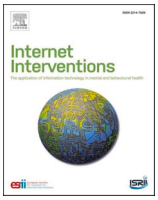
Internet Interventions 33 (2023) 100657

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.elsevier.com/locate/invent



Automatic patient functionality assessment from multimodal data using deep learning techniques – Development and feasibility evaluation

Emese Sükei [a,*], Santiago de Leon-Martinez [a,b,c], Pablo M. Olmos [a,d], Antonio Artés [a,e]

[a] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Av. de la Universidad 30, Leganés 28911, Madrid, Spain
[b] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[c] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[d] Gregorio Marañón Health Research Institute, Madrid 28009, Spain
[e] eB2 - Evidence-based Behavior, Leganés 28919, Spain

ARTICLE INFO

Keywords:
In-situ patient monitoring
Digital phenotyping
Ecological momentary assessment
Time-series modelling
Attention models
Transfer learning

ABSTRACT

Wearable devices and mobile sensors enable the real-time collection of an abundant source of physiological and behavioural data unobtrusively. Unlike traditional in-person evaluation or ecological momentary assessment (EMA) questionnaire-based approaches, these data sources open many possibilities in remote patient monitoring. However, defining robust models is challenging due to the data's noisy and frequently missing observations.

This work proposes an attention-based Long Short-Term Memory (LSTM) neural network-based pipeline for predicting mobility impairment based on WHODAS 2.0 evaluation from such digital biomarkers. Furthermore, we addressed the missing observation problem by utilising hidden Markov models and the possibility of including information from unlabelled samples via transfer learning. We validated our approach using two wearable/mobile sensor data sets collected in the wild and socio-demographic information about the patients.

Our results showed that in the WHODAS 2.0 mobility impairment prediction task, the proposed pipeline outperformed a prior baseline while additionally providing interpretability with attention heatmaps. Moreover, using a much smaller cohort via task transfer learning, the same model could learn to predict generalised anxiety severity accurately based on GAD-7 scores.

1. Introduction

1.1. Background

The ubiquity of mobile devices and their advanced sensors have allowed the passive collection of physiological and behavioural data in real-time, such as step count, location, sleep, and phone usage, which has opened new avenues for mental health research (Froehlich et al., 2007; Cornelius et al., 2008; Kukkonen et al., 2009). By monitoring physical activity, social interactions, and mental health biomarkers, mobile sensing enables continuous, unobtrusive evaluation of patients' functions and well-being, providing opportunities to improve care timeliness, treatment adherence, and health outcomes while minimising participant burden and recall bias (Noah et al., 2018).

To establish reliable frameworks for collecting, processing, and analysing behavioural sensor data, recent research has focused on monitoring and diagnosing data sources and developing robust models for accurate and effective predictions (Jara et al., 2013; Mohr et al., 2014; Sakr and Elgammal, 2016). Various health monitoring systems can be built based on mobile sensing and wearable data sources, such as human activity and posture monitoring, general well-being, fall monitoring in the elderly, and Parkinson's disease management, to name a few (Miranda et al., 2022). Mobile phone-based interventions have also been proposed as early-prevention approaches for reducing the disease burden associated with mental illness (Goldberg et al., 2022).

Such frameworks would be extremely valuable for passive follow-up of the evolution of patients' mobility impairment. Mobility impairment is a limitation in a person's ability to move and perform physical activities independently. It can be caused by various factors, such as ageing, chronic illnesses, and injuries. Mobility impairment can significantly impact an individual's daily life, leading to a decline in physical and mental health, reduced social interaction, and increased healthcare costs. Predicting mobility impairment using various approaches such as machine learning, statistical modelling, and clinical assessments can be

* Corresponding author.
E-mail address: esukei@ing.uc3m.es (E. Sükei).

https://doi.org/10.1016/j.invent.2023.100657
Received 13 November 2022; Received in revised form 24 July 2023; Accepted 7 August 2023
Available online 8 August 2023

found in the literature. However, the main focus group is usually the elderly (Demiris et al., 2004; Van Grootven et al., 2020). Moreover, the existing methods are obtained from data collected in a smart home environment.

Our prior work (Sükei et al., 2022) examined the feasibility of predicting patient functionality outcomes based on the World Health Organisation Disability Assessment Schedule (WHODAS 2.0), utilising passively collected digital biomarkers as model inputs to regression models. We performed feature engineering by extracting statistical measures (minimum, maximum, mean, median, standard deviation, IQR) from 30-day long time-series data sequences, followed by a simple linear regression for simplicity and interpretability of the biomarker features as predictors. Deep learning models have proven successful in many prediction tasks (LeCun et al., 2015), including mobile sensing (Servia-Rodríguez et al., 2017; Yao et al., 2017; Bahador et al., 2021), and could be employed for mental health outcome prediction.

1.2. Aims

Our objective is to develop a comprehensive deep-learning framework that predicts functional mobility impairment by integrating digital biomarkers and socio-demographic data. Due to the noisy and missing nature of mobile sensed data, we will employ appropriate time-series models and imputation techniques to capture underlying patterns in the data. Furthermore, including socio-demographic data can significantly enhance the model's performance by providing information strongly correlated to the individuals' functionality. Based on the shared behavioural patterns of patient groups, this data can effectively capture the complexity of human mobility.

Additionally, we suggest a simple task transfer learning approach that fine-tunes the model for predicting anxiety outcomes. Anxiety can significantly impact an individual's behavior, emotions, and thoughts, affecting their daily activities (Al-Lawati et al., 2000; Roemer and Orsillo, 2002). It can also lower motivation and prevent individuals from participating in physical activities (Otto and Smits, 2011; W. H. Organization, 2022). Therefore, we hypothesise that mobility descriptor variables can effectively predict anxiety outcomes.

