Smartphone-tracked digital biomarkers of stress: An idiographic machine learning perspective

George Aalbers, Andrew T. Hendrickson, Mariek M.P. Vanden Abeele, Loes Keijsers

Submitted to: JMIR mHealth and uHealth
on: February 22, 2022

Table of Contents

Smartphone-tracked digital biomarkers of stress: An idiographic machine learning perspective

George Aalbers[1] BSc, MSc; Andrew T. Hendrickson[2] PhD; Mariek M.P. Vanden Abeele[3] PhD; Loes Keijsers[4] PhD

[1]Department of Cognitive Science & Artificial Intelligence Tilburg University Tilburg NL
[2]Media, Innovation and Communication Technologies Department of Communication Sciences Ghent University Ghent BE
[3]Clinical Child and Family Studies Erasmus School of Social and Behavioural Sciences Erasmus University Rotterdam Rotterdam NL

Corresponding Author:
George Aalbers BSc, MSc

Abstract

Background: Stress is an important causal factor in common mental disorders such as burnout and depression. To aid in the early detection of chronic stress, machine learning models are increasingly trained to learn mathematical mappings from digital footprints to self-reported stress. Earlier work has studied general principles in population-wide studies, but the extent to which findings apply to individuals is understudied.

Objective: We investigated 1) if features of smartphone usage log data (e.g., Messenger application use frequency) are digital biomarkers that can be used to predict momentary subjective stress, 2) if these biomarkers are positively or negatively related to momentary subjective stress (at the group and individual levels), and 3) how accurate these potential digital biomarkers are at recognizing momentary subjective stress on a person-by-person basis in out-of-sample data.

Methods: Using a large-scale, intensive longitudinal dataset (N = 224, 44,381 observations), we trained machine learning models to predict momentary subjective stress, utilizing explainable artificial intelligence to identify potential digital biomarkers.

Results: We identified prolonged use of Messenger and Social Network site applications and sleep proxies as valid digital biomarkers. The relationships of these markers with momentary subjective stress as well as predictive accuracy of models differed from person to person. In the majority of individuals, model predictions correlated positively and significantly with self-reported stress.

Conclusions: Our findings indicate smartphone log data can be utilized as digital biomarkers of momentary subjective stress, but the relationship differs from person to person.

(JMIR Preprints 22/02/2022:37469)
DOI: https://doi.org/10.2196/preprints.37469

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

Smartphone-tracked digital biomarkers of stress: An idiographic machine learning perspective

George Aalbers[1,3], Andrew T. Hendrickson[1], Mariek M. P. vanden Abeele[2,3], & Loes Keijsers[3]

[1] Department of Cognitive Science & Artificial Intelligence, Tilburg University

[2] imec-mict-UGent, Department of Communication Sciences, Ghent University

[3] Department of Cognition & Communication, Tilburg University

[4] Department of Psychology, Education & Child Studies/Clinical Child and Family Studies, Erasmus University Rotterdam

Author note

Correspondence concerning this article should be addressed to George Aalbers, Warandelaan 2, Tilburg. E-mail: h.j.g.aalbers@tilburguniversity.edu

Abstract

Background: Stress is an important causal factor in common mental disorders such as burnout and depression. To aid in the early detection of chronic stress, machine learning models are increasingly trained to learn mathematical mappings from digital footprints to self-reported stress. Earlier work has studied general principles in population-wide studies, but the extent to which findings apply to individuals is understudied.

Objectives: We investigated 1) if features of smartphone usage log data (e.g., Messenger application use frequency) are *digital biomarkers* that can be used to predict momentary subjective stress, 2) if these biomarkers are positively or negatively related to momentary subjective stress (at the group and individual levels), and 3) how accurate these potential digital biomarkers are at recognizing momentary subjective stress on a person-by-person basis in out-of-sample data.

Methods: Using a large-scale, intensive longitudinal dataset (N = 224, 44,381 observations), we trained machine learning models to predict momentary subjective stress, utilizing explainable artificial intelligence to identify potential digital biomarkers.

Results: We identified prolonged use of Messenger and Social Network site applications and sleep proxies as valid digital biomarkers. The relationships of these markers with momentary subjective stress as well as predictive accuracy of models differed from person to person. In the majority of individuals, model predictions correlated positively and significantly with self-reported stress.

Conclusions: Our findings indicate smartphone log data can be utilized as digital biomarkers of momentary subjective stress, but the relationship differs from person to person.

Keywords: passive sensing, mobile sensing, personal sensing, context sensing, digital phenotyping, digital biomarkers, machine learning, experience sampling method, ESM, smartphone

Digital biomarkers of stress: An idiographic machine learning perspective on digital phenotyping

Introduction

Stress is an important predictor of common mental disorders such as depression [1] and a key causal agent in neurobiological theories of these disorders [2]. How stress influences mental health depends on its duration: Whereas stress with a duration of minutes to hours (*acute stress*) is commonly considered an adaptive psychophysiological response, stress lasting for weeks to months or even years (*chronic stress*) is believed to have adverse psychological and physiological consequences [3]. Given its potential effects, early detection and treatment of chronic stress is important to prevent mental and physiological illness.

Because asking individuals to consistently self-monitor and self-report stress over extended periods is a difficult, costly, and time-intensive procedure [4], researchers have started developing algorithms to unobtrusively detect stress from passively logged data, such as smartphone application usage log data [5-6]. If successful, such algorithms open opportunities for early detection of the tipping point where acute stress turns into chronic stress, possibly unlocking earlier possibilities for scalable interventions, such as moderate aerobic exercise [7], mindfulness meditation [8], or smartphone-based cognitive behavioral therapy with chatbots [9].

Previous studies in this domain suggest passively logged data might be used to detect how stressed a person feels on a given day [10-14] and forecast how stressed they will feel the next day [15-18]. However, it is relatively unclear which aspects of passively logged data are indicative of stress. The present study therefore investigates (1) which features of smartphone usage log data might be used to predict momentary subjective stress (i.e., *digital biomarkers*), (2) whether these features are positively or negatively related to momentary subjective stress at both the group and individual level, and (3) if individuals differ in these relationships. Following machine and deep learning research on mood prediction [19-21], we also assessed how important digital biomarkers are relative to temporal features: time of day, day of the week, day of the month, and before COVID-19 versus during COVID-19 lockdown.

Digital biomarkers

Digital biomarkers are digital footprints, such as features of smartphone usage log data, that are related to an underlying biological state [22], such as stress[1]. Such features might represent any quantification of raw log data of digital devices, ranging from simple (e.g., time spent using a device) [23] to more complex (e.g., daily life patterns derived from device use) [24-25]. In this study, we investigate two types of potential digital biomarkers: (1) usage of different types of smartphone applications (e.g., duration and frequency of Social Networks, Messengers, and Video applications) and (2) sleep proxies derived from smartphone application log data.

Smartphone Use Behavior. Smartphones enable individuals to perform a wide range of behaviors with profound psychological meaning and relevance to momentary subjective stress. One smartphone-mediated behavior with particular relevance to stress is social interaction. Smartphone users can socially interact through phone calls (e.g., Viber; Whatsapp; default dialing application), direct messaging (e.g., Whatsapp, Facebook Messenger, Telegram, Signal), and social media (e.g., Facebook, Instagram). Recent work on passively logged data has demonstrated social application use is associated with depression, with depressed individuals using social applications more frequently than non-depressed individuals [26]. We believe logging such behavior is also useful in the present context as social interaction is related to stress [27-28]. Altogether, an accurate measure of social behavior could be informative about stress. We therefore test whether social application use is predictive of stress.

Furthermore, other application categories such as video streaming and gaming applications

1 Montag et al. [22] distinguish between direct and indirect digital biomarkers. A **direct** digital biomarker has a relationship with an underlying biological state (e.g., psychophysiological stress), whereas an **indirect** digital biomarker has a relationship with a proxy for this state (e.g., self-reported stress). As we did not measure psychophysiological stress but used a self-report methodology (i.e., experience sampling), our study focuses on identifying **indirect** digital biomarkers.

might be predictive of stress. For instance, evidence suggests that approximately one fifth of college students binge watch videos to de-stress [29]. We therefore explore whether applications other than messengers and social media are predictive of stress.

Sleep. Because the human sleep-wake cycle closely aligns with smartphone application usage patterns [30], different disciplines have leveraged smartphone log data (e.g., call records, screen on-off status, screen taps) to estimate sleep-related variables, such as sleep onset, offset, and duration [24-25,31]. Such sleep proxies could potentially be important for stress recognition, as sleep is a universal behavior that is related to stress [32]. We therefore applied a rule-based algorithm (similar to recent work [24-25]) to extract proxies for sleep duration and sleep onset from smartphone application log data.

