Investigating Feature Selection for Physical Activity Prediction Based on Ecological Momentary Assessments: Towards Tailoring the Timing of Behaviour Change Support Messages

Devender Kumar, David Haag, Jens Blechert, Josef Niebauer, Jan David Smeddinck

Submitted to: JMIR mHealth and uHealth
on: February 13, 2024

Table of Contents

Investigating Feature Selection for Physical Activity Prediction Based on Ecological Momentary Assessments: Towards Tailoring the Timing of Behaviour Change Support Messages

Devender Kumar[1] PhD; David Haag[1, 2, 3]; Jens Blechert[2, 4]; Josef Niebauer[1, 5]; Jan David Smeddinck[1, 6]

[1]Ludwig Boltzmann Institute for Digital Health and Prevention Salzburg AT
[2]Department of Psychology Paris-Lodron-University of Salzburg Salzburg AT
[3]Digital Health Information Systems, Center for Health & Bioresources AIT Austrian Institute of Technology GmbH Graz AT
[4]Centre for Cognitive Neuroscience Paris-Lodron-University of Salzburg Salzburg AT
[5]University Institute of Sports Medicine, Prevention and Rehabilitation, Paracelsus Medical University Salzburg AT
[6]LMU Munich Munich DE

Corresponding Author:
Devender Kumar PhD
Ludwig Boltzmann Institute for Digital Health and Prevention
Lindhofstraße 22, 5020 Salzburg
Institut f Sportmedizin d Landes Salzburg
Salzburg
AT

Abstract

Background: There has been a surge in the development of applications that aim to improve health, physical activity (PA), and well-being through behavior change. These apps often focus on creating a long-term and sustainable impact on the user. Just-in-time adaptive interventions (JITAIs) based on passive sensing of the current user context (e.g., from smartphones and wearables) have been devised to enhance the effectiveness of these apps and foster PA. JITAIs aim to provide personalized support and interventions, such as encouraging messages, in a context-aware manner. However, based on a limited range of passive sensing capabilities, getting the timing and context right for delivering well accepted and effective interventions is often challenging. Ecological Momentary Assessment (EMA) can provide personal context by directly capturing user assessments e.g. moods and emotion. Thus, EMA might be a useful complement to passive sensing in determining when JITAIs are triggered. Yet, extensive EMA schedules need to be scrutinized as they can increase user burden.

Objective: Use machine learning (ML) to balance feature set size of EMA questions with prediction accuracy regarding likelihood of enacting PA.

Methods: A total of 43 healthy participants (ages 19-67) completed four EMA surveys daily for three weeks. These surveys prospectively assessed different states including both motivational and volitional variables of PA preparation (e.g. intrinsic motivation, self-efficacy, perceived barriers) alongside stress and mood/emotions. PA enactment was assessed retrospectively via EMA and served as the outcome variable

Results: The best performing ML models predicted PA engagement with an AUC score of 0.87 ± 0.02 SD in 5-fold cross validation and 0.87 on test set. Particularly strong predictors included self-efficacy, stress, planning, and perceived barriers, indicating that a small set of EMA predictors can yield accurate PA prediction.

Conclusions: A small set of EMA based features like self-efficacy, stress, planning and perceived barriers can be enough to predict PA reasonably well and can thus be used to meaningfully tailor JITAIs such as sending well-timed and context aware support messages.

(JMIR Preprints 13/02/2024:57255)
DOI: https://doi.org/10.2196/preprints.57255

Preprint Settings

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

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

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

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

Original Manuscript

Investigating Feature Selection for Physical Activity Prediction Based on Ecological Momentary Assessments: Towards Tailoring the Timing of Behaviour Change Support Messages

Devender Kumar, David Haag, Josef Niebauer, Jens Blechert, Jan Smeddinck

Abstract

Background: There has been a surge in the development of applications that aim to improve health, physical activity (PA), and well-being through behavior change. These apps often focus on creating a long-term and sustainable impact on the user. Just-in-time adaptive interventions (JITAIs) based on passive sensing of the current user context (e.g., from smartphones and wearables) have been devised to enhance the effectiveness of these apps and foster PA. JITAIs aim to provide personalized support and interventions, such as encouraging messages, in a context-aware manner. However, based on a limited range of passive sensing capabilities, getting the timing and context right for delivering well accepted and effective interventions is often challenging. Ecological Momentary Assessment (EMA) can provide personal context by directly capturing user assessments e.g. moods and emotion. Thus, EMA might be a useful complement to passive sensing in determining when JITAIs are triggered. Yet, extensive EMA schedules need to be scrutinized as they can increase user burden.

Objective: Use machine learning (ML) to balance feature set size of EMA questions with prediction accuracy regarding likelihood of enacting PA.

Methods: A total of 43 healthy participants (ages 19-67) completed four EMA surveys daily for three weeks. These surveys prospectively assessed different states including both motivational and volitional variables of PA preparation (e.g. intrinsic motivation, self-efficacy, perceived barriers) alongside stress and mood/emotions. PA enactment was assessed retrospectively via EMA and served as the outcome variable.

Results: The best performing ML models predicted PA engagement with an AUC score of 0.87 ± 0.02 SD in 5-fold cross validation and 0.87 on test set. Particularly strong predictors included self-efficacy, stress, planning, and perceived barriers, indicating that a small set of EMA predictors can yield accurate PA prediction.

Conclusions: A small set of EMA based features like self-efficacy, stress, planning and perceived barriers can be enough to predict PA reasonably well and can thus be used to meaningfully tailor JITAIs such as sending well-timed and context aware support messages.

Keywords: digital health, behavior change, tailoring, personalization, adaptive systems, ecological momentary assessments, sensing, questionnaires, machine learning, feature selection, situated research, physical activity, implementation intentions, barriers, intention-behavior gap

Introduction

Due to pressures on societal and health-care systems linked to aging populations and increased prevalence in chronic diseases related to sedentary lifestyle and other behavioral patterns [1], a growing number of digital health applications aim at promoting positive lifestyle changes [2]. Such

applications have the potential to improve health outcomes [3], prevent diseases [4], and enhance the quality of life for individuals [5]. For instance, encouraging physical activity (PA) and heart-healthy habits can help prevent serious health issues like coronary heart disease, diabetes, and cancer [6]. Yet, long-term adherence to PA presents a significant challenge [7]. Despite the well-documented benefits of regular exercise for overall health and disease prevention, people struggle to maintain consistent PA [7].