Our contributions consist of the following:

- a hidden Markov model (HMM)-based method for handling the missing observations without resorting to classical imputation techniques,
- a deep neural network model that combines the self-attention method with long short-term memory (LSTM) networks to obtain patient embeddings over time,
- and a general multimodal pipeline to model questionnaire outcomes that can be trained using transfer learning techniques

2. Materials & methods

2.1. Data set

2.1.1. Study participants

The data used in this study were collected from two ongoing studies (Barrigón et al., 2017; Berrouiguet et al., 2018) involving passive smartphone monitoring of clinical outpatients. The studies received approval from the Institutional Review Board at the Psychiatry Department of Fundación Jimenez Diaz Hospital. All participants provided written informed consent.

2.1.2. Eligibility criteria

The study recruited patients who were at least 18 years old and were clinical outpatients diagnosed with mental disorders or attending therapy groups at the institutions mentioned above. To participate, patients had to own a smartphone with either Android or iOS operating systems, which they used to connect to a Wi-Fi network at least once a week. Only patients who provided written informed consent were included, and they did not receive payment for their participation.

2.1.3. Data sources

Passive data of participants was collected using eB2 Mindcare (eb2 evidence-based behaviour, n.d.; Bonilla-Escribano et al., 2019; Carretero et al., 2020), an eHealth platform validated for clinical use. The app runs in the background of the individual's smartphone, providing an e-diary for mental health. The data consisted of 48 half-hour daily summaries of four observations collected passively, namely step count, distance travelled, time spent at home, and exercise time. The data used in this work was collected from January 2016 to April 2022 from 2348 individuals, resulting in 516,604 entries, of which 31.5 % were collected in 2019. However, the final dataset had many missing observations, with overall missingness percentages of over 60 % for each variable (see Supplementary Figs. A.1 and A.2 for details). Missingness can be caused by varying sampling frequencies of sensors or sensor non-collection due to technological and behavioural factors, such as forgetting to charge the phone, disabling GPS, or uninstalling the study application.

On patient enrolment, *socio-demographic information*, such as age, gender, cohabitation status, and employment status, and initial completion of the *functionality evaluation* questionnaires and resulting scores were recorded using the *MeMind* tool (Barrigón et al., 2017; Muñoz-Navarro et al., 2017). The follow-up scores were recorded biannually at an in-person appointment or via a phone call (Fig. 1A). The two health outcomes that we focus on are the *World Health Organisation Disability Assessment Schedule 2.0 (WHODAS 2.0)* (McKibbin et al., 2004) and the *Generalised Anxiety Disorder Assessment 7-item (GAD-7)* (Spitzer et al., 2006) scores.

The WHODAS 2.0 questionnaire consists of 36 items that assess disability in 6 domains: cognition, mobility, self-care, getting along, life activities, and participation. Patients are asked to report difficulties experienced in the past 30 days while performing tasks in each domain and rate the difficulty level. Scores range from 0 to 100 %, with higher values indicating higher disability levels. Functional impairment categories can be defined at different cut-offs, as explained in (de Pedro-Cuesta et al., 2013). The GAD-7 screening tool is a 7-item questionnaire that assesses the severity of generalised anxiety disorder, with scores ranging from 0 to 21. The cut-off points for mild, moderate, and severe anxiety are 5, 10, and 15, respectively, and a score of 10 or greater indicates a need for further evaluation (Spitzer et al., 2006); therefore, we consider this cut-off value to dichotomise the GAD-7 outcome.

A subgroup of 2011 patients from the two studies had clinical evaluations for the outcomes of interest. Table 1 provides an overview of the distribution of socio-demographic information at baseline and the mental health outcome scores in the two study groups.

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.

2.2. Data preprocessing

2.2.1. Input data

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 (Fig. 1B). 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.

