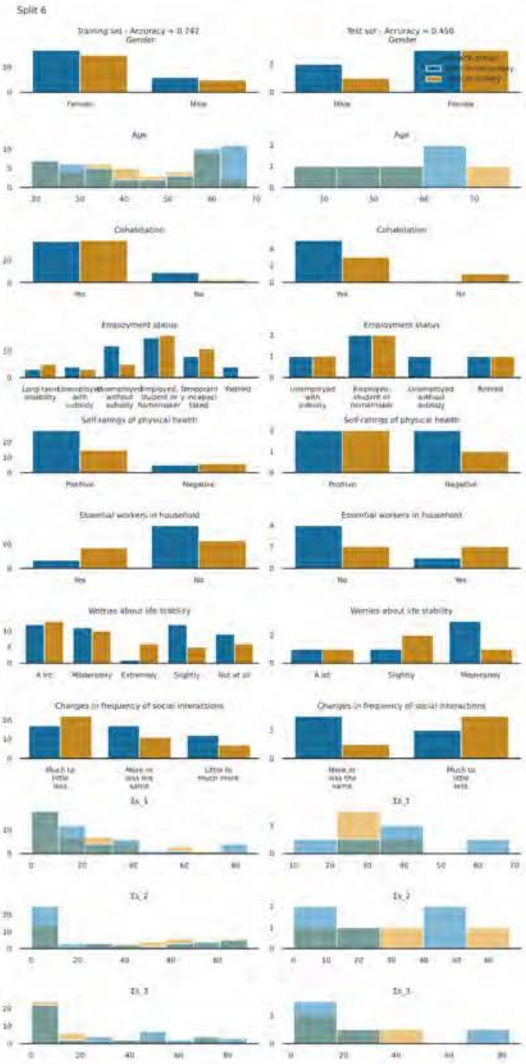
Split 3



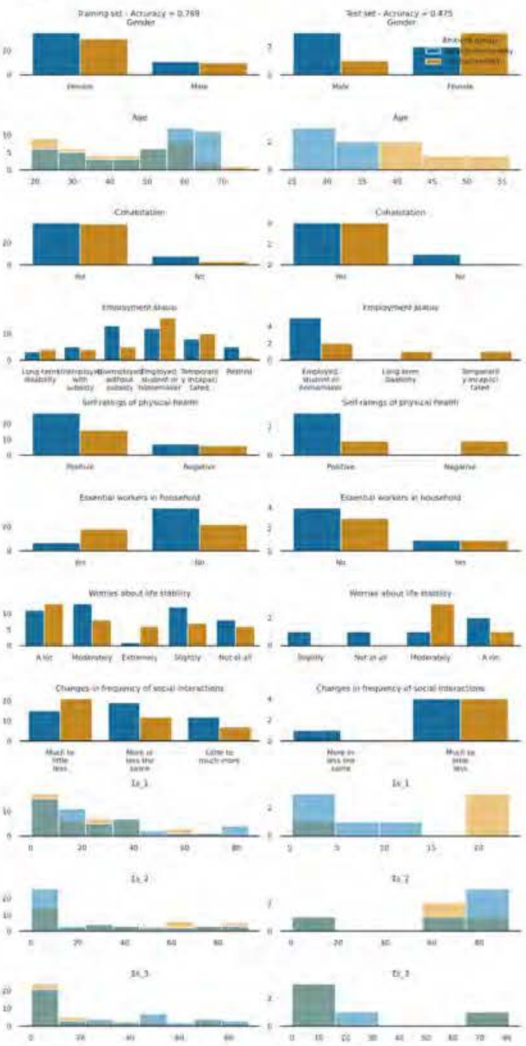
Split 4



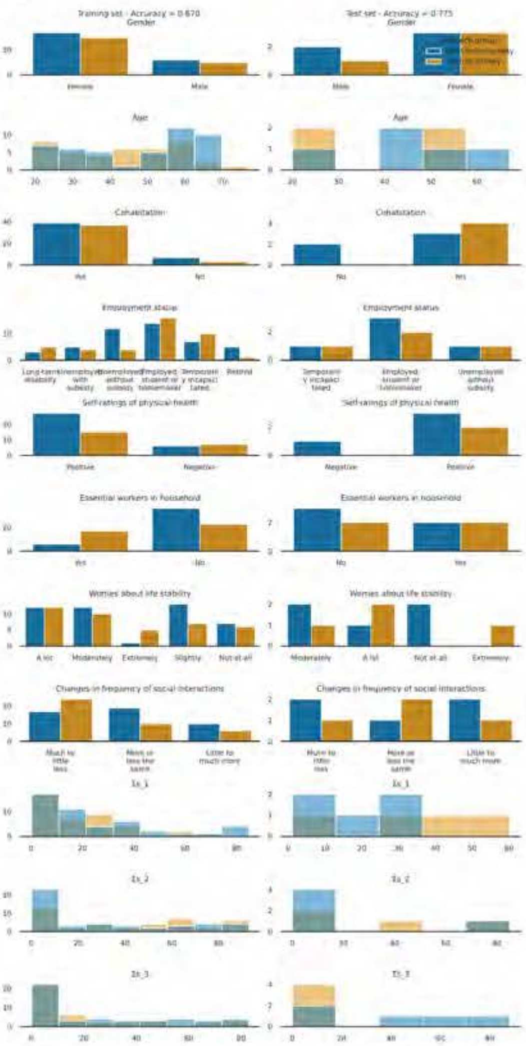

Split 5
Training set - Accuracy = 0.892
Test set - Accuracy = 0.701
Gender
Female
Male
Age
Cohabitation
Yes
No
Employment status
Long-term disability
Unemployed with subsidy
Unemployed without subsidy
Employed, student or homemaker
Temporarily incapacitated
Retired
Self-ratings of physical health
Positive
Negative
Essential workers in household
Worries about life stability
A lot
Moderately
Extremely
Slightly
Not at all
Changes in frequency of social interactions
Much to little less
More or less the same
Little to much more
Ex_1
Ex_2
Ex_3


Split 6
Training set - Accuracy = 0.742
Test set - Accuracy = 0.450
Gender
Female
Male
Age
Cohabitation
Yes
No
Employment status
Long-term disability
Unemployed with subsidy
Unemployed without subsidy
Employed, student or homemaker
Temporarily incapacitated
Retired
Self-ratings of physical health
Positive
Negative
Essential workers in household
Worries about life stability
A lot
Moderately
Extremely
Slightly
Not at all
Changes in frequency of social interactions
Much to little less
More or less the same
Little to much more
Ex_1
Ex_2
Ex_3