To improve the effectiveness and adherence to PA, *Just-in-Time Adaptive Interventions* (JITAIs) are being investigated for tailoring personalized and contextualized digital health support, often enabled by mHealth technologies, such as wearable sensing devices and smartphones [8] [9]. JITAIs offer a promising solution to tackle issues around physical inactivity enabling effective behavior change and habitualization by providing personalized and timely support to individuals, e.g. sending motivational messages to incentivize movement after prolonged periods of inactivity. JITAIs are configured to use real-time data and context-awareness to deliver the 'right' interventions precisely when they are needed the most [8][9]. JITAIs based on passive sensing have already been applied to various areas, including eating disorders [10], mental health conditions, obesity and weight management, physical activity promotion [11], and smoking cessation [12]. Tailoring of JITIAs based on *passively sensed contextual factors* (PSCF), such as location, activity type/levels, daily weather conditions, or an individual's heart rate over time is commonly observed in the literature. However, these passively sensed features face the challenge of capturing telling aspects, which enable characterizations of the contexts that accurately represent what matters to a user at a given time. To meet this challenge, it would be necessary to validate passively sensed feature sets against the subjective user experience, which is very difficult in the context of emotions and self-regulation that play a central part in the enactment of health behaviors such as PA [13]. Therefore, tailoring PA fostering JITAIs based on PSCF comes with a considerable risk of misaligning intervention timing and content with users' current states. This may annoy users or could even hinder engagement up to dropping out from JITAI use [14][9]. Yet, self-report measures of momentary affect and motivation are closely related to actual PA [15] [16]. This begs the question whether these self-reports – referred to as *Ecological Momentary Assessment* (EMA) – could directly be used as an addition or alternative to tailoring JITAIs via PSCF.

EMA can present a valuable option for collecting near-real-time information on the experiences, interests, abilities, needs, behaviors, or other contextual circumstances of individuals in their natural context [17][18]. The advantages of EMA lie in providing rich, context-specific data with minimized biases, facilitating a deeper understanding of human behaviors and experiences that are also directly tied to the individual. EMAs are therefore, widely used across various fields to gain insights into psychological and behavioral dynamics [18]. Although EMA has shown tremendous potential in capturing individuals' momentary experiences, its utility in PA adherence prediction has not been widely explored. Additionally, EMA also comes with a set of challenges such as participant burden and reactance [18]. Participant burden refers to the issue that asking too many questions in an EMA can leave participants annoyed, potentially causing them to ignore EMA prompts or discontinue their participation entirely [19]. Reactance, on the other hand, can be another outcome of overly frequent or extensive inquiries, which might also unduly influence participants' perceptions or behaviors [18]. Compared to EMA deployment in limited duration study settings, these concerns are even more relevant when EMA is intended to inform JITAI tailoring over prolonged periods of time. Thus, alongside investigating if EMA can at all be used to accurately predict PA, it is critical to

identify the most predictive EMA questions that allow for a sufficiently accurate prediction of PA without overburdening the user.

Accordingly, we investigated if EMA can be utilized to predict PA and thereby inform the contextualized tailoring of JITAIs. Further, from a practical perspective, we focused on understanding design implications and how to balance practical concerns of 'EMA fatigue' [20][21] with traditional optimization and feature selection techniques in ML. This work therefore carries novelty both in exploring the viability of EMA for the timewise tailoring of JITAIs – in terms of 'when a JITAI should optimally be delivered' – and in preparing such tailoring not directly based on e.g. observed variable thresholds, but on prediction outcomes. These overarching aims are reflected in the following guiding research questions:

RQ1: *How accurately can EMAs predict whether short term intentions to carry out PA are put into action?*

RQ2: *Which motivational, emotional or volitional psychological states captured through EMAs will best predict PA? And which is the smallest set of these predictors that balances acceptable user burden with practically sufficient prediction accuracy?*

Based on the existing EMA literature [22], we expect that EMA can inform behavioural predictions as they closely reflect a participant's perceived state, which arguably can have a close relation to their subsequent behavior. Regarding the constructs that could inform prediction of PA from EMA, health behavior models such as the Integrated Behavior Change Model [23], the Health Action Process Approach (HAPA) [24] or the Temporal Self-Regulation Theory (TST) [13] alongside our prior work on predictors of PA [15] would suggest motivational and volitional qualities of self-regulation to be closely related to PA engagement. This includes constructs such as intrinsic motivation, intention, action planning, or self-efficacy alongside anticipated contextual barriers and momentary affect (e.g. current mood or stress). Therefore, we also expect these factors to be strong predictors in this study.

Methods

Study Design:

The EMA data used in this article were collected in our prior study [15] aiming to understand determinants and barriers of physical activity engagement. Following informed consent and after completing psychometric and demographic (see supplement 1) questionnaires in an online survey, participants were setup to complete the study's EMA phase. During this phase, four EMA prompts were sent to the participants on a daily basis at fixed time points (9 am, 1 pm, 5 pm, and 9 pm) over the course of a three-week study period. Additionally, participants were also able to report activities independently of those fixed EMA prompts. After that, another online questionnaire was sent to the participants, assessing compliance and reactivity. However, the analyses in this paper focus on the EMA data. The study received ethics approval at the University of Salzburg, Austria [GZ 11/2020].

EMA measures:

In selecting EMA items, we intended to map the structure proposed in so-called *dual-process models of health behavior*, such as HAPA [27], the *Integrated Behavior-Change Model* for PA [23], or TST [13]. These models propose a two-step process to explain the implementation of health behaviours

like PA. This process consists of (1) a *motivational phase*, which leads to the formation of intentions and (2) a *volitional phase* bridging the gap between intention and health behavior enactment.

To map these two steps in our study design, we first sampled predictors that the above-mentioned models propose as relevant in the motivational phase at each EMA prompt. For this, we selected measures of momentary *mood* (ten items from the *Positive and Negative Affect Schedule*; PANAS [28]: Happy, Relaxed, Active, Irritated, Concerned, Depressed, Nervous, Stressed, Energetic, Tired), *stress* (two items from the *Perceived Stress Scale*; PSS [29]; German version by Schneider et al. [30]), anticipated *barriers to PA* (*"How well would your given circumstances allow you to be physically active at the moment?"*; BarrPA) and also specifically *pain* as a barrier to PA (*"At the moment, do you have physical complaints that impede physical activity?"*). Furthermore, the morning prompt contained items assessing *sleep quality* (*"How good was your sleep?"; SleepQlt*), the time of falling asleep (*"When did you fall asleep?"; ST), and the waking time* (*"When did you wake up in the morning?"; WT*), which could also be barriers or resources to PA engagement.