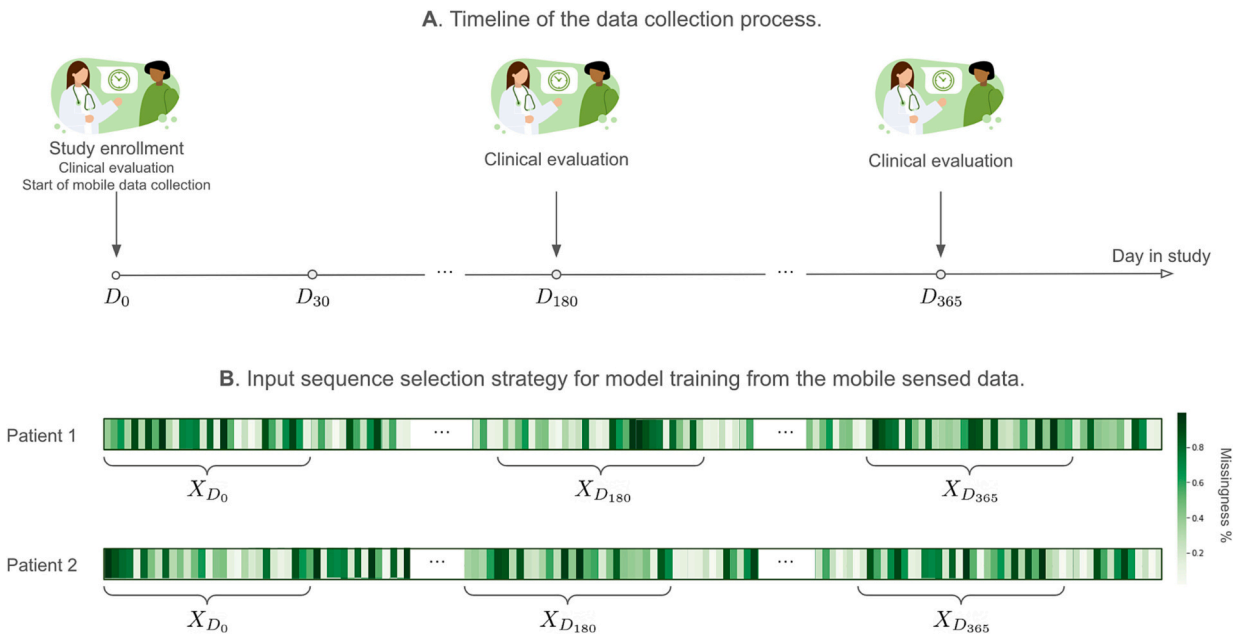


Fig. 1. Outline of the data collection process and input sequence selection. Notation: D_i - the i^{th} day in the study for a given patient, X_{Di} - the extracted 30-day mobile sensing data sequence around an evaluation.

2.2.2. Output data

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: for the WHODAS 2.0 mobility impairment outcome, in Study A, there are 1394/2233 0-labels, while in Study B, 165/283, and for the GAD-7 score 176/283, respectively (see the distributions in Supplementary Fig. A.3).

2.3. Proposed pipeline

This section introduces our proposed pipeline for leveraging the passively sensed time-series sequences and the socio-demographic data in functionality estimation. Fig. 2 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. The code used to produce the results presented in this paper is available at the following GitHub repository: https://github.com/mlfpm/functionality_prediction.

2.3.1. Dealing with missing data

Missing observations form a common problem with data collected from wearable devices and smartphones. They can occur due to varying degrees of compliance behavior, sensor failure, or non-collection (Kiang et al., 2021). Given the high percentage of missing data, imputation using statistical measures such as the mean, median, or even interpolation is insufficient, as they underestimate the variance and ignore the relationship between variables (Jäger et al., 2021). The imputed values would not generalise to wearable characteristics or participant behavior, can reduce variability in the data set, and introduce bias. Probabilistic generative models, such as hidden Markov models (HMMs) (Rabiner, 1989) 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 (Speekenbrink and Visser, 2021).

HMMs are commonly used generative models in time-series analysis, characterised by observable sequences and a set of discrete states, which are assumed to have been generated by a first-order Markov chain process. The learnable parameters of an HMM comprise an initial state probability distribution, a state transition probability distribution, and a symbol emission probability distribution. They can be trained unsupervised using the expectation-maximisation (EM) (Moon, 1996) algorithm and marginalisation to deal with the missing data.

Only those 48-slot daily patient sequences with at least 80 % of observations were considered for HMM training. It is important to note that the sequential data is fully leveraged for the HMM, and we do not just restrict the sequences to the 30-day windows. After this elimination process, 91,047 sequences were used to train the HMMs with different numbers of hidden states $n = \{2, 3, \ldots 23\}$. The best model was selected using the Bayesian and Akaike information criteria (Dridi and Hadzagic, 2018) on a randomly selected subset of 10,000 sequences with varying missingness. Given this model, we imputed the missing observations repeatedly during the mini-batch stochastic gradient descent. Every time a new batch of data was generated, the sequences were decoded using the Viterbi algorithm (Forney, 1973), and the missing observations were imputed by samples generated from the corresponding most probable state.

2.3.2. Predictive model

Our proposed pipeline is illustrated in Fig. 2. It performs feature encoding for the daily information by applying Time2Vec (Kazemi et al., 2019), followed by two LSTM (Hochreiter and Schmidhuber, 1997) encoders with self-attention (Vaswani et al., 2017) for the 30-day input sequence. A feed-forward layer on top of the second attention layer's outputs concatenated with 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 $t2v(\tau)$, is a vector of size $k+1$ defined as $t2v(\tau)[i] = \omega_i\tau + \phi_i$ if $i = 0$, and $t2v(\tau)[i] = F(\omega_i\tau + \phi_i)$ if $1 \leq i \leq k$. Here $t2v(\tau)[i]$ is the i^{th} element of $t2v(\tau)$, 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

Table 1
The study cohorts.

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

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, which within-day and within-month temporal patterns contribute to correct predictions for the model, is complicated due to the multitude of non-linear operations 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 significance 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 the worsening of a patient's state.

3. Experiments

The data from Study A was used for cross-validation and testing in all experiments except the one described in Section 3.5. We kept the data from Study B as a held-out test set. Therefore we refer to the results of models trained and evaluated on the same dataset as in-distribution and those tested on a different dataset as out-of-distribution (OOD). All models were trained using an Adam optimiser with a learning rate of $1e-3$ and batch size of 64.

We evaluated prediction performance using accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) scores (Saito and Rehmsmeier, 2015) 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 OOD data set, except for the experiment described in Section 3.5.

3.1. Defining the baseline

We started by re-using the pipeline defined in (Sükei et al., 2022) 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

