Using Smartphone App Use and Lagged-Ensemble Machine Learning for the Prediction of Work Fatigue and Boredom

Damien Lekkas[1,4], **George D. Price**[1,4], **Nicholas C. Jacobson**[1,2,3,4]

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College

[2]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College

[3]Department of Psychiatry, Geisel School of Medicine, Dartmouth College

[4]Quantitative Biomedical Sciences Program, Dartmouth College

Abstract

Intro: As smartphone usage becomes increasingly prevalent in the workplace, the physical and psychological implications of this behavior warrant consideration. Recent research has investigated associations between workplace smartphone use and fatigue and boredom, yet findings are not conclusive.

Methods: To build off recent efforts, we applied an ensemble machine learning model on a previously published dataset of $N = 83$ graduate students in the Netherlands to predict work boredom and fatigue from passively collected smartphone app use information. Using time-based feature engineering and lagged variations of the data to train, validate, and test idiographic models, we evaluated the efficacy of a lagged-ensemble predictive paradigm on sparse temporal data. Moreover, we probed the relative importance of both derived app use variables and lags within this predictive framework.

Results: The ability to predict fatigue and boredom trajectories from app use information was heterogeneous and highly person-specific. Idiographic modeling reflected moderate to high correlative capacity ($r > 0.4$) in 47% of participants for fatigue and 24% for boredom, with better overall performance in the fatigue prediction task. App use relating to duration, communication, and patterns of use frequency were among the most important features driving predictions across

lags, with longer lags contributing more heavily to final ensemble predictions compared with shorter ones.

Conclusion: A lag- specific ensemble predictive paradigm is a promising approach to leveraging high-dimensional app use behavioral data for the prediction of work fatigue and boredom. Future research will benefit from evaluating associations on densely collected data across longer time scales.

Keywords

passive sensing; EMA; app use; machine learning; lag; digital phenotyping; fatigue; boredom

1. Introduction

The incorporation of smartphone usage into the workplace has become ubiquitous as it allows for consistent engagement with work-related matters without the necessity of being located in front of a computer. This increase in work engagement via smartphone usage has been argued to benefit workplace communication, cooperation, and the ability to share information in real-time (Kossek & Lautsch, 2012; Lanaj et al., 2014). A 2017 study showed that young Chinese workers rely on smartphones for work-related communication and tasks (Li & Lin, 2018), suggesting that smartphone usage at work is not only beneficial, but has become necessary to meet various work-related needs (Li & Lin, 2018).

However, these perceived benefits of workplace smartphone usage are met with a number of concerns related to both productivity at the office and well-being of the individual. Notably, workplace smartphone usage may lead to increased distraction stemming from engagement with non-work related content such as personal conversations and social media usage (Middleton, 2007). The implications of smartphone distraction in the workplace may be most prevalent in students and working young adults. In 2015, the Pew Research Center found that in smartphone owners aged 18–29, text messaging, social networking and internet use were among the most frequently used applications (apps) on their smartphone (Smith, Aaron, 2015). Thus, smartphone usage in the workplace can prove to be a form of distraction, and is shown to reduce focus and increase reaction time (David et al., 2015; Porter, 2010). Furthermore, considering that there has been an increase in smartphone usage over the past decade (OECD, 2017), the immediate effects of increased smartphone usage are also met with increased consideration for downstream physical and psychological concerns. Frequent phone usage has been associated with anxiety, depression, and decreased sleep quality (Demirci et al., 2015; Patalay & Gage, 2019). These findings suggest that further research is required to continue identifying the predictors and consequences of smartphone overuse.

1.1 Smartphone Overuse & Boredom

As smartphone overuse persists, the association between boredom and smartphone use has become more widely researched (Diefenbach & Borrmann, 2019). Mikulas and Vodanovich (1993) define boredom as "a state of relatively low arousal and dissatisfaction, which is attributed to an inadequately stimulating situation" (Mikulas & Vodanovich, 1993). The