Then, we prospectively asked the participants for their *intentions* to be physically active ("Do you intend to be physically active in the next 4h?", yes or no). Only if they responded with 'yes', we also assessed *volitional determinants of health behavior enactment* such as *planning specificity* (*"How specifically did you plan this physical activity?"*; ActPlan), *self-efficacy* (*"How strongly do you believe you can enact your plan under the given circumstances?"*; ActPlanSE), and momentary *intrinsic motivation* (*"Independent of the circumstances, how motivated are you right now to be physically active?"*).

All of these items were developed or adjusted to fit the specific needs of this study. Except for intention and sleep duration, all of them were measured on a horizontal slider from 0 ('not at all') to 100 ('very much'). Within each EMA prompt, participants finally reported their *PA* retrospectively (*"Have you been physically active in the last 4 h?"*, yes or no), which we used as the primary outcome in our analyses (see Haag et al. [15] for cross validation of these PA self-reports against wearable data). For this EMA sampling, we first used the Smarteater App and in a later recruitment phase switched to the m-Path platform [31]. However, each participant only used one platform and assessments were identical on both platforms. Therefore, data quality was not impaired by the switch. For the full list of EMA items, please refer to the supplementary material.

Study Participants

Participants were healthy individuals who had no limitations in their ability to perform PA, but neither described themselves as competitive athletes. Until June 2022, a total of 49 participants were enrolled in the data collection, which was conducted in a phased manner. Ten out of the 49 participants dropped out before completing all three weeks. However, the data from 4 of the 10 dropouts were still used for this investigation since they filled out enough (over 1 week) EMA prompts to be included. The final sample included 43 participants (31 female, 12 male) aged between 19 and 67 (M = 39.14, SD = 15.53) years. 16 reported high, 22 moderate and 5 low activity levels in the IPAQ (*International Physical Activity Questionnaire*; [25]) completed at the individual study start. Based on participants self-reported height and weight, their body mass index (BMI) ranged from 19.0 to 39.6 (M = 24.4kg/m^2, SD = 3.4kg/m^2) kg/m^2, with 2 participants being obese (BMI ≥ 30.0kg/m^2), 11 overweight (BMI = 25.0-29.9kg/m^2), and 29 of normal weight (BMI = 18.5-24.9kg/m^2). The BMI for one participant was omitted due to being unrealistic.

Study participants were compensated with 30€ and received personalised feedback based on their data during the study [26]. For further details on study setup, procedure, as well as the primary outcomes, please cf. Haag et al. [15].

Data cleaning and feature engineering:

Each EMA prompt contained at least 16 questions. Some questions in the EMA prompts included conditional sub-questions (e.g. asking for *planning specificity* only when a PA was intended) and questions that only required per-day-level measurements were only presented at one prompt per day (e.g. sleep quality and duration were only gauged in the morning prompts). In total, this led to 41 different questions which were assessed in the EMA. For practical reasons we only considered the prospective and momentary questions as predictors of retrospectively assessed PA in the following analysis since they could reasonably be employed for a model tasked to predict the likelihood of upcoming PA adherence at runtime. Further details of the EMA questions can be found in the supplementary material and in our previous article [15].

In the data cleaning process, EMA questions with less than 30% response compliance rate were dropped, because data augmentation on a low compliance rate could lead to bias in the models. EMA prompts for which participants did not report whether or not they performed the PA were also discarded. The times participants fell asleep and woke up were self-reported once only in the first EMA of the day. Therefore, empty cells for ST and WT for EMA prompts 2, 3, and 4 were populated with values from the first EMA of each day. Any remaining missing values for the sleep hours, wakeup time, and sleep quality – cases in which a participant did not respond to a day's first EMA prompt – were replaced with their respective median values. The StandardScaler of the Scikit-learn package [32] was used to scale all the continuous variables to unit variance. This standardization of input features can be beneficial for the performance and convergence of ML algorithms. The categorical EMA questions with 'yes/no' answers were recoded as 1/0.

Model Training and validation:

For modeling, a representative selection of common ML methods, including logistic regression (LR), decision trees (DT), support vector machines (SVM) [33], k-nearest neighbors (KNN) [34], random forest (RF) [35], and gradient boosting (XGBoost/XGB) [36] were used to compare the model prediction performances. The decision to choose traditional ML models for exploration over neural networks (NN) was made given the relatively small size of the dataset, since NN typically require an extensive dataset for training. The data was split into training/test sets (in a ratio of 80/20) using the Scikit-learn package [32]. A stratified 5-fold cross validation method on the training dataset was used to evaluate model performance and hyperparameter tuning. The stratified k-fold was chosen to ensure that each fold's class distribution is similar to the overall class distribution in the dataset. This is particularly useful when dealing with imbalanced datasets, where one class may have significantly more instances than the others. All models were then evaluated on the test dataset. The GridSearchCV scoring method from Scikit-learn [32] was used to find the optimal hyper-parameters.