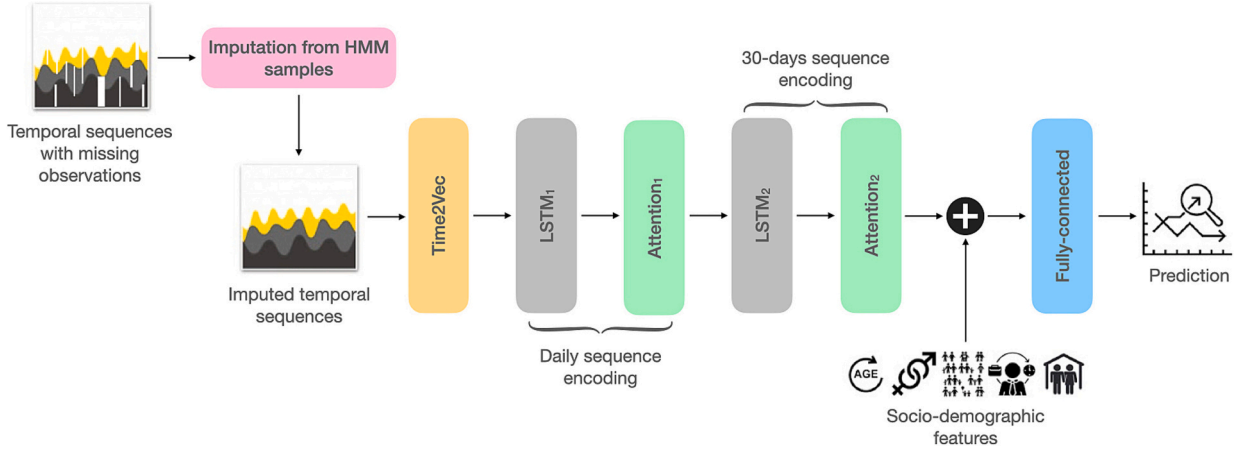


Fig. 2. The overall structure of the designed pipeline.

feature), which we imputed using simple mean imputation. These values occur because we do not discard sequences with a single measurement in the features to be able to compare the results directly.

3.2. Nested cross-validation

We first performed model hyper-parameter optimisation and model selection using a nested cross-validation approach (Krstajic et al., 2014). As such, a k-fold cross-validation procedure for model hyper-parameter optimisation is nested inside a k-fold cross-validation procedure for model evaluation and model selection for testing on the OOD dataset. 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 (Bergstra and Bengio, 2012), we try to discover a set of hyper-parameters that perform well on the data set in the sense of the AUPRC score.

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 2). Nested cross-validation with $k = 5$ folds in the outer loop would fit and evaluate 250 models. We trained the models for 50 epochs. The final model is configured by applying the outer loop to the entire data set. This model trained on the complete in-distribution data is then used to predict on the unseen OOD data.

3.3. Ablation study

When creating a complex machine learning model, it is helpful to understand the impact of each component separately (Meyes et al., 2019). 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 seen 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 Study A data set for the found number of epochs and evaluated on the held-out test set (Study B) in each case.

3.4. 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 pre-training 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 adapt better to the specific 30-day periods around the evaluation.

Table 2
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	$\{x + 8 \mid x \in N \cap [32, 128]\}$
Hidden dimension - Block 2	$\{x + 8 \mid x \in N \cap [64, 256]\}$
Bidirectional	{True, False}
Number of layers	{1, 2, 3}
Dropout	{0.1, 0.2, 0.3}

3.5. Task transfer learning

The core symptom of general anxiety disorder is chronic, excessive, and uncontrolled worry (Rowa and Antony, 2008), which is reflected in individuals' behavioural patterns. Previous studies have shown the positive impact of regular activity on patients' anxiety-related outcomes (Anderson and Shivakumar, 2013; Aylett et al., 2018). 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 smaller labelled sample size, which makes it difficult for such complex models to learn to generalise well to unseen OOD 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 (Olivas et al., 2009).

4. Results

4.1. 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, as described in 2.4.1.

The hyper-parameter tuning 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

4.2. Baseline, ablation and temporal pre-training

In Table 3, 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 in-distribution AUROC score but achieved slightly worse performance in AUPRC scores. On the OOD test set, the DL model outperformed the baseline in AUPRC.

We will now examine the results of the ablation study in reference to the entire DL pipeline. Removing the attention layer but keeping the Time2Vec layer led to a significant performance decrease on the in-distribution test, with an increase only in AUROC in the OOD test. Removing the Time2Vec block led to lower AUROC and slightly higher AUPRC for the in-distribution test, while in OOD test provided the best performance. The model without the Time2Vec and self-attention layer provided the best AUROC of the ablation variants and the best overall AUPRC in-distribution; however, while OOD did improve upon the entire model, it performed worse than the former model. Ablation, in general, led to lower AUROC in-distribution, but this trade-off could be desirable, particularly in the case of the removal of Time2Vec. Removing Time2Vec from the pipeline allowed for a more stable model with lower in-distribution standard deviation and relevant improvements in all areas except the formerly mentioned.

Pre-training the temporal encoder block of the entire model using all the available data sequences and then freezing led to a slight decrease in the model performances except in-distribution AUPRC. With the fine-

Table 3
Model performance comparison for the binary WHODAS 2.0 mobility impairment prediction task trained on study A. The highest achieved performance is indicated with boldin each setup. Notation: SD = standard deviation. AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve.

Experiment	Model	In-distribution test - 3-fold mean (SD)		Out-of-distribution test - all held-out	
		AUROC	AUPRC	AUROC	AUPRC
Baseline	Random	0.500 (0.000)	0.369 (0.000)	0.500	0.411
	SFS + LR (Sükei et al., 2022)	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
Encoder pre-training	DL pipeline (no pre-train)	**0.693 (0.023)**	0.532 (0.051)	**0.586**	**0.542**
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

tuning approach, the in-distribution performance increased to 0.558 AUPRC, as opposed to the 0.532 AUPRC achieved without pre-training. In contrast, the model only reaches 0.545 average AUPRC after training with the feature extraction approach. Nonetheless, the feature extraction approach reaches a slightly higher AUPRC on the OOD test.

As Fig. 3 shows, more attention is paid on average to the activity in the evening hours (slots 36–47), very low attention weights are associated with late-night activity, and varying patterns can be seen during the day in both cohorts. As for the monthly sequences, the attention weights are fairly uniform over the 30-day interval in both groups, with occasionally more attention being assigned to the last days of the period.