concept of sensation-seeking has been related to the tendency to be bored (Zuckerman & Neeb, 1979). This logic can be extended to the relationship between the habitual use of smartphones and boredom, where checking one's smartphone for instant gratification becomes a method for the user to avoid boredom and a lack of sufficient surrounding stimuli (Leung, Louis, 2007; Oulasvirta et al., 2012). Boredom is associated with a variety of negative behaviors, which further highlight its ties to smartphone usage. For example, younger people, who are more likely to experience boredom, have a higher likelihood of internet addiction, which may partly be manifested in smartphone internet usage (Biolcati et al., 2018; Spaeth et al., 2015). Furthermore, a study examining the relationship between boredom and problem internet use found that students prone to boredom were more likely to seek out the internet for stimulation, which had a negative effect on their academic performance (Skues et al., 2016). This relationship is also seen in the absence of smartphone social media engagement, where disengagement from smartphone social apps results in increased reported boredom (Stieger & Lewetz, 2018). With increased smartphone dependency resulting in more pervasive smartphone addiction (Lin et al., 2015), the mental and physical effects of the relationship between boredom and smartphone use should be further investigated.

1.2 Smartphone Overuse & Fatigue

Unfortunately, the consequences of smartphone addiction extend beyond considerations of boredom, as increased smartphone use, particularly as it pertains to social media engagement, has also been linked to fatigue (Whelan et al., 2020). Fatigue can be considered both as a physical concern, as it relates to an inability to best perform a physical task (Hagberg, 1981), and a mental concern, where a prolonged cognitive task results in reduced performance (Mizuno et al., 2011). Indeed, increased smartphone use may affect both physical and mental fatigue. In young and normal adults, consistent smartphone use results in increased forward head posture, which has been shown to worsen fatigue in the neck and shoulders based on smartphone use duration (Kim & Koo, 2016; Lee et al., 2017). The effects of smartphone overuse have also been considered in the context of mental fatigue. A study in Saudi Arabia found that frequent mobile phone use resulted in fatigue, accompanied by headache and tension (Al-Khlaiwi & Meo, 2004). This influence of smartphones on mental fatigue may be related to workplace focus and efficiency, as a relationship has been shown between fatigue and a dearth of motivation to maintain task performance (Boksem et al., 2006). Additionally, smartphone app exposure has been linked to mental fatigue in decision-making performance (Gantois et al., 2020).

1.3 Machine Learning & Psychological Research

To appropriately address the relationships between smartphone use and fatigue or boredom, deviating from traditional statistical inference in behavioral research may be efficacious. When implemented appropriately, a machine learning approach may yield more accurate and unbiased results for psychological research (Orrù et al., 2020). Recent efforts to leverage machine learning in psychological research with clinical characteristics and neuroimaging have shown promise in predicting: depression (Wang et al., 2019), anxiety (Mellem et al., 2020), and obsessive compulsive disorder (Kushki et al., 2019). However, smartphone data specifically has been utilized in a classification machine learning framework to predict

personality traits and emotional states (Chittaranjan et al., 2013; Sultana et al., 2020). Unfortunately, few studies have utilized machine learning to specifically predict boredom and fatigue (Seo et al., 2019; Zuñiga et al., 2020), respectively. Overall, this body of research is generally sparse and disease-specific.

1.4 Current Study and Hypotheses

A recent study in the Netherlands considered the relationships between both fatigue and boredom and smartphone use in PhD students (Dora et al., 2020). A Bayesian mixed-effects models approach found associations between smartphone use and both boredom and fatigue. Given the promise of (ensemble) machine learning as a tool to interrogate mental health and behavioral constructs (Papini et al., 2018; Pearson et al., 2019; Srividya et al., 2018) and the unique affordances of this temporal dataset, we have expanded upon the original manuscript's analyses by leveraging an ensemble machine learning approach on this same study population and data set to test app use associations with fatigue and boredom within a predictive framework. Accordingly, this study was driven by the following hypotheses:

i. Passively collected app use information can be effectively leveraged to predict work fatigue and boredom ($r > 0.4$) across a cohort when analyzed within an ensemble machine learning paradigm.

ii. Given the high degree of subjectivity associated with states of fatigue and boredom as well as variation in associated behavioral responses, predictive models will perform well ($r > 0.6$) on a subset of individuals while performing poorly ($r < 0.3$) on others.