Model prediction performances were compared using the *Area Under the Receiver Operating Characteristic Curve* (AUROC) as it is a more appropriate measure compared to simple accuracy, especially for imbalanced dataset [37]. It visualizes the performance of the model at various probability thresholds for classification and helps in selecting an appropriate threshold that balances the trade-off between true positives and false positives. To overcome the data imbalance problem, we also attempted up-sampling with SMOTE (the *synthetic minority oversampling technique*) [38] that is implemented in the Python library *imbalanced-learn* [39] on the training data. We trained above-mentioned models with the up-sampled data and compared the internal validation results of the models on the test dataset. For finding the top features contributing to predicting the binary outcome

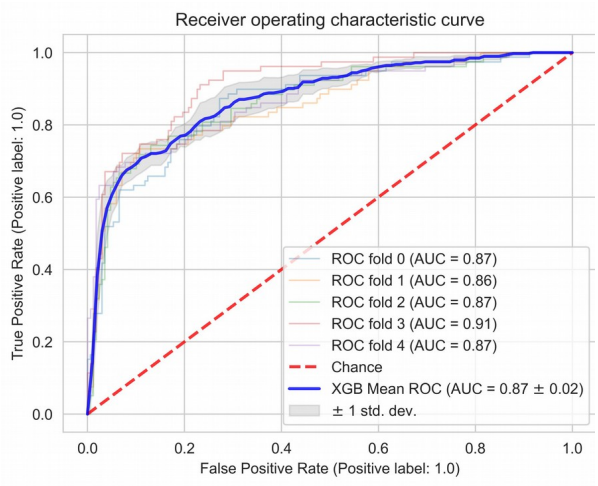
whether PA did take place or not within a given EAM slot, we employed the *recursive feature elimination* (RFE) technique using Scikit-learn [32]. RFE is primarily used in ML to select the most relevant and important features from a given dataset. It recursively trains the model using subsets of features and ranks them based on their contribution to improving prediction outcome success. By iteratively eliminating less important features, RFE gives an optimal subset of features that maximizes the model's performance. Besides the RFE, we also used the SHAP (*Shapley Additive exPlanations*) framework [40] for understanding the feature importance within the best-performing models.

Results

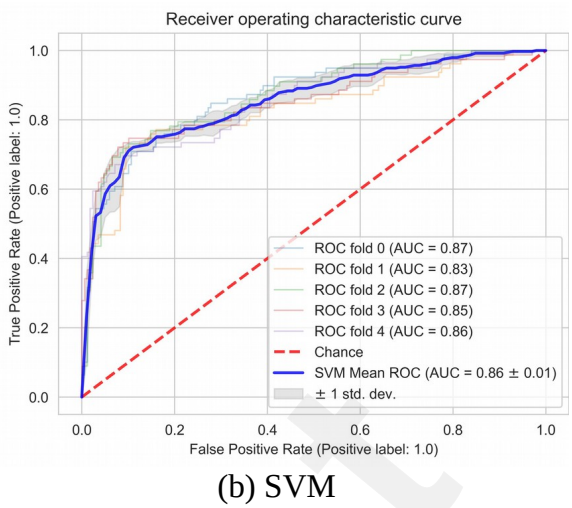
Following data cleaning, 23 out of 41 EMA questions fulfilled the minimum data compliance requirements and were used as features for building the models and predicting the compliance or non-compliance of PA. In this section, we first outline the performance of the range of selected candidate ML models in predicting PA when all the features were used in training. Thereafter, we will show the results of feature importance in predicting the PA using the RFE technique and a SHAP value visualization. We conclude with a practically informed discussion around selecting a particular model or ensemble and feature set contextualizing the work around more general concerns in model training and selection for dynamic personalization based on EMA.

PA prediction performance of various model using all available features:

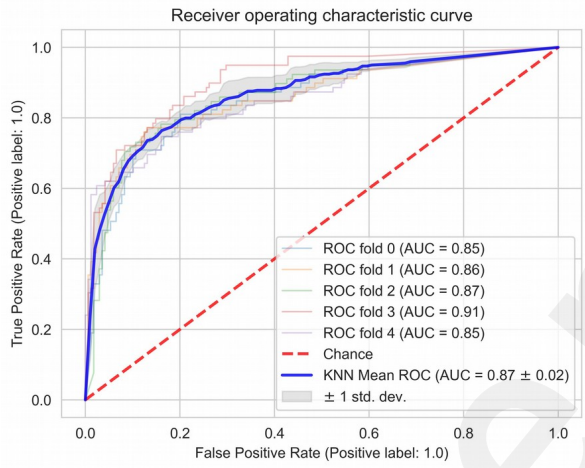
Figure 1 shows the AUROC of various models in *a 5-fold* cross-validation on training data, and Table 1 presents their AUC scores on training and test set. XGB and RF models have shown the best performance and achieved an AUC score of 0.87 ± 0.02 on 5-fold cross-validation and XGB achieves a slightly higher AUC score (0.87) on the test set. We also tested the up-sampling technique SMOTE (*synthetic minority oversampling technique*) on the training data; however, as shown in the two rightmost columns in Table 1, this did not significantly improve test set AUC score. Since SMOTE does not improve the perfomance on test set, therefore, it was not used for further analysis.



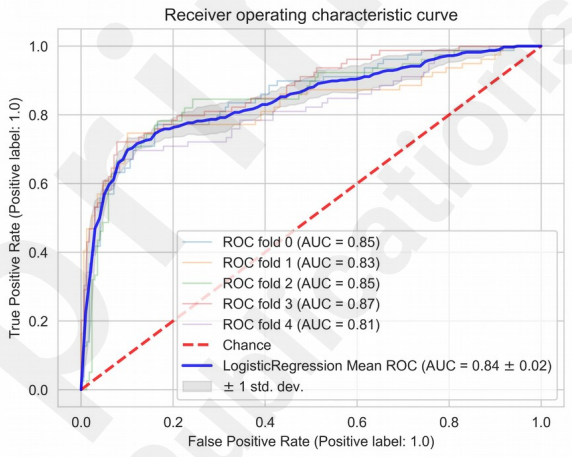
(a) XBG



(b) SVM

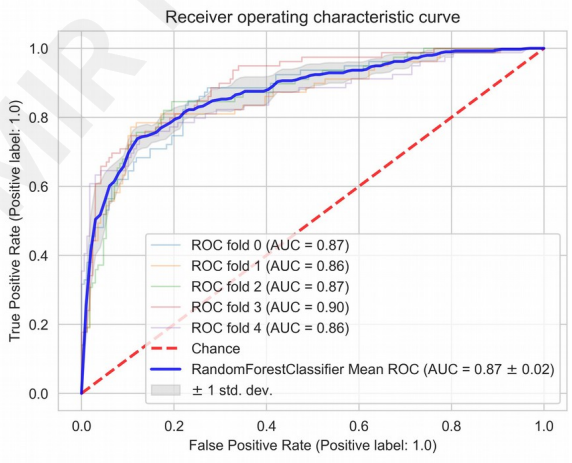


(c) KNN



(d) LR

Figure 1: (a), (b), (c), (d), and (e) show the performance of selected ML models in 5-fold cross validation using all the features.