When analysing the attention weights at the patient level (Fig. 4), 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 late 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.

4.3. *Task transfer learning*

Table 4 shows the dichotomised GAD-7 classification performance scores with and without transfer learning, respectively. As expected, the 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, and the

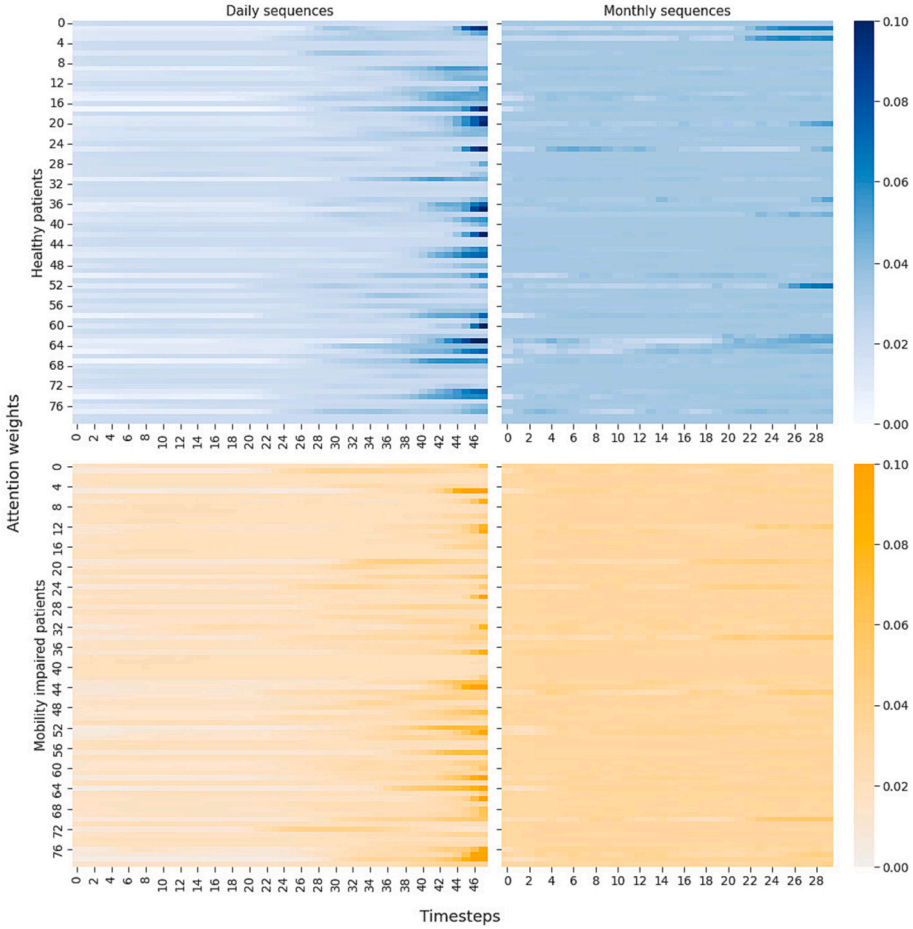


Fig. 3. Daily and monthly average attention weights for 80 randomly selected patients from each mobility difficulty.