2. Methods

2.1 Study Population & Original Data Set

Participants ($N = 83$, 21 male, 62 female, $mean_{age} = 26.78$) were included in the final study population based on a set of specified inclusion criteria. Participants had to be: (i) employed as a PhD candidate, (ii) own an Android smartphone, (iii) self-report high job autonomy, and (iv) self-report smartphone use more so for private use rather than work-related matters to initially qualify for the study (Dora et al., 2020). Recruitment occurred over an 18-month period. Participants varied in their PhD year ($N_{Year1} = 30$, $N_{Year2} = 21$, $N_{Year3} = 17$, $N_{Year4} = 14$, $N_{Year5} = 1$) and discipline of study (e.g., $N_{social\ sciences} = 31$, $N_{medical\ sciences} = 15$, $N_{sciences} = 11$). All such demographic information was collected via a self-report general questionnaire. Data collection was tracked from the 'App Usage - Manage/Track Usage' app across a span of three contiguous work days that were unique to the participant. This data collection time frame reflected the desire to assess continuous, densely collected app use behavioral information and self-report fatigue and boredom across consecutive days where the researchers could ensure that participants were not engaged in non-dependent work-related activities that would take them away from their office (e.g., meetings with students). The original study was approved by Radboud University's Ethics Committee Social Science (ECSW2017–1303-485), and conducted in accordance with local guidelines.

2.2 Outcome Metrics: Fatigue and Boredom

Ecological momentary assessment (EMA) data on fatigue and boredom was collected from participants between the hours of 8 AM and 6 PM. Participants were prompted hourly with "How fatigued do you currently feel?" and "How bored do you currently feel", and answered using a 100-point visual analog sliding scale to gauge their response ("not at all" to "extremely"), a measurement adapted from the single-item fatigue measure proposed by Van Hooff et al (Van Hooff et al., 2007). These self-reported metrics are therefore linked to the specific hour in which the EMA was completed with data sporadically available across three consecutive work days. Availability of information across participants is variable both in terms of the absolute number of fatigue and boredom assessments (ranging from 9 to 24) as well as in terms of the times represented by the responses. As indicated in Figure 1.1, the raw outcome data comprised 3,235 rows of long-formatted (i.e., one participant spans multiple rows), hour-based fatigue and boredom self-report outcome across three days.

2.3 Feature Engineering: Time-dependent App Use Data

The current study used passively collected app log information to derive several time-series-based features (see Figure 1.1). This app log data contained detailed information including de-identified subject ID, app name, day (1,2, or 3), time-stamped initiation of app, and total consecutive duration of use for each app event. For $N = 86$ participants across a continuous 3-day span of time, this equated to 25,308 rows of data across five columns of features in a long data format (i.e., each subject is represented across multiple rows based on the number of app events logged for that subject). To parse each participant's log, all events were binned into one-hour representations of app use behavior. For each hour, ranging from 6 AM to 11 PM across three days, (i) total number of app events, (ii) total number of unique apps used, (iii) total duration of app use (in minutes), (iv) root mean square of successive differences (RMSSD) in app use events (binned in discrete five-minute windows), and (v) RMSSD of unique app use events (binned in discrete five-minute windows) were calculated. Quantiles (5th, 10th, 35th, 50th, 65th, 90th, and 95th) of these five-minute window count distributions for both total use events and unique use events were also calculated and used as additional features. With app names available in the logs, the Google Play score was programmatically queried with the *play_scraper* (Version 0.60) Python library under the MIT license (Liu, 2019) to identify corresponding app categories. For each of "Social", "Communication", "Game", "Video", "Music", and "Shopping", additional features were created that documented the (i) total number of app use events, (ii) total duration of app use events (in minutes), (iii) percentage of total event counts, and (iv) percentage of total app use minutes belonging to the category within the designated hour. Taken together, this resulted in the creation of 43 time-dependent predictors. All data wrangling and associated feature derivations were performed with custom scripts in Python using the raw app log data provided as public access in the initial study (Dora et al., 2020).

2.4 Lagged Derivative Datasets

To more thoroughly investigate long-term behavioral relationships, this work generated five separate, long-format (multiple, time-based rows for each participant) datasets that reflect lagged pairings of app use information with fatigue and boredom outcomes. For each

participant, app use data across a previous hour (*t-1* for lag 1, *t-2* for lag 2, etc.) were matched to each hour (*t*) in which a fatigue/boredom measure was available. For example, an available fatigue score at 12 PM was matched with app use information in the 11 – 12 PM time frame for the lag 1 dataset, while this same fatigue score was matched with app use information in the 10 – 11 AM time frame for the lag 2 dataset. Conditioning these lagged datasets to only include time points for which outcome metrics were available ultimately resulted in trajectories of fatigue and boredom that reflect *relative* rather than absolute time; points in these derivative data spaces are not equivalent in interval. See Figure 1.2 for a qualitative representation of this process.