(e) RF

Model Name	AUC score 5-fold cross valid. (M±SD)	AUC score on Test set	SMOTE AUC score 5-fold cross validation (M±SD)	SMOTE AUC score on test set
XGB	0.87 ± 0.02	0.87	0.93 ± 0.01	0.86
RF	0.87 ± 0.02	0.86	0.94 ± 0.01	0.86
KNN	0.87 ± 0.03	0.86	0.92 ± 0.01	0.85
LR	0.84 ± 0.02	0.85	0.86 ± 0.02	0.83
SVM	0.86 ± 0.02	0.86	0.86 ± 0.03	0.84
DT	0.83 ± 0.01	0.83	0.86 ± 0.01	0.81

Table 1: Performance of various models on 5-fold CV and test set. M = mean; SD = standard deviation, contrasted with performance following synthetic minority oversampling (SMOTE) [8] of training data.

Feature importance:

Aiming to reduce potential questionnaire fatigue [20] due to extensive EMA survey length and finding the smallest but accurate EMA questions subset for predicting the PA, after finalizing the best-performing model on all 23 features, we focused on understating the feature importance in predicting PA. Since the XGB model showed slightly higher performance on the test dataset (cf. Table 1) compared to RF and KNN, explorations into feature importance were executed using XGB only.

Recursive Feature Elimination: Figure 2 shows the selected top-n (n = number 1,2,3 etc.) features identified by RFE together with a box plot of the corresponding AUC scores for the XGB classifier with the given set of n features. As Figure 2 depicts, for the XGB models, the highest mean AUC score (0.87) is achieved when using 13 features and then remains the same even when all 23 features are used. At the same time, a competitively performing AUC score of 0.85 is already achieved by the XGB model using 3 features only, in this case combining ActPlan and ActPlanSE as indicators of volitional intention with PSS2 as a marker of stress. An increase to an AUC score of 0.86 is only achieved when using 10 or more features.

SHAP values-based feature importance: Figure 3 shows a SHAP values graphs of the top 13 features and their contributions in PA prediction by the XGB model. For brevity we only show SHAP value of 13 features as after which subsequent features don’t improve the AUC score (as apparent in Figure 2). SHAP value graph indicates the correlations through which different features are contributing to the model's output for individual predictions.

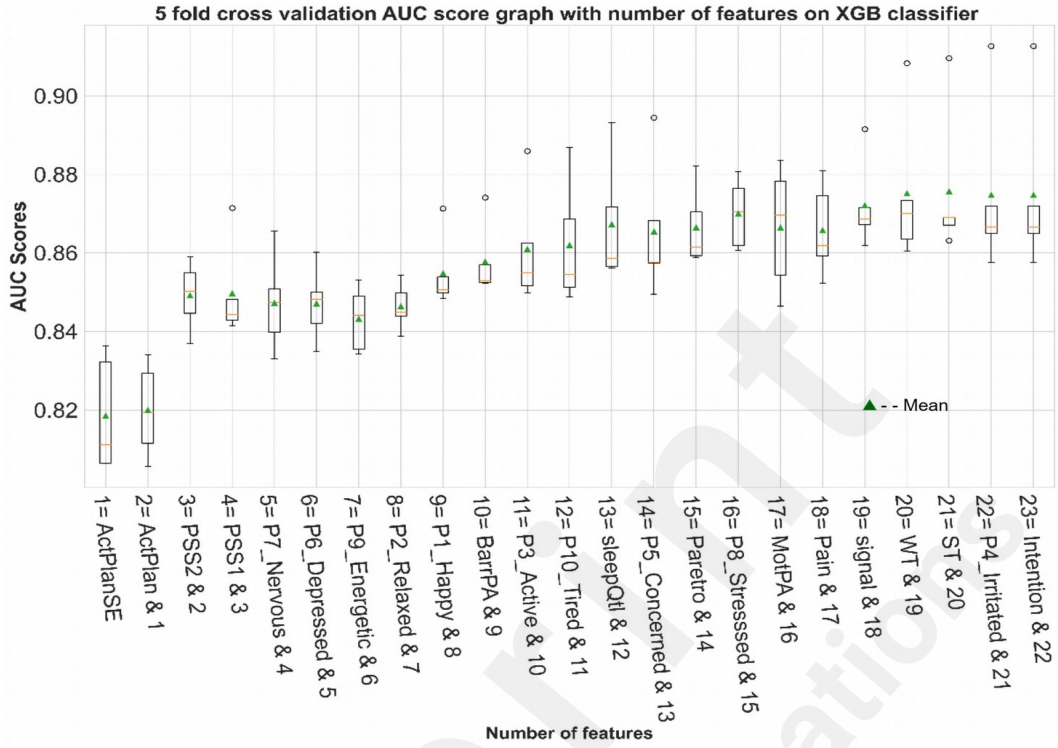


Figure 2. XGB models with growing feature set sizes: Boxplots of AUC scores with number of features using RFE. The numbers (1,2,3 etc.) indicate the top n-number of features and their names. Feature sets are listed as “additonal_feature & X (identifier of compounded prior features)”.

Similar to the RFE (Figure2), items ActPlanSE (self-efficacy) and ActPlan (PA planning specificity) remain the top 2 features with most notable impact on model output (indicated by wide impact score spread combined with decisive directionality on the horizontal axis). The positive SHAP values indicate that the feature pushes the prediction higher, while negative values indicate the opposite. The color of the bar represents the feature value (red for high, blue for low). Note that the slight difference in the features ranking by RFE and SHAP value is due to difference in their feature selection methodology and objective of feature selection. RFE focuses on improving overall model performance metrics whereas SHAP values aim to provide interpretable explanations for individual prediction by a given feature.

Discussion

In this study, we explored the viability of predicting PA execution based on common supervised ML methods with features drawn from EMA data collected in the participants’ natural contexts. Conceptually, if PA execution or non-execution can be predicted, this can inform the timing of issuing JITAIs, such as motivational messages of encouragement, or to foster more specific planning or replanning. Since EMA requires active involvement of the user, we also explored which features/questions of the EMA are key determinants for effectively predicting PA. This exploration offers an example procedure for informing trade-off decisions that can help in designing JITAI systems that

take valuable information derived from EMA without overburdening the users.