Explainable Artificial Intelligence

One potential avenue to identify digital biomarkers is to (1) train machine learning models to find a mathematical mapping from digital biomarkers to self-reports of momentary subjective stress and (2) apply explainable artificial intelligence (XAI; e.g., Shapley values [33]) to clarify how these models make predictions. Applying XAI is necessary because the structure of some powerful models (e.g., random forest) preclude the straightforward interpretation of parameters that linear statistical models have.

By taking this approach, we aim (1) to identify digital biomarkers by testing which features of smartphone usage log data a model uses to predict momentary subjective stress and (2) to understand the nature of the relationship between these digital biomarkers and stress. For instance, time spent on Messenger applications might be important for the prediction of stress and negatively related to stress, suggesting individuals tend to spend less time on these applications when feeling stressed. This could potentially indicate that stress reduces social interaction or *vice versa*.

Nomothetic Versus Idiographic

When training machine learning models on human-subjects data, it is important to take into account that individuals differ from another. Hence, both machine learning [19, 34] and behavioral scientists [35-37] have underlined the importance of ensuring that models of human-subjects data are applicable to the individuals they pertain to. This is certainly also important in the domain of digital biomarkers of stress. Behavioral scientists have shown relationships between digital trace data and psychological self-reports differ across individuals [23]. In parallel, machine learning researchers have demonstrated personalized models – which are known as idiographic models in behavioral science – tend to predict subjective stress more accurately than non-personalized models [34] – also referred to as nomothetic models. Naturally, these findings go hand in hand: A personalized (idiographic) model will more adequately capture person-specific dependencies between digital trace data and stress and therefore should make more accurate predictions than a non-personalized (nomothetic) model. In this work we implement, evaluate, and compare both approaches.

Objectives

This study has three complementary aims, namely to determine 1) if features of smartphone usage log data might be *digital biomarkers* that can be used to predict momentary subjective stress, 2) if these biomarkers are positively or negatively related to momentary subjective stress (at the group and individual level), and 3) how accurately these potential digital biomarkers can be used to recognize momentary subjective stress on a person-by-person basis in out-of-sample data.

Methods

Participants

We follow reporting guidelines recommended for experience sampling studies [38]. For a preregistered data collection (AUTHORS), we used the university participant pool to recruit 247 student participants, 224 of whom we included for analysis. Their average age was 21.97 (*SD* = 3.04) years and the majority were female (52.47%). We excluded (1) participants with operating systems other than Android on their primary phone and (2) participants with insufficient survey responses for training idiographic machine learning models (less than six).

Procedure

Ethical approval was issued by the [institution name blinded] Ethics Committee (approval code [blinded]). Participants were recruited through the university participant pool. After receiving online information through Qualtrics, having been offered a possibility to ask questions, and signing an informed consent form, participants followed online instructions to install two applications on their smartphone. After completing these instructions, participants attended an onboarding session in which we provided additional information, offered further opportunity to ask questions, and motivated participants to participate to the best of their ability.

Technology. All participants installed two applications on their Android device: Ethica Data [39] and mobileDNA [40]. Ethica Data is an application that prompts participants to complete brief surveys on their smartphone (i.e., *experience sampling*). MobileDNA is an application that unobtrusively logs a person's smartphone application use, smartphone notifications, and location (i.e., *passive logging*).

Sampling scheme. In a four-month period, Ethica notified participants five times a day for a maximum of 60 days (30 days in month one; 30 days in month four) at pseudo-random times between 8:30 and 22:30 to complete a 10-item survey (approximately one minute to complete) on stress and other constructs (fatigue, procrastination, mood), while mobileDNA continuously logged smartphone application use. Following an initial push notification, each survey was available to the participant for 50 minutes. After 45 minutes, they received a reminder notification. After 50 minutes, the survey expired. Participants were allowed to catch up on one missed survey per day by starting and completing a new survey.

Monitoring Protocol. We actively monitored participant compliance and motivated participants with weekly emails containing personalized feedback. When a participant failed to complete many consecutive surveys, we sent an email to inquire why they could not comply with the study protocol and how any issues might be resolved. In a limited number of cases, participants did not respond to such emails, in which case we contacted them through a phone call. Participants were compensated with course credits for research participation and were entered into a raffle comprising twenty prizes of 15 euro.

Compliance. Participants (N = 224) completed a total of 44,381 surveys (198 per person on average). Though data collection spanned the introduction of COVID-19, lockdown, the majority of participants remained in the study across four months, with a median study adherence of 222 surveys (74.00%; SD = 60.24). Reasons for noncompliance ranged from technical difficulties (e.g., not receiving any notifications; broken or lost smartphone) to not being able to complete a survey (e.g., waking up too late; receiving a notification during work or lecture) to personal reasons (e.g., attrition due to COVID-19-related personal problems or collecting sufficient course credits). The median time difference between receiving the initial notification and completing the survey was 7.27 minutes (SD = 14.08). In the sample we analyzed, the median participant completed t = 205 surveys (SD = 87.42) and had 3,072 hours of smartphone usage log data.

Measures

Stress Experience Sampling Scale-2 (Str-ESS-2). We used the Str-ESS-2 scale to measure participants' current (i.e., in-the-moment) subjective level of stress (AUTHORS). Str-ESS-2 consists of two items on a 7-point Likert scale ranging from 1 ("Not at all") to 4 ("Moderately) to 7 ("Very much"): "Right now, I feel relaxed", and "Right now, I feel stressed (tense, restless, nervous or anxious)". The two items have an adequate intra-class coefficient (>50% of variance due to within-person fluctuations), and high within-person reliability (ω = 0.95). We calculated an unweighted average of the two items and within-person centered the resulting values.

Passively Logged Features. Table 1 provides an overview of all features included in this study. We analyzed three categories of features: (1) smartphone use behavior, (2) sleep, and (3) time. Because the raw values of these features are on vastly different scales, which can dramatically impact model performance, we scaled all features to a range between 0 and 1 using MinMaxScaler in *sklearn*, based on the minimum and maximum values in the training data.

Table 1. Overview of features included in models.

Feature category	Features
Smartphone use behavior	Duration (seconds) spent on smartphone application category X in the past 60 minutes Frequency (count) of opening smartphone application category X in the past 60 minutes
Sleep	Sleep onset (hours, post-midnight hours > 24) Sleep duration (hours)
Time	Hour of day (0 to 23, starting at midnight) Day of week (0 = weekday, 1 = weekend) Day of month (0 to 31) COVID-19 (pre-COVID-19 = 0, during COVID-19 = 1)

***Smartphone Use Behavior*.** We categorized applications using a coding scheme that maps application names (e.g., com.whatsapp) to one of 18 major categories (see Table 2; these core categories represent 78.38% of all application events). Then, during the 60 minutes before each stress observation, we calculated 1) the total time spent on all applications in a category (*duration*) and 2) the total number of times applications in a category had been accessed (*frequency*).

***Sleep*.** We extracted sleep duration and sleep onset from the raw smartphone application log timestamps. The algorithm we used is similar (but not identical) to a previously validated algorithm [21-22] (see Online Supplementary Materials for a description of our approach), an algorithm that was found to be strongly associated with actigraphy-based and self-reported sleep duration and onset.

Time. We used pandas.datetime in Python 3.9.9 to extract time features (i.e., hour of day, day of week, day of month, and lockdown status) from the raw self-report timestamps. In the interest of model simplicity, we recoded day of the week into a binary variable (weekday = 0, weekend = 1) rather than treating this variable as a categorical variable. We further coded lockdown status as a binary variable (before COVID-19 lockdown = 0, during COVID-19 lockdown = 1), based on the lockdown timing in the country of the study.

Table 2. Smartphone application categories and example applications

Category	**Examples**
Browser	Chrome, Opera
Calling	Default dial applications
Camera	Default camera applications
Dating	Tinder, Grindr
E-mail	Gmail, Outlook
Exercise	RunKeeper
Food & Drink	UberEATS
Gallery	Default gallery applications
Game	CandyCrush
Messenger	Whatsapp
Music & Audio	Spotify
Productivity	Microsoft Word
Shared transportation	9292OV (Dutch public transport)
Social networks	Facebook, Instagram, Twitter
Video	Youtube, Netflix
Weather	Default weather applications
Work	StudentJob, EmployeeApp

Stress Recognition Models

We trained three different types of machine learning models: LASSO regression (LASSO), support vector regression (SVR), and random forest regression (RF). We trained all models using a nomothetic as well as an idiographic approach (explained below). For brevity and clarity, we prepend an "N" to the abbreviations of our nomothetic models (i.e., N-LASSO, N-SVR, N-RF) and an "I" to the abbreviations of our idiographic models (i.e., I-LASSO, I-SVR, I-RF).