2.5 Data Preprocessing

All features were individually standardized such that data had a mean of 0 and standard deviation of 1 for use in subsequent models.

2.6 Machine Learning Model Framework

The machine learning pipeline was built and run in R (v3.6.1) using the *caret* package (Kuhn, 2008). Prediction was carried out as separate regression modeling tasks for fatigue and boredom. For both outcomes, the work sought to build subject-specific, idiographic models using a nested leave-one-subject-out (LOSO) cross-validation framework. Using each of the lagged datasets, a Random Forest model (Liaw & Wiener, 2002) was trained, validated, and hyperparameter tuned on $N = 82$ participants to predict the held-out participant's ($N = 1$) series of self-reported outcome measures (see Figure 1.3). Each Random Forest model was optimized based on the number of features randomly selected for sampling at each decision split in the tree (mtry) via an automatic grid search procedure within the *caret* package. The number of trees to generate was kept at the default of 500 trees. The resulting outcome predictions from the $N = 83$ idiographic models were then saved as higher-level features for an ensemble Random Forest model. As this LOSO approach was conducted independently on five separate lagged datasets (lag 1, lag 2, lag 3, lag 4, and lag 5), the ensemble predictor space consisted of five predictors that incorporated lower-level predictions across 415 (5×83) uniquely-tuned Random Forest models. With this derived, time-dependent ensemble of prediction values, a second round of ensemble model generation with LOSO cross-validation was conducted. As with the lower-level modeling, a Random Forest model was trained, validated, and hyperparameter tuned on $N = 82$ participants to predict the held-out participant's ($N = 1$) series of self-reported outcome measures (see Figure 1.4).

The resulting outcome predictions for each of the $N = 83$ idiographic ensemble models were evaluated against the observed self-report values and assessed for correlative strength (see Figure 1.5). Where R^2 represents the proportion of the variance in the true fatigue or boredom outcome scores for a subject that is predictable from the model's corresponding predicted scores, the square root of this explained variance, or *r*, was used to represent the correlation. Correlative strength between the predicted and observed values of fatigue or boredom for each subject over time was qualitatively transformed into strata that represent "very strong" ($r > 0.8$), "strong" ($0.8 \quad r > 0.6$), "moderate" ($0.6 \quad r > 0.4$), "weak" ($0.4 \quad r$ 0.2), and "no correlation" ($r < 0.2$). Such thresholds for more precise notions of strength are

variable, subfield-dependent, and context-dependent; however, the above scheme represents a more conservative adaptation of Cohen's widely known and applied guidelines for the behavioral sciences (Cohen, 1988; Hemphill, 2003). Under this framework, an r in the range of 0.1 represents a "small" correlation, an r in the range of 0.3 represents a "medium" correlation, and an r within range of 0.5 represents a "large" correlation.

To additionally evaluate the nomothetic capabilities of the subject-specific models across the cohort, three linear mixed-effect models were tested using the *lme4* package for both fatigue and boredom, and the conditional R^2 was calculated for each model using the *performance* package (Bates et al., 2015; Daniel et al., 2020). The first model only considered the subject as a random effect to assess the conditional R^2 when model predictions were omitted. In the second model, outcome predictions were added as a fixed effect, and the variance explained by these predictions as a fixed effect was calculated by taking the difference between the aforementioned conditional R^2 values. In the third model, predictions were included as both a fixed effect and a random effect. The variance explained by the outcome predictions as a random effect was calculated by taking the difference between the conditional R^2 values of the second and final model (Table 1).

2.7 Feature Importance

For each of the lower-level and ensemble Random Forest models generated in the pipeline, the *varImp* function in *caret* was applied to determine a scaled relative measure of feature importance. As mentioned, the lower-level models were trained on 43 time-dependent predictors, while the ensemble models were trained on five lag-based variables reflecting lower-level model predictions of the outcome. In practice, the *varImp* function evaluates the relationship between each predictor and the outcome in the model. For evaluation, a loess smoother was used to fit the outcome and predictor. The R^2 was then calculated against an intercept-only null model to arrive at a measure of feature importance. These values were averaged across all $N = 83$ participant-specific models (for both lower-level and ensemble) to arrive at summative trends.