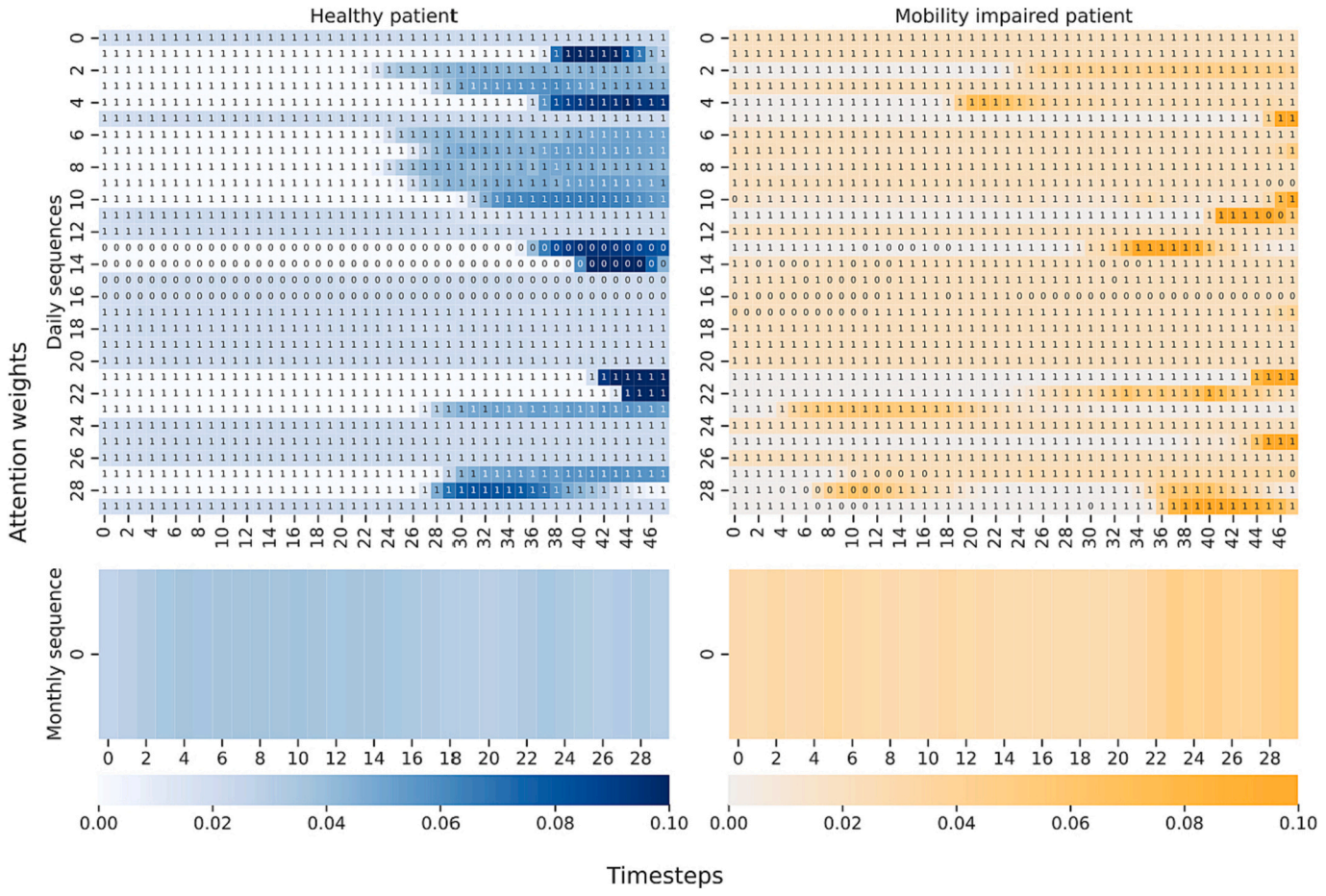


Fig. 4. Daily and monthly attention weights for 2 randomly selected patients with different mobility difficulty levels. We indicate with 1 the presence of a sample, while with 0, its missingness.

Table 4
Model performance comparison for the binary GAD-7 anxiety prediction task trained on study B, with and without transfer learning. The highest achieved performance is indicated with boldin each setup. Notation: SD = standard deviation. AUROC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve.

Experiment	Model	In-distribution test score 5-fold mean (SD)	
		AUROC	AUPRC
Baseline	Random	0.500 (0.000)	0.392 (0.011)
	SFS + LR (Sükei et al., 2022)	0.505 (0.077)	0.450 (0.068)
	DL pipeline	**0.518 (0.143)**	**0.463 (0.148)**
Transfer & pre-training	DL pipeline (no pre-train, no transfer)	0.518 (0.143)	0.463 (0.148)
Feature extraction approach		0.530 (0.107)	0.504 (0.148)
Fine-tuning approach		**0.603 (0.121)**	**0.556 (0.148)**

variance between splits is relatively large.

For transfer learning, weights were copied from the trained WHODAS DL pipeline without pre-training, as encoder pre-training was not helpful for performance, then both transfer learning approaches were applied. Overall, the transfer learning was successful as both approaches improved the baseline. A significant performance increase was reached when we fine-tuned the model trained on the mobility impairment classification task. This process works because the features are suitable for both base and target tasks, and the significantly larger dataset A covers a broader range of covariates. Hence the starting weights of the network are more representative. It could be argued that there is insufficient data to train a sophisticated model from scratch. However, even with a small amount of data, this baseline pipeline still outperforms the simpler sequential feature selection and logistic regression. This comes at the cost of higher variance, which is also improved upon in the transfer learning approaches. The pre-trained model is leveraging the additional information for improvements, which confirms the relationship between the tasks of WHODAS 2.0 prediction and GAD-7 prediction and between the datasets of study A and B. It can be said that the overall model performance still leaves room for improvement, but it is important to recognise the difficulty of the problem of GAD-7 prediction with such a small dataset.

5. Discussion

5.1. Principal results

This work tackled common problems in modelling mental health outcomes from passively sensed digital biomarkers. In this case, the data was passively collected with mobile phones, which have inherent limitations in sensor technology (they are not designed for research quality), leading to noisy and missing data. However, they are the ideal device for data collection, as most people carry a mobile phone throughout their day. Thus, if these limitations can be overcome, it provides a promising way to passively collect data without any subject interaction, which is better for retention and provides more ecological data (i.e.

representative of real life). We present a method to overcome this limitation by gathering large amounts of data (data mining) and then applying machine learning techniques to fill in missing values and capture the relevant information for predicting the desired task, in this case, psychiatric target prediction.

One of the main difficulties we faced was dealing with the large amounts of missing data meaningfully. We used HMMs trained on the 48-half-hour time slot sequences describing patients' daily activities, which were then used for imputation. This allowed for marginalising the missingness by repeated imputations from the distributions best describing each feature in the sense of likelihood.

Deep learning techniques automatically learned 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 proved 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/transfer learning step for the temporal encoder block of the model helped with a more meaningful initialisation of the model weights for the classification task of GAD7. However, it was unhelpful in the case of predicting WHODAS 2.0 mobility impairment. Pre-training can be viewed as beginning gradient descent from a different initialisation on the objective space, and thus improvements are due to reaching better optima that minimise the loss. In the case of the WHODAS 2.0 mobility scores, it was clear that this initialisation led to a finding of poorer optima, while in GAD-7 prediction, it led to significantly better optima. Finding a worse optimum is primarily due to either being stuck in a worse local optimum or descending the objective space slower. The pre-training initialisation is based on predicting the average next-day mobility markers, and this can be generally helpful when there is a lack of data. However, with more data, the optimisation problem solved in the predictive task strays further and further away from the one defined in the pre-training. Moreover, as the data set from study A 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 B.

The self-attention heatmaps showed different general within-day and within-month patterns emerging in the healthy versus mobility-impaired cohorts. Additionally, when paired with a relevant time-point, the heatmaps can be used to analyse the different emerging patterns concerning certain events, for example, the beginning or discontinuation of a prescribed treatment or comparing patterns between visits. These simple visualisations provide a helpful tool for clinicians to gain insights into the individuals' activity patterns and what led to an improvement, and even more importantly, what led to a decline. This can provide invaluable information when deciding to trigger proper interventions to help slow down or reverse the decline.