Cross-validation strategies

Generally, when we perform cross-validation, we (1) split the data into train and test data, (2) specify a range of values for a model's different hyperparameters (i.e., researcher-specified parameters that control model complexity), (3) identify the best hyperparameters per model using *k*-fold cross-validation within the training set (4) evaluate each model based on predictive accuracy for the test data. In what follows, we outline how we cross-validated nomothetic and idiographic models.

Nomothetic models. We applied a user split to distinguish between train and test data. We first used group *k*-fold cross-validation (GroupKFold in *sklearn*) to partition the data into five subsets. Four subsets contained data from 45 participants, and one subset contained data from 44 participants. We then selected four subsets (*train data*) to train a model. For training the model, we used five-fold grid search cross-validation (GridSearchCV in *sklearn*) to minimize each model's default *sklearn* error metric (see Online Supplementary Materials for tuned hyperparameters and minimized error metrics). After training the model, we let the model make predictions on the data subset we did not include in training (*test data*; i.e., all observations of participants excluded from training). Finally, we evaluated the accuracy of these predictions. We repeated this process until each subset of the data had been left out of training once and did so for all models.

Idiographic models. We applied a time split to distinguish between train and test data. We iteratively selected one participant's data to train and test models only on these data. For each person, (1) we assigned each participant's first 80% observations to a train dataset and their final 20% observations to a test dataset, and (2) trained each model (SVR, RF, LASSO). We applied five-fold grid search cross-validation to each participant's train data to optimize hyperparameters[2]. (3) Finally, we let trained models make predictions on the individual's test data (i.e., final 20% observations) and evaluated the accuracy of these predictions.

2 Because the number of idiographic models to train is much larger, we applied five-fold randomized search cross-validation (CV) rather than grid search CV for training random forests, which are more computationally expensive than LASSO and SVR. Grid search CV always uses a larger number of hyperparameters than randomized search CV because the latter trains models on a random subset of all the hyperparameter settings used by the former. Randomized search CV considerably speeds up training time and makes training a large number of random forest models more feasible.

Model evaluation

Spearman rho rank-order correlation. Spearman rho rank-order correlation indicates whether a model tends to predict greater values when an individual feels more stressed without assuming a linear relationship. We consider a model to perform above-chance for a given individual if the Spearman rho between predictions and self-reports between predictions and self-reports is sufficiently large that the corresponding p-value is below .05.

Mean absolute error (MAE). Lower MAE values indicate a more accurate model. We consider the MAE to be an intuitive metric to assess predictive accuracy on a target variable measured on a 7-point Likert scale, as it allows us to make statements such as “on average, the model mispredicts momentary subjective stress by ±0.80 points on a 7-point Likert scale)”. To evaluate if our models perform better than random guessing, we compare against the MAEs of a “naive” but person-specific baseline model. This “naive” baseline model always predicts an individual’s average level of stress.

Model explanation

One of the challenges of machine learning approaches is to understand the results - as they are more complex and therefore less intuitive than standard statistical approaches. To explain models, we use the SHAP library implemented in Python 3.9.9 to calculate and visualize Shapley values. Shapley values can be used to determine (1) which features are most important in a model and (2) how features are related to model predictions [30].

Results

Recognizing Momentary Subjective Stress

Nomothetic models. Table 3 provides an overview of how accurately nomothetic models predict out-of-sample data. Model predictions correlated positively and significantly with self-reports in the majority of participants for N-LASSO and N-RF, with N-RF performing best in terms of %

significant results (55.80%). The median correlation between predictions and self-reports was weak for all models (between .15 and .18). Correlations also differed between participants: in 60 participants, the positive correlation was moderate or larger ($\rho > 0.3$), while being negative ($p < .05$) in 5 people (2.23%). In the median participant, models on average mispredicted momentary subjective stress by approximately 0.8 points on a 7-point Likert scale (MAE of .84; scale range: 1 "Not at all", 4 "Moderately", 7 "Very much"). The MAE of nomothetic models varied across individuals but for 89.28% of participants the person-specific baseline outperformed all nomothetic models (see sixth column in Table 3) followed by N-LASSO (8.93%) and N-SVR (1.79%). N-RF did not outperform the other models in terms of MAE.

Thus, nomothetic models make predictions that weakly and positively correlate with actual stress self-reports for the majority of participants. This means that when these participants, who were not included in the training data set, feel stressed, the model tends to output a higher value, and when a participant does not feel stressed, the model tends to output a lower value. These models are not highly accurate however, as they often do not outperform a person-specific baseline. For most participants, the association between predictions and actual self-reported stress is positive, but it is significantly negative for a very small group ($n = 5$). Thus, for a very small group of participants, for whom the model tends to detect subjective stress when the person does not feel stressed, and *vice versa*.

Idiographic models. We next evaluated how well idiographic models recognize momentary subjective stress in out-of-sample observations. For all models, model predictions of momentary subjective stress correlated positively and significantly with self-reports of momentary subjective stress, but only in a minority of participants. In 60 participants, this correlation was moderate or larger ($\rho > 0.3$), whereas the overall median correlation was weak (range of median correlations per model, ρ = [0.05, 0.1]). In a minority ($n = 9$), model predictions of stress and self-reported stress were significantly negatively associated. Similar to the nomothetic models, the median person-specific MAE for each idiographic model was slightly above 0.8 points. These MAE scores are compared to a person-specific baseline based on the person-specific average level of stress in the participant's train data (i.e., first 80%) rather than all their data to prevent data leakage from train to test data. MAE varied from individual to individual, but for 80.63% of participants at least one of the idiographic models outperformed the person-specific baseline. The best model was most often the I-SVR, but both I-RF and I-LASSO were the best models for a sizable proportion of people.

Table 3. Central tendency and range of out-of-sample predictive accuracy (Spearman rho correlation and mean absolute error) for nomothetic and idiographic models on a person-by-person basis.

	Spearman rho rank-order			**Mean absolute error (MAE)**			
Model	**Median**	**Range**	**% significant**	**Median**	**Range**	**% better than baseline**	**% best model**
N-baseline	-	-	-	**0.83**	[0.00, 3.39]	-	**89.28**
N-LASSO	0.15	[-0.45, 0.65]	51.79	0.84	[0.11, 2.04]	10.27	8.93
N-SVR	0.17	[-0.31, 0.62]	50.89	0.84	[0.17, 2.04]	4.91	1.79
N-RF	**0.18**	[-0.41, 0.56]	**55.80**	0.84	[0.25, 2.04]	3.57	0.00
I-baseline	-	-	-	**0.83**	[0.00, 3.40]	-	19.64
I-LASSO	0.08	[-0.83, 1.00]	14.29	0.85	[0.00, 3.36]	34.82	23.21
I-SVR	0.05	[-0.79, 1.00]	18.30	0.85	[0.00, 3.50]	**47.32**	**37.95**
I-RF	**0.10**	[-1.00, 1.00]	**20.54**	0.84	[0.00, 3.26]	41.52	19.20

Note. N = nomothetic, I = idiographic. Bolded values represent best performance.

Identifying Digital Biomarkers of Momentary Subjective Stress

To demonstrate which aspects of the model have the strongest predictive value, Figure 2 provides an overview of the ten most important features per nomothetic and idiographic model. In all six models, temporal features, COVID-19, Messenger and Social Network use, and sleep proxies were most important to the prediction stress. The biggest disagreement between nomothetic and idiographic models was the importance of Messenger and Social Network use relative to sleep duration and onset. The former features were more important in nomothetic models, whereas the latter were more important in idiographic models. Feature importance was relatively consistent for models across data splits (see Online Supplementary Materials for beeswarm plots per model per data split).