3. Results

3.1 Ensemble Model Performance - Fatigue

Performance of the idiographic lagged-ensemble models varied significantly from person to person. Of the $N = 83$ built to predict self-reported fatigue scores across a three-day period of time, approximately 47% ($N = 39$) were capable of tracking observed fatigue with at least a moderate correlation ($r > 0.4$). Moreover, approximately 17% ($N = 14$) of the models had a strong correlation ($r > 0.6$) with observed values of subject fatigue, while approximately 24% of the models had a very weak or no correlation ($r < 0.2$) with the outcome. Figure 2A illustrates the predicted and observed trajectories of fatigue on a subset of participants (see Supplementary Table 1 for a complete list of idiographic model performances).

To assess the relative influence of the ensemble model fatigue predictions on variance explained, a linear mixed-effects model incorporating the ensemble model predictions as a fixed effect explained an additional 5.8% of the variance of the entire model compared

to the machine learning prediction-agnostic model ($R^2 = 0.369$). Additionally, when the ensemble model predictions were included as both a random and fixed effect, the linear mixed-effects model explained an additional 2.6% compared to when model predictions were only included as a fixed effect ($R^2 = 0.427$).

3.2 Variable Importance - Fatigue

In the final ensemble prediction models of fatigue, Lag 5 was, on average, the most important feature driving the predictions, while Lag 2 showed no relative importance when compared to the other lags. However, when considering the individual, lower-level lag models, the top three most important features were uniform across all five lags: 'Hour', 'Total App Duration', and 'Duration Communication', respectively. Additionally, although there are minor differences in the relative rank of these top ten most important features, the features themselves are largely recapitulated across all lags. Notably, 'Hour' was substantially and uniformly more important in driving the individual lag model predictions compared to the subsequent top features across the individual lag models. Table 2 reports the average relative importance of the top ten features across the lower-level lag models while also reporting the relative importance of each lag toward fatigue prediction in the final ensemble model.

3.3 Ensemble Model Performance - Boredom

Much like with the models designed to predict fatigue, performance of the idiographic lagged-ensemble models for boredom varied significantly from person to person. However, general performance was much less robust overall. Of the $N = 83$ subject-specific models predicting self-reported boredom scores across a three-day period of time, approximately 24% ($N = 20$) were capable of attaining at least a moderate correlation ($r > 0.4$) with observed values. In addition, only about 2% ($N = 2$) of the models expressed strong correlations ($r > 0.6$) with more than half (54%; $N = 45$) presenting weak or no correlations ($r < 0.4$) with the outcome. Figure 2B illustrates the predicted and observed trajectories of boredom on a subset of participants (see Supplementary Table 2 for a complete list of idiographic model performances).

A linear mixed-effects model incorporating the ensemble model's boredom predictions as a fixed effect explained an additional 0.5% of the variance of the entire model compared to the machine learning prediction-agnostic model ($r^2 = 0.306$). When the ensemble model predictions were included as both a random and a fixed effect, the linear mixed-effects model performed worse compared to when model predictions were only included as a fixed effect, with a 0.2% lower model contribution ($r^2 = 0.309$ vs $r^2 = 0.311$, respectively).

3.4 Variable Importance - Boredom

In contrast to the final ensemble for fatigue, Lag 1, Lag 4, and Lag 5 had comparable importance as features in the final ensemble prediction for boredom. Lag 2 again showed no relative importance when compared to the other individual lag features. The relative importance of the features in the individual lag models for boredom were more comparable across the top ten features, although 'Hour' and 'Total App Duration' were the top two most important features across all lag models. Table 3 reports the average relative importance

of the top ten features across the lower-level lag models while also reporting the relative importance of each lag toward boredom prediction in the final ensemble model.

4. Discussion

This study utilized a dataset consisting of passively collected smartphone app use data and self-reported fatigue and boredom scores across three days in a population of Dutch PhD students. Where this data was collected and analyzed previously using a Bayesian (generalized) linear mixed methods approach for the description of the bidirectional associations of app use and fatigue/boredom (Dora et al., 2020), the current study aimed to leverage the data to interrogate potential predictive power within larger, extended temporal windows. To this end, this work engineered time-based predictors of app use and interrogated the capabilities of smartphone engagement at work to forecast fatigue and boredom trajectories. Uniquely, this research sought to leverage an ensemble machine learning approach that combined the insights of lag-based derivations of time-anchored data to ultimately arrive at predicted consensus trajectories of fatigue and boredom.