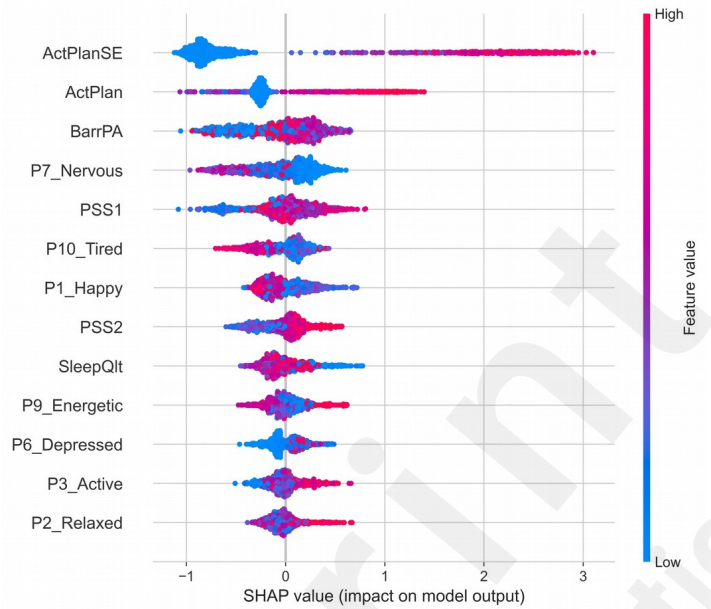


Figure 3: SHAP values graph of top 13 features indicating how each feature is contributing to the XGB model's output.

EMA and PA prediction

Regarding RQ1 on how accurately EMAs and ML models can help predict PA, all the ML models (cf. Table 1) produce AUC scores of 80 or higher on test data that was not seen by the models in training. When considering smaller performance differences, both XGB and RF models performed best on both cross-validation (0.87±0.2) and test sets (AUC 0.87 and 0.86 respectively), broadly indicating that EMA could be a viable option for PA prediction and can complement the tailoring of passive sensing based JITAIs to foster PA.

While such performance levels would not satisfy strong reliability requirements e.g. in health diagnostics, they can arguably be reasonably employed in JITAI settings around e.g. fostering long-term PA, where occasional Type1/Type2 errors do not have grave consequences. Compared with previous work on the same data set, which used non-ML models to investigate determinants of PA enactment in a theory-driven manner [15], this work takes a practical angle on the potential of operationalizing EMA for actively "driving JITAIs" independently of preconditions, such as the PA being explicitly intended. Thereby, favoring model performance over analyses intended to inform theory-building for feature composition.

Still, the outcomes of automated feature selection corroborate earlier findings (give reference) with non-ML modeling, as feature importance analysis indicates that self-efficacy (ActPlanSE) and planning specificity (ActPlan) are the key relevant predictors of PA engagement. Here self-efficacy represents a participant's confidence in their ability to engage in the intended activity and planning specificity refers to how specifically they had planned that activity. This result also falls in line with other previous publications

indicating a close relationship between PA engagement and these volitional components [41] [42] [43]. Further, analysis based on RFE on the XGB model (see Figure 2) indicates stress (PSS1 and PSS2) as another potentially relevant factor that can be a qualifier for PA outcomes. This association of stress and PA is also consistent with previous EMA literature [16].

Psychological considerations and contextualization in theory

In the light of the dual process models of health behavior [13] [23] which the present data collection was based upon, it is not surprising to see the volitional constructs (ActPlanSE and ActPlan) ending up at the top of our feature selection. However, even though we found a resemblance to these psychological models, our results are not to be confused with an evaluation of such models. Instead, our findings represent a practical investigation of the ML methodology being applied to determine, with a given set of candidate EMA items, which ones could be employed for a runtime system to predict PA adherence and control adaptive interventions accordingly. This implies that the proposed feature selection might be very different in another set of potential predictors and necessitates the interpretation of our findings in the light of our specific EMA design.

For example, considering the intention item/feature, which was ranked at the very bottom of our RFE (see Figure 2): From a theoretical point of view, intention would often be seen as the pivotal point in health behavior engagement [24]. Therefore, it may seem very surprising that intention is being ranked the least predictive feature in our data set. This discrepancy originates in the structure of our EMA with ActPlanSE and ActPlan only being assessed if participants report to have the intention for PA. Therefore, these features would have missing values for each episode where no intention is reported. In these cases, values of the volitional determinants were replaced with zeros. Thus, the binary intention data is encoded in ActPlanSE and ActPlan. In the RFE process, the presented models will have picked up on this characteristic and ranked intention itself very low since it doesn't contain additional information if ActPlan/ActPlanSE are included. However, this does not imply that we can forgo the assessment of intention, since it is only meaningfully possible to assess planning and self-efficacy when an intention is given. In practical terms this would imply that the precondition of intention being given would be wrapped into the wording for an ActPlan (SE) item if it were to be presented without an explicit item on intention preceding it (i.e. "In case you are intending to exercise, how specifically have you planned ").

Nonetheless, the indicated feature combination including ActPlan and stress items corresponds well with expectations that can be derived from the aforementioned dual-process models of health behavior [27][23][13], representing elements that can be seen to capture aspects of both a motivational phase and a volitional phase.

Trade-off between EMA burden and PA prediction-accuracy

The potential viability of the sets of EMA questionnaires selected for this study was informed by psychological theories. However, the list of questions as a whole is clearly too burdensome for being of practical use in guiding JITAIs. As indicated in the EMA

literature [19][20][21], questionnaire fatigue/overload is a real concern for practicality. As shown in Figure 2, an AUC of over 0.85 is achieved by just using the top 3 features (i.e., ActPlanSE, ActPlan, PSS2) as compared to maximal achievable AUC of 0.87 when all the 23 features were used. Moreover, as evidenced by the RFE process, the maximal AUC score of 0.87 is already achieved with 13 features indicating that beyond these 13 features rest other features are irrelevant from the PA predation point of view. Even within these 13 features, after adding an item from the stress questions (PSS1 and PSS2), PANAS mood questions (P7_nerves, P5_concern, P1_happy, P6_depressed, P7_energetic) are arguably not adding significant improvements in AUC score to justify their inclusion as frequently asked items in a practical longer-term deployment. The PA-friendly external circumstances (BarrPA), i.e. the absence of barriers and Sleep Quality (SleepQlt) do slight lifting of the AUC score and could be candidates for inclusion with the EMA for research purposes, but would likely not be included in a production system built on the available foundations with the models built on the currently available data.