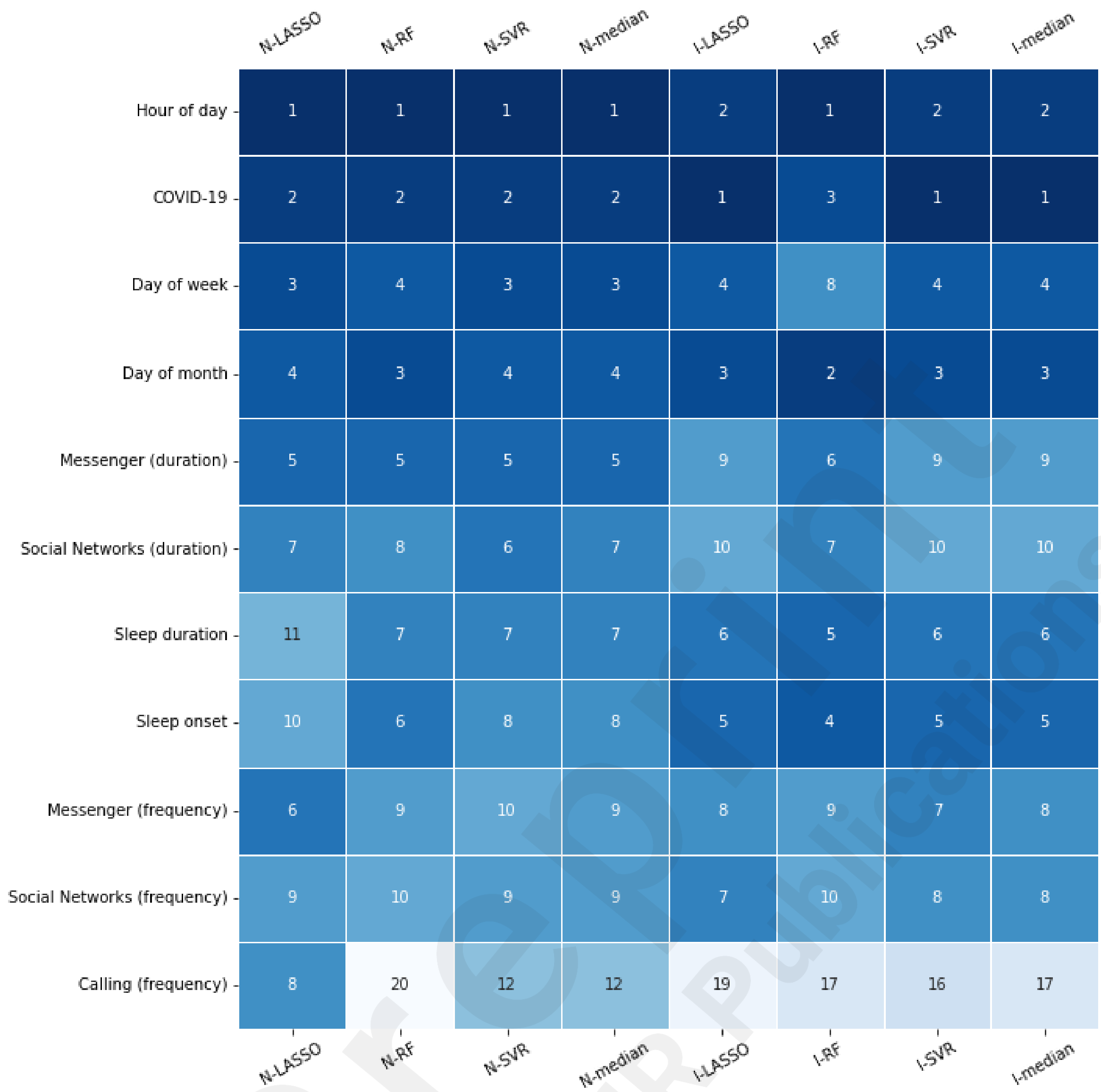


Figure 2. Feature importance ranking for each nomothetic and idiographic model, containing features that appeared in the top 10 feature of any model. Numeric values represent the ranking of one feature for one model. Dark (top) cells represent more important features. N-Median and I-Median represent the median ranking of a feature across nomothetic and idiographic models, where double values indicate a tie between two features. Features are ordered by N-Median scores.

Understanding Digital Biomarkers of Momentary Subjective Stress

Feature importance provides relevant information about which features contribute most to predictions. However, it does not tell us whether small or large values of a feature are indicative of momentary subjective stress. For instance, although we have identified hour of the day as a relatively important predictor of momentary subjective stress, it is still unclear whether people tend to feel more stressed in earlier or later hours of the day. To clarify potential relationships between features and momentary subjective stress, we present a beeswarm plot for the nomothetic random forest (Figure 1; see Online Supplementary Materials for all beeswarm plots) to visualize how different features are related to model predictions in one split of the data. For instance, Figure 1 shows N-RF predicts greater momentary subjective stress values (1) in earlier hours of the day (indicated in blue), (2) during COVID-19 lockdown, (3) on weekdays, and (4) later in the month. Similarly, when individuals spend more time on Messenger (red), these models output greater values for momentary subjective stress.

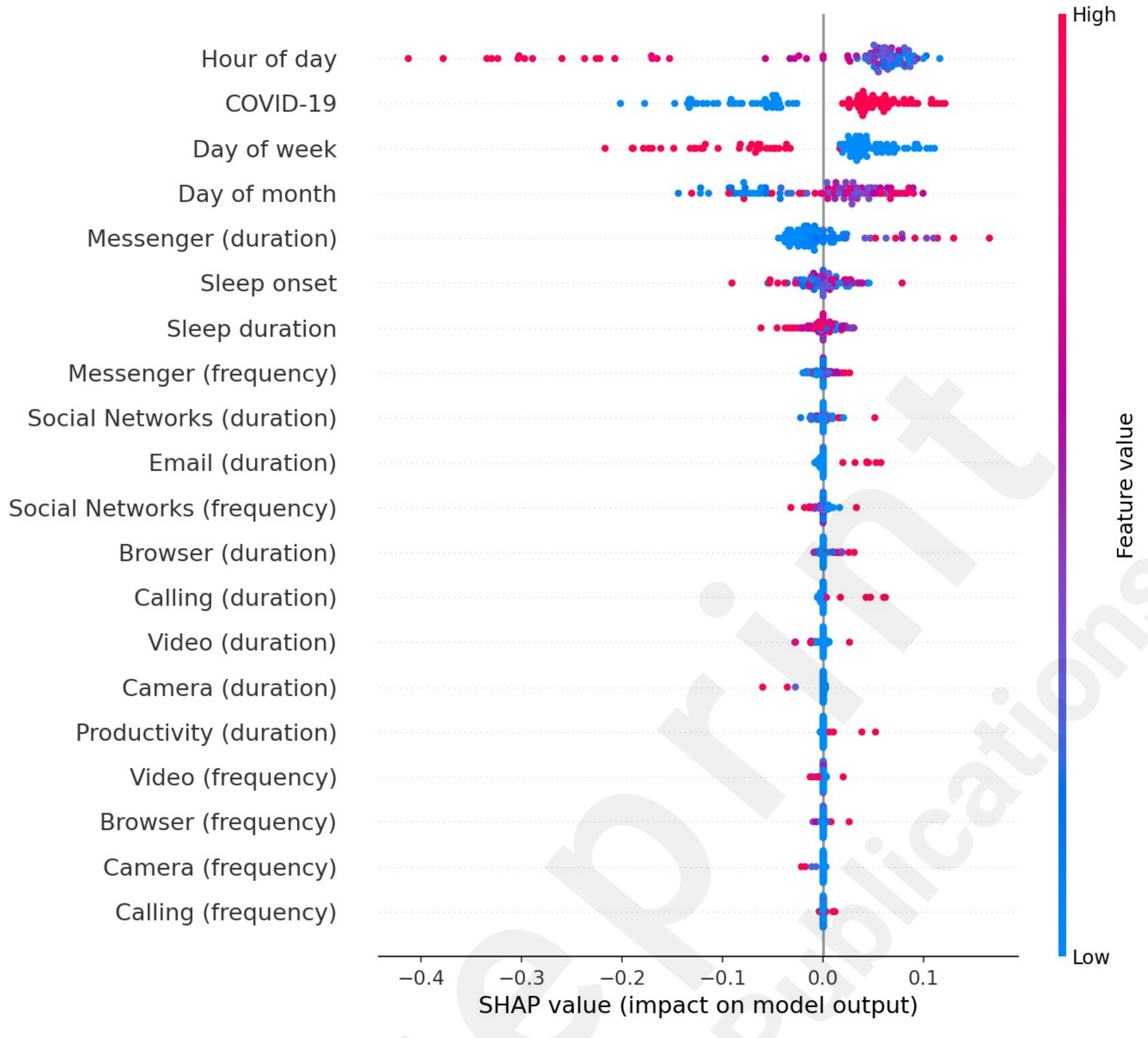


Figure 2. SHAP beeswarm plot indicating the relative importance of each feature and the relationship between feature values and model prediction for the nomothetic random forest model (N-RF). For each feature (listed on the y-axis in decreasing order of importance) a point represents one test trial, the color of a point indicates the value of the feature (red for higher values [e.g., later hour of the day], blue for low values [e.g., earlier hour of the day], and the position along the x-axis indicates the SHAP value (positive values correspond with higher stress predictions, larger magnitude values indicate stronger impact). For any one feature, red points on the left side of the plot (high feature values and negative SHAP values) indicate a relationship where increasing the feature value results in the model predicting lower outcome values (e.g., higher hour predicts lower stress), while red points on the right side of the plot (high feature values and positive SHAP values) indicate a positive relationship (e.g. covid lockdown predicts higher stress). COVID-19 is coded as 0 = pre and 1 = post.