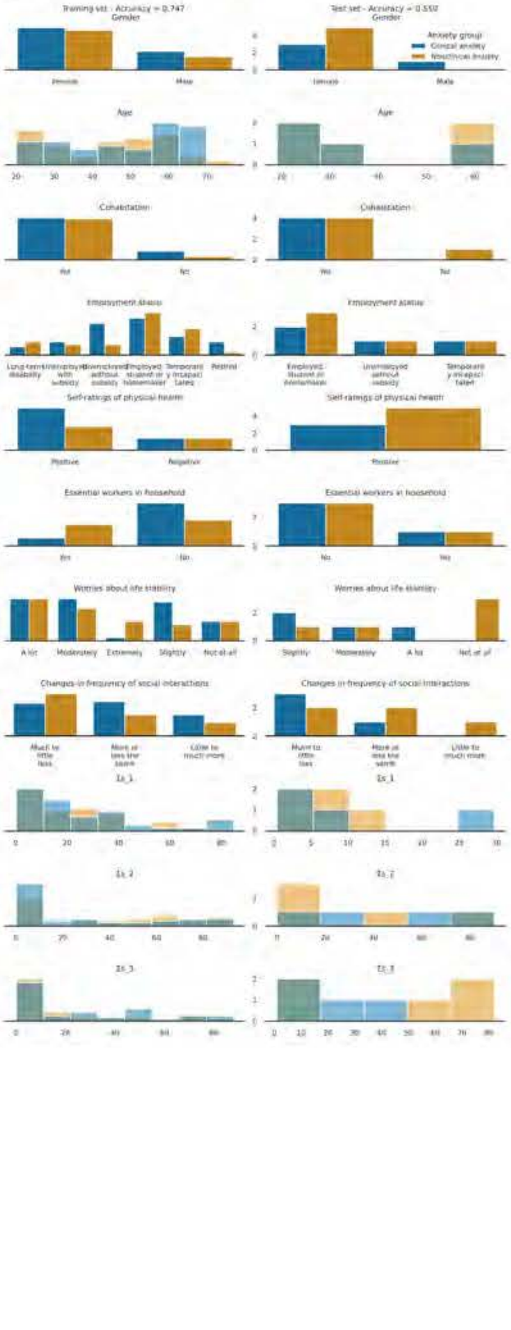
Split 7



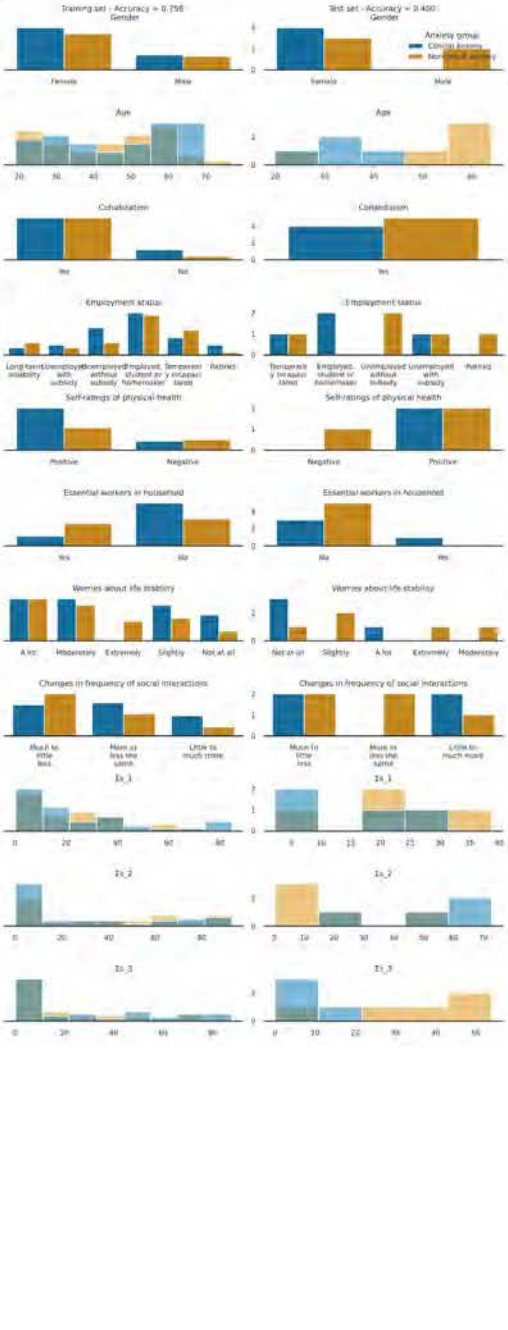
Split 8



Split 9



Split 10



APPENDIX C

Table A3: Lists of selected features per domain after applying sequential feature selection. Notations: SD = standard deviation, Q25/50/75 = 25th, 50th, 75th quantile.

WHODAS 2.0 domain	Number of features selected	Selected features
Cognition	19	Mean distance travelled
		Minimum distance travelled
		Number of time spent at home entries
		Q75 of time spent at home
		Maximum time spent at home
		Minimum step count
		Q25 of step count
		Q50 of step count
		Q75 of step count
		Number of exercise time entries
		Minimum time spent exercising
		Number of vehicle time entries
		SD of vehicle time
		Minimum vehicle time
		Number of walking time entries
		Q75 of time spent walking
		Q50 of sleep duration
		Q75 of sleep duration
		Maximum sleep duration
Mobility	19	Number of distance travelled entries
		Maximum distance travelled
		SD of number of visited locations
		Q75 of number of visited locations
		Minimum time spent at home
		Number of step count entries
		SD of step count
		Minimum step count
		Maximum step count
		Mean time spent exercising

Table A3 continued from previous page

WHODAS 2.0 domain	Number of features selected	Selected features
		Minimum time spent exercising
		Q50 of time spent exercising
		Number of vehicle time entries
		Mean vehicle time
		Minimum vehicle time