5.2. Strengths and limitations

The strength of this work lies in the novelty of a pipeline for passively determining patient functionality scores. We focus on psychiatric targets, but this idea could be extended to many areas. Not only is the idea and implementation novel to the domain of passive mobile psychiatric evaluation but also this work presents a way to deal with the limitations of passively sensed mobile data in any domain to achieve reasonable results. Despite high percentages of missing values and low variability in labels, the model captured patient outcomes.

This study shows that it is feasible to set up machine learning pipelines for passive patient evaluation, and there are ways to address problems in mobile sensed data. With the work presented here and future improvements, the limitations of passive data collection from mobile phones can be mitigated, leaving researchers and clinicians with a data collection tool that is almost always carried by subjects, collects data in real-time, requires no patient interaction, and can gather multiple data per day across modalities.

Although our approach showed promising results, it faces additional challenges and leaves room for improvement in future work. A common limitation in health applications is the relatively small labelled sample size. A larger patient cohort and multiple different valued 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. Secondly, to help address the missingness problem, we introduce another limitation in finding the least missing containing a 30-day window for each label, which can include days after questionnaire completion. However, it is not an unreasonable assumption that the patient's state would not change within a few days or weeks of questionnaire completion. This assumption is further supported by the data's lack of within-individual label differences.

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.

6. 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. This pipeline results in better 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.

Funding

ES received funding from the European Union Horizon 2020 research and innovation program (Marie Sklodowska-Curie grant 813533). AA and PO are supported by the Spanish government MCI under grants PID2021-123182OB-I00 and PID2021-125159NB-I00, by Comunidad de Madrid under grant IND2022/TIC-23550, by the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program under Grant 714161, and by Comunidad de Madrid and FEDER through IntCARE-CM.

Ethics approval and consent to participate

All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The Ethics Committee of the University Hospital Fundación Jiménez Díaz approved all procedures involving human patients. All participants provided written informed consent to participate in the study.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Antonio Artés is the co-founder of Evidence-Based Behavior (eB2).

Acknowledgements

The authors would like to thank Dr. Enrique Baca-Garcia for providing expert advice on clinical and patient-related matters. We would also like to thank all participants and therapists for their valuable contributions to the studies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.invent.2023.100657.

References

Al-Lawati, J., Al-Lawati, N., Al-Siddiqui, M., Antony, S.X., Al-Naamani, A., Martin, R., Kolbe, R., Theodorsson, T., Osman, Y., Al-Hussaini, A., et al., 2000. Psychological morbidity in primary health care in Oman: a preliminary study. Sultan Qaboos Univ. Med. J. 2 (2), 105–110.

Anderson, E., Shivakumar, G., 2013. Effects of exercise and physical activity on anxiety. Front. Psychiatry 4, 27.

Aylett, E., Small, N., Bower, P., 2018. Exercise in the treatment of clinical anxiety in general practice–a systematic review and meta-analysis. BMC Health Serv. Res. 18 (1), 1–18.

Bahador, N., Ferreira, D., Tamminen, S., Kortelainen, J., et al., 2021. Deep learning–based multimodal data fusion: case study in food intake episodes detection using wearable sensors. JMIR mHealth uHealth 9 (1), e21926.

Barrigón, M.L., Berrouiguet, S., Carballo, J.J., Bonal-Giménez, C., Fernández-Navarro, P., Pfang, B., Delgado-Gómez, D., Courtet, P., Aroca, F., Lopez-Castroman, J., et al., 2017. User profiles of an electronic mental health tool for ecological momentary assessment: Memind. Int. J. Methods Psychiatr. Res. 26 (1), e1554.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimisation. J. Mach. Learn. Res. 13 (2).

Berrouiguet, S., Ramírez, D., Barrigón, M.L., Moreno-Muñoz, P., Camacho, R.C., Baca-García, E., Artés-Rodríguez, A., et al., 2018. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioural changes detection: a case series of the evidence-based behaviour (eb2) study. JMIR mHealth uHealth 6 (12), e9472.

Bonilla-Escribano, P., Ramirez, D., Sedano-Capdevila, A., Campaña-Montes, J.J., Baca-Garcia, E., Courtet, P., Artes-Rodriguez, A., 2019. Assessment of e-social activity in psychiatric patients. IEEE J. Biomed. Health Inform. 23 (6), 2247–2256.

Carretero, P., Campana-Montes, J.J., Artes-Rodriguez, A., 2020. Ecological momentary assessment for monitoring the risk of suicide behaviour. In: Behavioral Neurobiology of Suicide and Self Harm. Springer, pp. 229–245.

Cornelius, C., Kapadia, A., Kotz, D., Peebles, D., Shin, M., Triandopoulos, N., 2008. Anonysense: privacy-aware people-centric sensing. In: Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services, pp. 211–224.

de Pedro-Cuesta, J., Garcia-Sagredo, P., Alcalde-Cabero, E., Alberquilla, A., Damián, J., Bosca, G., López-Rodríguez, F., Carmona, M., de Tena-Dávila, M.J., García-Olmos, L., et al., 2013. Disability transitions after 30 months in three community-dwelling diagnostic groups in Spain. PLoS One 8 (10), e77482.

Demiris, G., Rantz, M.J., Aud, M.A., Marek, K.D., Tyrer, H.W., Skubic, M., Hussam, A.A., 2004. Older adults' attitudes towards and perceptions of 'smart home' technologies: a pilot study. Med. Inform. Internet Med. 29 (2), 87–94.

Dridi, N., Hadzagic, M., 2018. Akaike and Bayesian information criteria for hidden Markov models. IEEE Signal Process Lett. 26 (2), 302–306.

eb2 evidence-based behaviour. https://eb2.tech/?lang=en.