Weekday is coded as 0 = weekday, 1 = weekend day.

Interindividual differences in the relationship between digital biomarkers and stress

We also investigated whether the nature of relationships between features and stress differed from person to person. Because the zero-order Spearman rank-order correlations between features and stress are relatively weak (see Online Supplementary Materials), we instead calculated these correlation between feature values and Shapley values for each idiographic model. A positive correlation between feature and Shapley value indicates that when this feature has a higher value, the model predicts that an individual feels more stressed. Shapley values of more complex models have the potential benefit of capturing the non-linear and interactive relationships learned by the model. Correlations with *p*-values above .05 are not included.

Figure 3 shows the frequency of significant positive and negative correlations between digital biomarkers and predicted stress for each ideographic model. Interestingly, most bars are either mostly red or mostly blue, which indicates a relatively homogeneous relationship across people between that feature and stress for most features. Notably, the relationship is more varied between stress and some of the most important features (see Figure 1), including Messenger use, day of the month, and even COVID-19 lockdown. A few features rarely show any correlation with predicted stress, and this is likely because that model has learned to disregard those features or because the correlation between feature values and SHAP values is either not reliable or monotonic.

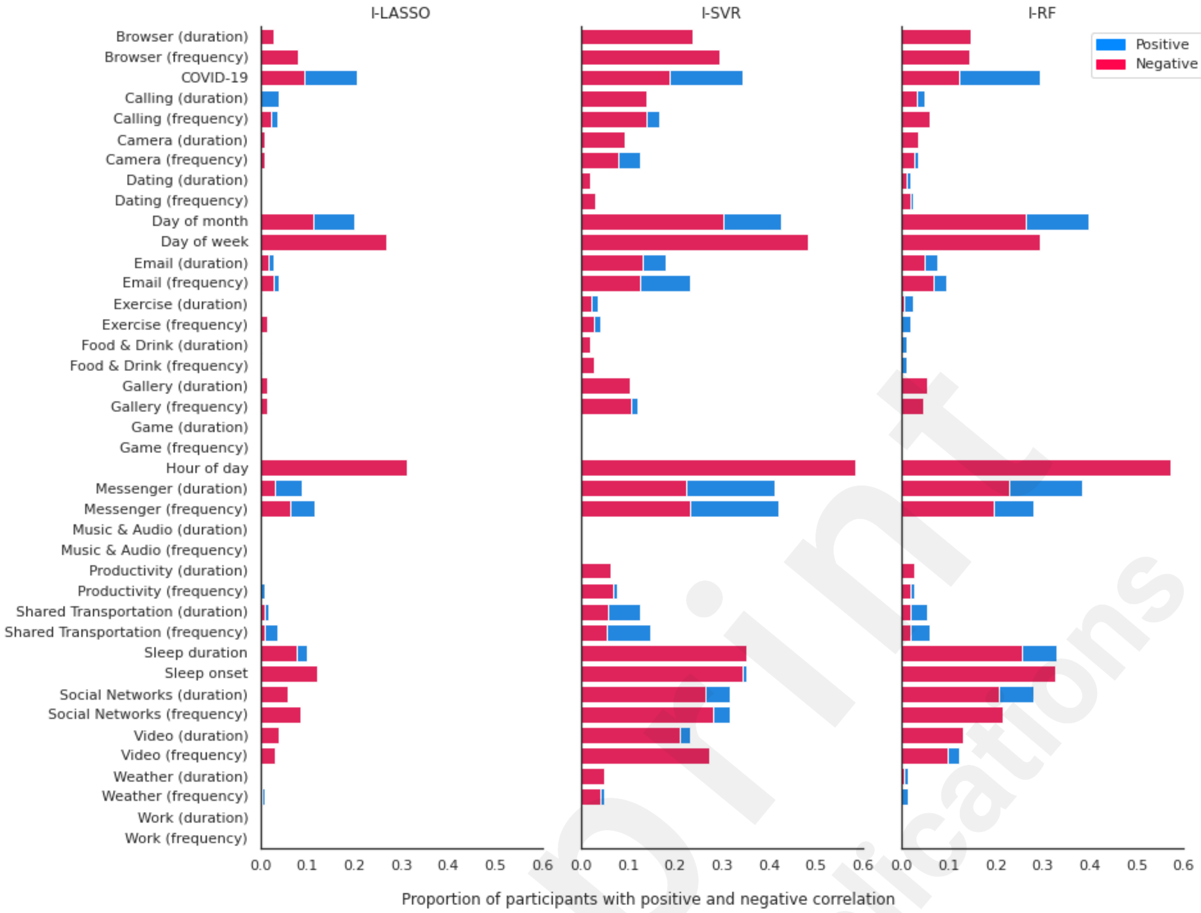


Figure 3. Stacked bar plots representing the proportion of participants showing a significant positive (blue bar) or negative (red bar) correlation between feature values and Shapley values for each idiographic model. A negative correlation indicates that when a given feature has a higher value, then the model predicts that an individual feels more stressed. For instance, the I-RF predicts a higher level of stress at later hours of the day in approximately 60% individuals.

Discussion

Principal Findings

The three aims of our study were to test 1) if features of smartphone usage log data might be *digital biomarkers* that can be used to predict momentary subjective stress, 2) if these biomarkers are positively or negatively related to momentary subjective stress (at the group and individual level), and 3) how accurately these potential digital biomarkers can be used to recognize momentary subjective stress on a person-by-person basis in out-of-sample data.

Using explainable artificial intelligence, we found that temporal features, prolonged Messenger and Social Media application use, and smartphone-tracked sleep proxies were the most important features of the best performing nomothetic and idiographic models. These models consistently ordered these features, with temporal features as most important and application usage as less important drivers of stress predictions. The biggest disagreement between nomothetic and idiographic models is the importance of application usage duration. In nomothetic models, these are more important than sleep features, but in idiographic models, the order is opposite. In sum, though, prolonged use of Messenger applications and sleep proxies might be valid digital biomarkers of stress.

We found that when individuals were more stressed, nomothetic models tended to predict a higher level of stress (and *vice versa*) in the majority of individuals. However, they generally did not predict stress with greater accuracy than a naive baseline model that always predicted a person was experiencing their average level of stress. We found these results to be similar for idiographic models, although these models predicted stress with greater accuracy than a naive baseline model. Even though performance should be improved to make clinical application feasible, this study does suggest that, in the absence of self-report or physiological data, digital biomarkers can be used to recognize momentary subjective stress on a person-by-person basis in out-of-sample data.

Our results suggest that, for the average person, in the earlier hours of the day, on weekdays, on later days of the month, and during COVID-19 lockdown (compared to before COVID-19 lockdown), individuals felt more stressed than they would usually do. Individuals also felt more stressed when spending more time on social applications (i.e., Messengers, Social Networks), having slept more briefly, and falling asleep later. The relationship between two feature categories (i.e., temporal features and sleep proxies) and (SHAP values for the prediction of) stress was rather consistent across individuals, suggesting these are universal principles for this population to monitor change in a person's momentary stress levels. These may even be explained by biological mechanisms (e.g., potentiation of the hypothalamic-pituitary-adrenal [HPA] axis due to sleep deprivation [41]; cortisol awakening response [CAR] [42]) that universally affect adolescents and young adults. Such features are likely useful for passive tracking in the context of stress-related common mental disorders such as depression and burnout.

One of the unique features of this study was to compare nomothetic and idiographic models in light of increasingly idiographic research practices in behavioral science [43]. Our findings show that model personalization is warranted especially when smartphone app usage features are added to the model. That is, the relationship between social applications usage and stress (see Figure 2) was negative for some individuals and positive or absent for others. We therefore find evidence that the idiographic approach of digital phenotyping research – i.e., “the moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices, in particular smartphones” [44] – is warranted in the context of stress.