		Q25 of vehicle time
		Number of walking time entries
		Q25 of time spent walking
		Maximum time spent walking
Self-care	5	SD of distance travelled
		Q50 of distance travelled
		Minimum time spent at home
		Number of vehicle time entries
		Minimum vehicle time
Getting along	6	Mean distance travelled
		Minimum distance travelled
		Minimum time spent at home
		Q25 of time spent at home
		Minimum time spent exercising
		Maximum time spent exercising
Life activities	17	SD of distance travelled
		Minimum distance travelled
		Minimum number of visited locations
		Maximum number of visited locations
		Q50 of time spent at home
		Q75 of time spent at home
		Minimum time spent exercising
		Q25 of time spent exercising
		Number of vehicle time entries
		Mean vehicle time
		SD of vehicle time
		Minimum vehicle time
		Q75 of vehicle time
		Maximum vehicle time
		Number of walking time entries

Table A3 continued from previous page

WHODAS 2.0 domain	Number of features selected	Selected features
		Minimum time spent walking
		Q25 of time spent walking
Participation	13	SD of distance travelled
		Maximum number of visited locations
		Mean time spent at home
		Minimum step count
		Minimum time spent exercising
		Maximum time spent exercising
		Number of vehicle time entries
		Q50 of vehicle time
		Q75 of vehicle time
		Number of walking time entries
		Q75 of walking time
		Maximum time spent walking
		Number of sleep duration entries