Forney, G.D., 1973. The viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A., 2007. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, pp. 57–70.

Goldberg, S.B., Lam, S.U., Simonsson, O., Torous, J., Sun, S., 2022. Mobile phone-based interventions for mental health: a systematic meta-review of 14 meta-analyses of randomised controlled trials. PLOS Digit. Health 1 (1), e0000002.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Jäger, S., Allhorn, A., Bießmann, F., 2021. A benchmark for data imputation methods. Front. Big Data 4, 693674.

Jara, A.J., Zamora-Izquierdo, M.A., Skarmeta, A.F., 2013. Interconnection framework for mHealth and remote monitoring based on the internet of things. IEEE J. Sel. Areas Commun. 31 (9), 47–65.

Kazemi, S.M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., Brubaker, M., 2019. Time2vec: Learning a Vector Representation of Time (arXiv preprint arXiv:1907.05321).

Kiang, M.V., Chen, J.T., Krieger, N., Buckee, C.O., Alexander, M.J., Baker, J.T., Buckner, R.L., Coombs, G., Rich-Edwards, J.W., Carlson, K.W., et al., 2021. Sociodemographic characteristics of missing data in digital phenotyping. Sci. Rep. 11 (1), 1–11.

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminformatics 6 (1), 1–15.

Kukkonen, J., Lagerspetz, E., Nurmi, P., Andersson, M., 2009. Betelgeuse: a platform for gathering and processing situational data. IEEE Pervasive Comput. 8 (2), 49–56.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

McKibbin, C., Patterson, T.L., Jeste, D.V., 2004. Assessing disability in older patients with schizophrenia: results from the WHODAS 2.0. J. Nerv. Ment. Dis. 192 (6), 405–413.

Meyes, R., Lu, M., de Puiseau, C.W., Meisen, T., 2019. Ablation Studies in Artificial Neural Networks (arXiv preprint arXiv:1901.08644).

Miranda, L., Viterbo, J., Bernardini, F., 2022. A survey on the use of machine learning methods in context-aware middlewares for human activity recognition. Artif. Intell. Rev. 55 (4), 3369–3400.

Mohr, D.C., Schueller, S.M., Montague, E., Burns, M.N., Rashidi, P., 2014. The behavioural intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J. Med. Internet Res. 16 (6), e3077.

Moon, T.K., 1996. The expectation-maximization algorithm. IEEE Signal Process. Mag. 13 (6), 47–60.

Muñoz-Navarro, R., Cano-Vindel, A., Moriana, J.A., Medrano, L.A., Ruiz-Rodríguez, P., Agüero-Gento, L., Rodríguez-Enríquez, M., Pizà, M.R., Ramírez-Manent, J.I., 2017. Screening for generalised anxiety disorder in Spanish primary care centres with the GAD-7. Psychiatry Res. 256, 312–317.

Noah, B., Keller, M.S., Mosadeghi, S., Stein, L., Johl, S., Delshad, S., Tashjian, V.C., Lew, D., Kwan, J.T., Jusufagic, A., et al., 2018. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomised controlled trials. NPJ Digit. Med. 1 (1), 1–12.

Olivas, E.S., Guerrero, J.D.M., Martinez-Sober, M., MagdalenaBenedito, J.R., Serrano, L., et al., 2009. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques. IGI Global.

Otto, M., Smits, J.A., 2011. Exercise for Mood and Anxiety: Proven Strategies for Overcoming Depression and Enhancing Well-being. OUP USA.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Roemer, L., Orsillo, S.M., 2002. Expanding our conceptualisation of and treatment for generalised anxiety disorder: integrating mindfulness/acceptance-based approaches with existing cognitive-behavioural models. Clin. Psychol. Sci. Pract. 9 (1), 54.

Rowa, K., Antony, M.M., 2008. Generalized anxiety disorder. Psychopathology: History, diagnosis, and empirical foundations 78–114.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432.

Sakr, S., Elgammal, A., 2016. Towards a comprehensive data analytics framework for smart healthcare services. Big Data Res. 4, 44–58.

Servia-Rodríguez, S., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., Lathia, N., Sandstrom, G. M., 2017. 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, pp. 103–112.

Speekenbrink, M., Visser, I., 2021. Ignorable and Non-ignorable Missing Data in Hidden Markov Models (arXiv preprint arXiv:2109.02770).

Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B., 2006. A brief measure for assessing generalised anxiety disorder: the gad-7. Arch. Intern. Med. 166 (10), 1092–1097.

Sükei, E., Romero-Medrano, L., de Leon-Martinez, S., Herrera López, J., Campaña-Montes, J.J., Olmos, P.M., Baca-Garcia, E., Artés, A., 2022. Assessing WHODAS 2.0 scores from behavioural biomarkers: a data-driven approach (preprint).

Van Grootven, B., Jeuris, A., Jonckers, M., Devriendt, E., Dierckx de Casterlé, B., Dubois, C., Fagard, K., Herregods, M.-C., Hornikx, M., Meuris, B., et al., 2020. Predicting hospitalisation-associated functional decline in older patients admitted to a cardiac care unit with cardiovascular disease: a prospective cohort study. BMC Geriatr. 20 (1), 1–7.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.

W. H. Organization, 2022. Physical activity. https://www.who.int/news-room/fact-sheets/detail/physical-activity (accessed on April 24, 2023).

Yao, S., Hu, S., Zhao, Y., Zhang, A., Abdelzaher, T., 2017. Deepsense: a unified deep learning framework for time-series mobile sensing data processing. In: Proceedings of the 26th International Conference on World Wide Web, pp. 351–360.