In clinical practice, the added value of digital phenotyping of stress is that it might help us to understand how, for a given individual, stress is related to temporal features, how sleep impacts their stress levels, and as a probe for qualitative investigation of how they respond to stress. For instance, if an individual spends less time on Messenger applications when stressed, this could suggest they avoid social contact, whereas seeking social support might be beneficial. If individuals are not aware of this pattern, personalized prediction models provide novel clinical insights and could potentially help to improve therapy outcomes (e.g., within personalized treatment modules [45]).

Limitations

Our study should be viewed in light of the following limitations. First, it is conceivable not all students self-reported stress accurately at every assessment. A significant proportion of variance might therefore represent noise that cannot be explained by any variable or model, irrespective of modeling decisions. This is especially an issue for idiographic models that rely on the participants' final observations, which might be observations of lower quality due to study fatigue [46].

Second, we applied within-person mean-centering to the self-reported stress. Although this corrects for differences in how people use a scale, it only allows nomothetic models to predict whether a person is currently experiencing more or less stress than usual and prevents models from predicting whether this person's stress level is very low or high relative to other people's stress level.

Third, we forced nomothetic models to learn one mapping from features to outcome for all individuals in our sample. This is problematic because such models learn one set of parameters that might be accurate for some individuals but highly inaccurate for others (i.e., *one-size-fits-all* fallacy) [47-48]. Truly idiographic models, which we also trained in this study, do not suffer from this issue by default. However, this comes at the cost of strongly reduced sample size, which limits model complexity and may lead to overfitting. As collecting more self-report data per individual is not feasible for samples of this size, future studies could 1) focus on smaller samples with exceptionally motivated participants for a longer sampling period, 2) use wearables to measure psychophysiological signals of stress (e.g., CortiWatch [49]), or 3) train machine learning models on a full dataset without losing sight of interindividual differences in feature-outcome relationships (e.g., using transfer learning [50]).

Conclusions

Our study has three main conclusions. First, temporal features, sleep proxies, and prolonged use of Messenger and Social Network applications are consistently identified as the most important digital biomarkers for predicting momentary subjective stress. Second, for the majority of people, these markers are not sufficiently informative to recognize momentary subjective stress with appreciable improvements accuracy over a baseline model, but do produce predictions that correlate with subjective stress. Third, the utility and relationship with stress of (some) digital biomarkers varies from person to person. On the one hand, one digital biomarker may be relevant to momentary subjective stress in one individual, but not in another. On the other hand, the increase of one digital biomarker might imply lower stress for one individual and higher stress for another. Our study thus provides evidence for phenotypic heterogeneity in the relationship between how we feel and the digital traces we leave behind. These findings are relevant for the implementation of algorithms in mHealth and uHealth applications to prevent, monitor, and treat stress-related mental disorders.

References

1. Maddock C, Pariante C. How does stress affect you? An overview of stress, immunity, depression and disease. *Epidemiol Psichiatr Soc*. 2001;10(3):153-162. doi:10.1017/s1121189x00005285
2. Menke A. Is the HPA axis as target for depression outdated, or is there a new hope?. *Frontiers in psychiatry*. 2019;10:101. doi:10.3389/fpsyt.2019.00101
3. Rohleder, N. Stress and inflammation–The need to address the gap in the transition between acute and chronic stress effects. *Psychoneuroendocrinology*. 2019;105:164-171. doi:10.1016/j.psyneuen.2019.02.021
4. Fukazawa Y, Ito T, Okimura T, Yamashita Y, Maeda T, Ota J. Predicting anxiety state using smartphone-based passive sensing. *Journal of Biomedical Informatics*. 2019;93:103151. doi:10.1016/j.jbi.2019.103151
5. Ferdous R, Osmani V, Mayora O. Smartphone app usage as a predictor of perceived stress levels at workplace. In: *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE; 2015:225-228. doi:10.4108/icst.pervasivehealth.2015.260192
6. Sano A, Phillips AJ, Amy ZY, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE:2015:1-6. doi:10.1109/BSN.2015.7299420
7. Bernstein EE, McNally RJ. Acute aerobic exercise hastens emotional recovery from a subsequent stressor. *Health Psychology*. 2017;36(6):560-567. doi:10.1037/hea0000482
8. Kabat-Zinn J. Mindfulness-based stress reduction (MBSR). *Constructivism in the Human Sciences*. 2003;8(2):73-83.

9. Williams R, Hopkins S, Frampton C, Holt-Quick C, Merry SN, Stasiak K. 21-Day Stress Detox: Open Trial of a Universal Well-Being Chatbot for Young Adults. *Social Sciences*. 2021;10(11):416. doi:10.3390/socsci10110416
10. Bogomolov A, Lepri B, Ferron M, et al. Daily stress recognition from mobile phone data, weather conditions and individual traits. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM; 2014:477-486. doi:10.1145/2647868.2654933
11. Jaques N, Taylor S, Sano A, Picard R. Multi-task, multi-kernel learning for estimating individual wellbeing. In: *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*. NIPS: 2015; 898:63.
12. Maxhuni A, Hernandez-Leal P, Morales EF, Sucar, L. E., Osmani, V., & Mayora, O. (2020). Unobtrusive stress assessment using smartphones. *IEEE Transactions on Mobile Computing*, *20*(6), 2313-2325. doi:10.1109/TMC.2020.2974834
13. Sano, A., & Picard, R. W. (2013, September). Stress recognition using wearable sensors and mobile phones. In: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE: 2013; 671-676. doi:10.1109/ACII.2013.117
14. Shaw A, Simsiri N, Deznaby I, et al. Personalized Student Stress Prediction with Deep Multitask Network. *arXiv*. Preprint posted online June 26, 2019. arXiv:1906.11356
15. Jaques N, Taylor S, Sano A, & Picard R. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In: *IJCAI 2017 Workshop on artificial intelligence in affective computing.* IJCAI: 2017; 17-33.
16. Li B, Sano A. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. ACM: 2020; 1-26. doi:10.1145/3397318
17. Umematsu T, Sano A, Taylor S et al. Forecasting stress, mood, and health from daytime

physiology in office workers and students. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE: 2020; 5953-5957. doi:10.1109/EMBC44109.2020.9176706

18. Yu H, Klerman EB, Picard RW, et al. Personalized wellbeing prediction using behavioral, physiological and weather data. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE: 2019; 1-4. doi:10.1109/BHI.2019.8834456
19. Bachmann A, Klebsattel C, Budde M, et al. How to use smartphones for less obtrusive ambulatory mood assessment and mood recognition. In: *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM: 2015; 693-702. doi:10.1145/2800835.2804394
20. Fisher AJ, Bosley HG. Identifying the presence and timing of discrete mood states prior to therapy. *Behaviour Research and Therapy*. 2020;128:103596. doi:10.1016/j.brat.2020.103596
21. Suhara Y, Xu Y, Pentland AS. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In: *Proceedings of the 26th International Conference on World Wide Web*. ACM: 2017; 715-724. doi:10.1145/3038912.3052676
22. Montag C, Elhai JD, Dagum P. On Blurry Boundaries When Defining Digital Biomarkers: How Much Biology Needs to Be in a Digital Biomarker? *Frontiers in Psychiatry*. 2021;12. doi:10.3389/fpsyt.2021.740292
23. Aalbers G, vanden Abeele MMP, Hendrickson AT, de Marez L, Keijsers L. Caught in the moment: Are there person-specific associations between momentary procrastination and passively measured smartphone use? *Mobile Media & Communication*. 2022:205015792199389. doi:10.1177/2050157921993896
24. Borger JN, Huber R, Ghosh A. Capturing sleep–wake cycles by using day-to-day smartphone touchscreen interactions. *npj Digital Medicine*. 2019;2(1). doi:10.1038/s41746-019-0147-4

25. Massar SAA, Chua XY, Soon CS, et al. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. *npj Digital Medicine*. 2021;4(1). doi:10.1038/s41746-021-00466-9

26. Wang Y, Ren X, Liu X, Zhu T. Examining the Correlation Between Depression and Social Behavior on Smartphones Through Usage Metadata: Empirical Study. *JMIR mHealth and uHealth*. 2021;9(1):e19046. doi:10.2196/19046

27. daSilva AW, Huckins JF, Wang W, Wang R, Campbell AT, Meyer ML. Daily perceived stress predicts less next day social interaction: Evidence from a naturalistic mobile sensing study. *Emotion*. Published online October 4, 2021. doi:10.1037/emo0000994

28. Aalbers G, McNally RJ, Heeren A, de Wit S, Fried EI. Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General*. 2019;148(8):1454-1462. doi:10.1037/xge0000528

29. Panda S, Pandey SC. Binge watching and college students: motivations and outcomes. *Young Consumers*. 2017;18(4):425-438. doi:10.1108/yc-07-2017-00707

30. Aledavood, T. Temporal patterns of human behavior. [Doctoral dissertation]. Aalto, Denmark: Aalto University; 2017. ISBN: 978-952-60-7724-6

31. Sano A, Chen W, Lopez-Martinez D, et al. Multimodal ambulatory sleep detection using LSTM recurrent neural networks. *IEEE journal of biomedical and health informatics*. IEEE: 2018; 1607-1617. doi:10.1109/JBHI.2018.2867619

32. Fischer D, McHill AW, Sano A, et al. Irregular sleep and event schedules are associated with poorer self-reported well-being in US college students. *Sleep*. 2019;43(6). doi:10.1093/sleep/zsz300

33. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems*. NIPS: 2017; 4768-4777.