For most practical purposes and keeping the EMA questionnaire burden to the minimum, in settings where an intention for PA is given (e.g. supporting the execution of a PA plan), a deployed JITAI model could be driven by just three EMA items (ActPlanSE, ActPlan, and PSS2). If intention is not given, the first item could be rephrased as indicated above.

Design Implications

For the practicality of JITAIs or any intervention intended for fostering PA to be driven by regular EMA in live deployment situations, it is essential to select EMA questions effectively and avoid causing EMA fatigue. In relation to RQ2 about which concepts will best inform PA prediction, as shown in the results, features (EMA questions) related to self-efficacy, planning, and stress have reasonable PA prediction power. These outcomes indicate that the long list of potentially relevant EMA questions used in PA prediction can be effectively cut short to a degree where real-world deployments of JITAIs with decision rules or trigger points [14] being informed by EMAs appear plausible.

Limitations and future work

This study is formative with regards to surveying potentially relevant EMA constructs and measures. It does not include observations of the application of the derived models in practice concerning performance as well as acceptability, which is a clearly indicated step for future work. The study also included a relatively small number of participants (43). Ensuring the generalizability of the ML model (XGB) for automated PA prediction and customization of JITAIs across broader demographics would necessitate further investigation.

In the subsequent research, we intend to gather data from larger and more diverse demographic cohorts to thoroughly test the chosen XGB model. Additionally, in the current study setup, the EMA frequency was established at four times a day. In future work, we plan to explore the impact of varying EMA frequencies per day on the predictive capabilities of the model for PA.

Conclusion

This paper presents a formative investigation into how well EMA can help with predicting PA implementation and whether EMA can be of practical use in gathering temporal context for JITAI decision-making. The outcomes with common supervised ML models, especially XGB with AUC score of 0.87 ± 0.02 indicate that EMA can offer relevant PA prediction power and thereby have the potential to complement more common passive sensing JITAI tailoring approaches. Furthermore, the investigation around finding a right trade-off between "EMA question load" and PA "prediction-accuracy" of ML models indicates that self-efficacy (ActPlanSE) and planning specificity (ActPlan) play the most important role in determining the PA prediction under the assumption that initial intention to perform PA was already given. Just three features ActPlanSE, ActPlan, and PSS2 (stress), allowed for achieving an AUC score of 0.85 as compared to the maximum AUC score of 0.87 when all the 23 EMA questions were used.

Acknowledgement:

We thank all study participants and all further contributors for their time and cooperation. Special thanks to Sebastian Gruber and Christoph Zeiner for their support during the design and early analysis of this study.

Conflicts of Interest:

The authors declare no conflicts of interest related to the research, authorship, or publication of this article.

Abbreviations

AUC: Area Under Curve
AUROC: Area Under the Receiver Operating Characteristic Curve
BMI: Body Mass Index
EMA: Ecological Momentary Assessment
DT: Decision Trees
JITAI: Just-In-Time Adaptive Interventions
KNN: K-Nearest Neighbors
LR: Logistic Regression
ML: Machine Learning
PA: Physical Activity
RF: Random Forest
RFE: Recursive Feature Elimination
SVM: Support Vector Machine
SD: Standard Deviation
SMOTE: Synthetic Minority Oversampling Technique
EMA: Ecological Momentary Assessment
XGB: Gradient Boosting

References:

1. WHO/NHI. Global Health and Aging. World Health Organization / National Institute on Aging, National Institutes of Health, U.S. Department of Health and Human Services; 2011. Available from: https://www.nia.nih.gov/sites/default/files/2017-06/global_health_aging.pdf

2. Woessner MN, Tacey A, Levinger-Limor A, Parker AG, Levinger P, Levinger I. The Evolution of Technology and Physical Inactivity: The Good, the Bad, and the Way Forward. Front Public Health 2021;9. doi: 10.3389/fpubh.2021.655491

3. Khan N, Marvel FA, Wang J, Martin SS. Digital Health Technologies to Promote Lifestyle Change and Adherence. Curr Treat Options Cardiovasc Med 2017 Jun 24;19(8):60. doi: 10.1007/s11936-017-0560-4

4. Gray R, Indraratna P, Lovell N, Ooi S-Y. Digital health technology in the prevention of heart failure and coronary artery disease. Cardiovasc Digit Health J 2022 Dec 1;3(6, Supplement):S9–S16. doi: 10.1016/j.cvdhj.2022.09.002

5. Victoria-Castro AM, Martin ML, Yamamoto Y, Melchinger H, Weinstein J, Nguyen A, Lee KA, Gerber B, Calderon F, Subair L, Lee V, Williams A, Shaw M, Arora T, Garcez A, Desai NR, Ahmad T, Wilson FP. Impact of Digital Health Technology on Quality of Life in Patients With Heart Failure. JACC Heart Fail 2023 Nov 8; doi: 10.1016/j.jchf.2023.09.022

6. Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. Can Med Assoc J 2006 Mar 14;174(6):801–809. doi: 10.1503/cmaj.051351

7. Collado-Mateo D, Lavín-Pérez AM, Peñacoba C, Del Coso J, Leyton-Román M, Luque-Casado A, Gasque P, Fernández-del-Olmo MÁ, Amado-Alonso D. Key Factors Associated with Adherence to Physical Exercise in Patients with Chronic Diseases and Older Adults: An Umbrella Review. Int J Environ Res Public Health 2021 Feb;18(4):2023. PMID:33669679

8. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Dec;16(1):31. doi: 10.1186/s12966-019-0792-7

9. Wunsch K, Eckert T, Fiedler J, Woll A. Just-in-time adaptive interventions in mobile physical activity interventions – A synthesis of frameworks and future directions. 22(4):10.

10. Ralph-Nearman C, Sandoval-Araujo LE, Karem A, Cusack CE, Glatt S, Hooper MA, Pena CR, Cohen D, Allen S, Cash ED, Welch K, Levinson CA. Using machine learning with passive wearable sensors to pilot the detection of eating disorder behaviors in everyday life. Psychol Med Cambridge University Press; 2023 Oct 20;1–7. doi: 10.1017/S003329172300288X

11. Nahum-Shani I, Smith SN, Tewari A, Witkiewitz K, Collins LM, Spring B, Murphy SA. Just-in-Time Adaptive Interventions (JITAIs): An Organizing Framework for Ongoing Health Behavior Support. 2014;(14).