34. Taylor S, Jaques N, Nosakhare E, et al.. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*. IEEE: 2017; 200-213. doi:10.1109/TAFFC.2017.2784832

35. Bülow A, Keijsers L, Boele S, , et al. Parenting adolescents in times of a pandemic: Changes in relationship quality, autonomy support, and parental control? *Developmental Psychology*. 2021;57(10):1582-1596. doi:10.1037/dev0001208

36. Molenaar PCM. A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*. 2004;2(4):201-218. doi:10.1207/s15366359mea0204_1

37. Verbeij T, Pouwels JL, Beyens I, et al. The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*. 2021;3:100090. doi:10.1016/j.chbr.2021.100090

38. Roekel E, Keijsers L, Chung JM. A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *Journal of Research on Adolescence*. 2019;29(3):560-577. doi:10.1111/jora.12471

39. Ethica Data. Accessed February 8, 2022. https://www.ethicadata.com/.

40. Anrijs S, Bombeke K, Durnez W, et al. MobileDNA: Relating Physiological Stress Measurements to Smartphone Usage to Assess the Effect of a Digital Detox. *HCI International 2018 – Posters' Extended Abstracts*. Published online 2018:356-363. doi:10.1007/978-3-319-92279-9_48

41. Minkel J, Moreta M, Muto J, et al. Sleep deprivation potentiates HPA axis stress reactivity in healthy adults. *Health Psychology*. 2014;33(11):1430-1434. doi:10.1037/a0034219

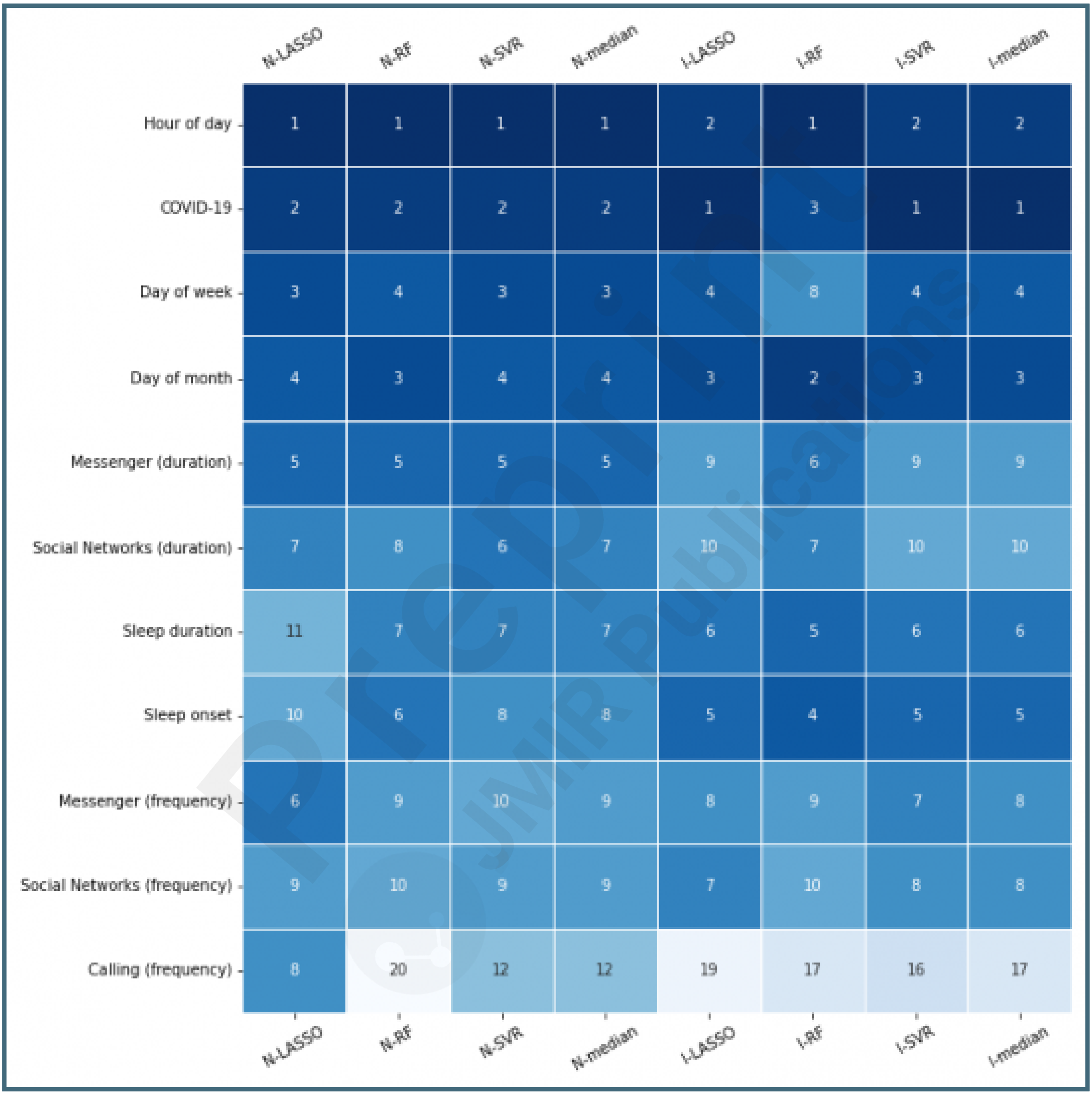
42. Steptoe A, Serwinski B. Cortisol awakening response. In: Fink G, eds. Stress: Concepts, cognition, emotion, and behavior. Cambridge, MA: Academic Press; 2016: 277-283. ISBN:978-0-12-800951-2

43. Beyens I, Pouwels JL, van Driel II, et al. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports*. 2020;10(1). doi:10.1038/s41598-020-67727-7
44. Onnela J-P. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*. Published online July 17, 2020. doi:10.1038/s41386-020-0771-3
45. Bos FM, von Klipstein L, Emerencia AC, et al. PErsonalized Treatment by Real-time assessment (PETRA): A web-application for personalized diaries in psychiatric care. doi:10.31234/osf.io/jh49z.
46. van Berkel N, Goncalves J, Koval P, et al. Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM: 2019; 1-12. doi:10.1145/3290605.3300281
47. Keijsers L, van Roekel E. Longitudinal methods in adolescent psychology: Where could we go from here? And should we? In: Hendry LB, Kloep M, eds. Reframing adolescent research. London, UK: Routledge; 2018: 56-77. ISBN: 9781138557079
48. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*. 2018;115(27):E6106-E6115. doi:10.1073/pnas.1711978115
49. Rice P, Upasham S, Jagannath B, Manuel R, Pali M, Prasad S. CortiWatch: watch-based cortisol tracker. *Future Science OA*. 2019;5(9):FSO416. doi:10.2144/fsoa-2019-0061
50. Woodward K, Kanjo E, Brown DJ, et al. On-device transfer learning for personalising psychological stress modelling using a convolutional neural network. *arXiv*. Preprint posted online April 3, 2020. arXiv:2004.01603

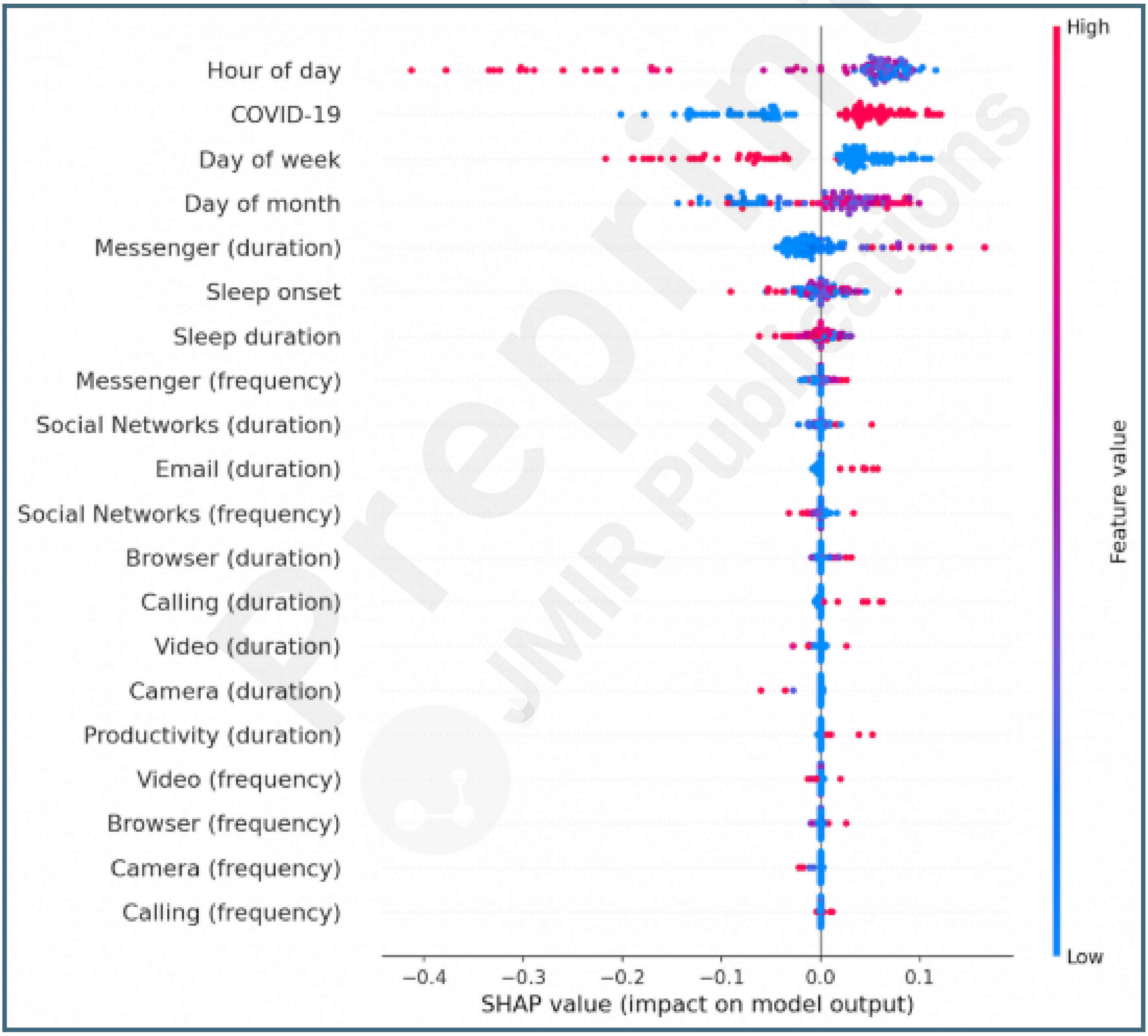
Supplementary Files

Figures

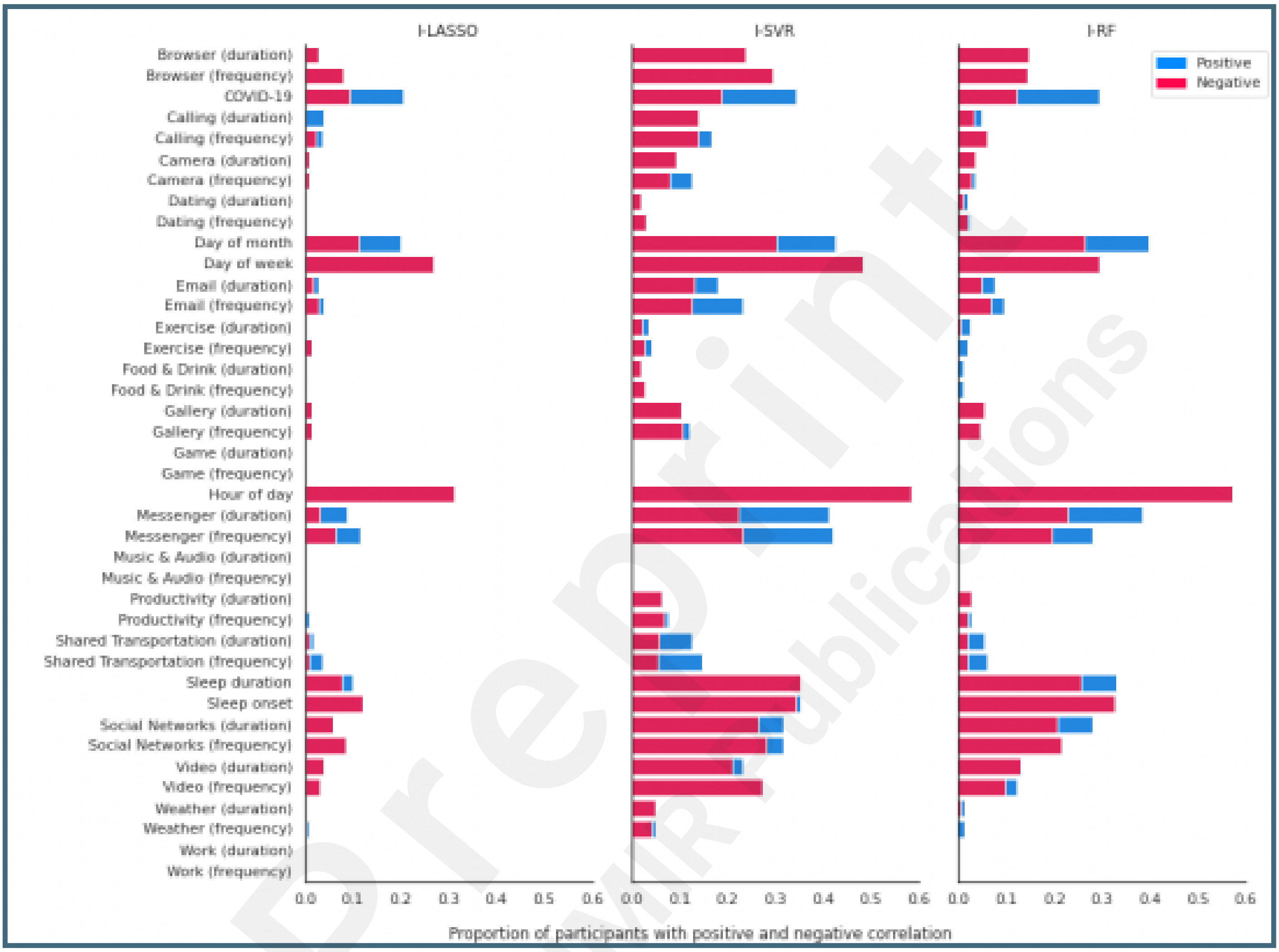
Feature importance ranking for each nomothetic and idiographic model, containing features that appeared in the top 10 feature of any model. Numeric values represent the ranking of one feature for one model. Dark (top) cells represent more important features. N-Median and I-Median represent the median ranking of a feature across nomothetic and idiographic models, where double values indicate a tie between two features. Features are ordered by N-Median scores.



SHAP beeswarm plot indicating the relative importance of each feature and the relationship between feature values and model prediction for the nomothetic random forest model (N-RF). For each feature (listed on the y-axis in decreasing order of importance) a point represents one test trial, the color of a point indicates the value of the feature (red for higher values [e.g., later hour of the day], blue for low values [e.g., earlier hour of the day], and the position along the x-axis indicates the SHAP value (positive values correspond with higher stress predictions, larger magnitude values indicate stronger impact). For any one feature, red points on the left side of the plot (high feature values and negative SHAP values) indicate a relationship where increasing the feature value results in the model predicting lower outcome values (e.g., higher hour predicts lower stress), while red points on the right side of the plot (high feature values and positive SHAP values) indicate a positive relationship (e.g. covid lockdown predicts higher stress). COVID-19 is coded as 0 = pre and 1 = post. Weekday is coded as 0 = weekday, 1 = weekend day.



Stacked bar plots representing the proportion of participants showing a significant positive (blue bar) or negative (red bar) correlation between feature values and Shapley values for each idiographic model. A negative correlation indicates that when a given feature has a higher value, then the model predicts that an individual feels more stressed. For instance, the I-RF predicts a higher level of stress at later hours of the day in approximately 60% individuals.



Multimedia Appendixes

Online supplementary materials.
URL: http://asset.jmir.pub/assets/0fa060452d86a96c7432127f53852744.docx