12. Naughton F. Delivering "Just-In-Time" Smoking Cessation Support Via Mobile Phones: Current Knowledge and Future Directions. Nicotine Tob Res 2017 Mar 1;19(3):379–383. doi: 10.1093/ntr/ntw143

13. Hall PA, Fong GT. Temporal self-regulation theory: A model for individual health behavior. Health Psychol Rev 2007 Mar;1(1):6–52. doi: 10.1080/17437190701492437

14. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. Ann Behav Med 2018 May 18;52(6):446–462. doi: 10.1007/s12160-016-9830-8

15. Haag D, Carrozzo E, Pannicke B, Niebauer J, Blechert J. Within-person association of volitional factors and physical activity: Insights from an ecological momentary assessment study. Psychol Sport Exerc 2023 Sep;68:102445. doi: 10.1016/j.psychsport.2023.102445

16. Schultchen D, Reichenberger J, Mittl T, Weh TRM, Smyth JM, Blechert J, Pollatos O. Bidirectional relationship of stress and affect with physical activity and healthy eating. Br J Health Psychol 2019 May;24(2):315–333. doi: 10.1111/bjhp.12355

17. Shiffman S, Stone AA, Hufford MR. Ecological Momentary Assessment. Annu Rev Clin Psychol 2008 Apr 1;4(1):1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415

18. Dunton GF. Ecological Momentary Assessment in Physical Activity Research. Exerc Sport Sci Rev 2017 Jan;45(1):48–54. doi: 10.1249/JES.0000000000000092

19. Eisele G, Vachon H, Lafit G, Kuppens P, Houben M, Myin-Germeys I, Viechtbauer W. The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. Assessment SAGE Publications Inc; 2022 Mar 1;29(2):136–151. doi: 10.1177/1073191120957102

20. Yang YS, Ryu GW, Choi M. Methodological Strategies for Ecological Momentary Assessment to Evaluate Mood and Stress in Adult Patients Using Mobile Phones: Systematic Review. JMIR MHealth UHealth 2019 Apr 1;7(4):e11215. doi: 10.2196/11215

21. Porras-Segovia A, Molina-Madueño RM, Berrouiguet S, López-Castroman J, Barrigón ML, Pérez-Rodríguez MS, Marco JH, Díaz-Oliván I, de León S, Courtet P,

Artés-Rodríguez A, Baca-García E. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. J Affect Disord 2020 Sep 1;274:733–741. doi: 10.1016/j.jad.2020.05.067

22. Mikus A, Hoogendoorn M, Rocha A, Gama J, Ruwaard J, Riper H. Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data. Internet Interv 2018 Jun 1;12:105–110. doi: 10.1016/j.invent.2017.10.001

23. Hagger MS, Chatzisarantis NLD. An Integrated Behavior Change Model for Physical Activity. Exerc Sport Sci Rev 2014 Apr;42(2):62–69. doi: 10.1249/JES.0000000000000008

24. Schwarzer R, Luszczynska A. How to Overcome Health-Compromising Behaviors: The Health Action Process Approach. Eur Psychol 2008 Jan;13(2):141–151. doi: 10.1027/1016-9040.13.2.141

25. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, Oja P. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381–1395. PMID:12900694

26. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res 2016 Dec 16;18(12):e5870. doi: 10.2196/jmir.5870

27. Schwarzer R. Modeling Health Behavior Change: How to Predict and Modify the Adoption and Maintenance of Health Behaviors. Appl Psychol 2008 Jan;57(1):1–29. doi: 10.1111/j.1464-0597.2007.00325.x

28. Watson D, Clark LA, Tellegen A. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. :8. doi: 10.1037/0022-3514.54.6.1063

29. Cohen S, Kamarck T, Mermelstein R. A Global Measure of Perceived Stress. J Health Soc Behav 1983 Dec;24(4):385. doi: 10.2307/2136404

30. Schneider EE, Schönfelder S, Domke-Wolf M, Wessa M. Measuring stress in clinical and nonclinical subjects using a German adaptation of the Perceived Stress Scale. Int J Clin Health Psychol 2020 May;20(2):173–181. doi: 10.1016/j.ijchp.2020.03.004

31. Mestdagh M, Verdonck S, Piot M, Niemeijer K, tuerlinckx francis, Kuppens P, Dejonckheere E. m-Path: An easy-to-use and flexible platform for ecological

momentary assessment and intervention in behavioral research and clinical practice. PsyArXiv; 2022. doi: 10.31234/osf.io/uqdfs

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–2830.

33. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol TIST Acm New York, NY, USA; 2011;2(3):1–27. doi: 10.1145/1961189.1961199

34. Kramer O. K-nearest neighbors. Dimens Reduct Unsupervised Nearest Neighbors Springer; 2013. p. 13–23.

35. Breiman L. Random forests. Mach Learn Springer; 2001;45(1):5–32. doi: 10.1023/A:1010933404324

36. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proc 22nd Acm Sigkdd Int Conf Knowl Discov Data Min 2016. p. 785–794. doi: 10.1145/2939672.2939785

37. Melo F. Receiver Operating Characteristic (ROC) Curve. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encycl Syst Biol New York, NY: Springer New York; 2013. p. 1818–1823. doi: 10.1007/978-1-4419-9863-7_242

38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–357. doi: 10.1613/jair.953

39. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res 2017;18(17):1–5.

40. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.

41. Scholz U, Keller R, Perren S. Predicting behavioral intentions and physical exercise: A test of the health action process approach at the intrapersonal level. Health Psychol US: American Psychological Association; 2009;28(6):702–708. doi: 10.1037/a0016088

42. Scholz U, Schüz B, Ziegelmann JP, Lippke S, Schwarzer R. Beyond behavioural intentions: Planning mediates between intentions and physical activity. Br J Health Psychol 2008 Sep;13(3):479–494. doi: 10.1348/135910707X216062

43. Sniehotta FF, Scholz U, Schwarzer R. Bridging the intention–behaviour gap:

Planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. Psychol Health Routledge; 2005 Apr 1;20(2):143–160. doi: 10.1080/08870440512331317670