4.1 Leveraging Lagged Datasets and Ensemble Machine Learning

The decision to create separate lagged datasets was made to allow for the deconstruction of time-lagged behavioral relationships between app use behaviors and fatigue/boredom outcome while also simultaneously permitting an ability to compare relative importance of specific behavioral operationalizations both within and between feature-outcome temporal pairings. Due to the large number of predictors (43), reorganization of the data into a single, wide-formatted dataset with $(43 \times 5 \text{ lags}) = 215$ features may have led to the well-appreciated "curse of dimensionality" problem whereby poor representation across feature value combinations would have impacted the model's ability to detect unique patterns and parse signal from noise. Moreover, treating each lagged dataset as a component within the greater temporal context of the association in question capitalized well on the advantages of an ensemble machine learning framework. Instead of employing this strategy to train different types of learners on the same data, we applied multiple, algorithmically identical (Random Forest) weaker learners trained on distinct, yet related temporal representations of the data. In turn, the predictions from these weak learners were combined to inform a stronger consensus learner with an ability to differentially utilize lag-specific predictions. The subsequent application of feature importance methods to introspect the ensemble model more holistically afforded an appreciation of the relative impact of different lagged framings in the overall prediction of fatigue and boredom from app use behavioral phenotypes. Taken together, the derived feature space for the ensemble model represented temporal abstractions of the raw feature data that was informative to varying degrees across the models.

More generally, a strength of the machine learning-based predictive approach is that it allowed validation of models within an external, out-of-sample context. Unlike traditional statistical approaches that build models using the entirety of the dataset, evaluation of a properly constructed machine learning pipeline can reflect unbiased performance since any one prediction on a subject is not influenced by information on that same subject within the dataset. The application of cross-validation during model training was selected specifically

for this reason. More specifically, LOSO cross-validation was chosen for two reasons. First, the stacked format of the data required that validation splits were performed by grouping all rows of data associated with the same subject prior to split to ensure no leakage of information caused by having any one subject's data in both the training and validation folds. Second, and more broadly, due to the relatively small sample size ($N = 82$), a k-fold approach would have been less appropriate as it would have removed a more significant portion of the data from any iteration of the training set.

The algorithm of choice for both the lower level and ensemble model was the Random Forest for a few reasons. It is robust to overfitting, relatively simplistic, interpretable, and easily comprehensible. Importantly, it is also highly conducive to quantification of relative predictor importance given that the tree-based algorithm operates through decision splits that are based on difference maximization of a random subset of features (Breiman, 2001; Fawagreh et al., 2014). Overall, due to the novelty of the lag-based ensemble approach, it was desirable to utilize a model that was simple and interpretable with a history of demonstrated utility within the mental health predictive space (Jacobson et al., 2020; Jacobson & Chung, 2020).

4.2 Interpretation of Results

The primary results of this application indicated highly heterogeneous subject-specific model efficacies in both fatigue and boredom prediction. For predictions of fatigue, nearly half (47%) of the models traced the observed trends with moderate correlations ($r > 0.4$). Performance in this regard was highest among 17% of the cohort ($r > 0.6$). Interestingly, Lag 5, reflecting a five-hour window, had the most influence on the final fatigue predictions. This may suggest that investigating larger time windows for predicting fatigue may be more efficacious than the currently considered lag times. Additionally, the duration of time participants interacted with both their smartphone overall, and with communication apps specifically, was more influential on predictions than the proportional use of any single app type on their smartphone across all lag models for fatigue. These results indicate that extended use of one's smartphone in the workplace is more influential on predicting fatigue, than numerous brief smartphone interactions. For predictions of boredom, a lower proportion of models (24%) was capable of tracing the observed trends with moderate correlative efficacy. Unlike fatigue, Lags 1, 4 and 5 had comparable influence on the final boredom predictions, providing less information on potential temporal windows to consider to optimize boredom predictability. However, similar to fatigue, overall app and communication app duration were more influential in predicting boredom across the individual lag models than proportional use of any single app type. In tandem, these findings suggest that the type of smartphone use may be less informative than smartphone interaction duration for predicting boredom and fatigue.

Despite the individual variation in model performances, linear mixed-effects modeling revealed that the ensemble model predictions for fatigue across the cohort explained an additional 8.4% of the variance (5.8% fixed effects; 2.6% random effects) compared to a baseline mixed-effects model without these predictions (Table 1). Ensemble model predictions for boredom across the cohort, following the above trend in the idiographic

analysis of the models, only explained an additional 0.5% of the variance in fixed effects and in fact explained less 0.2% of mixed effects compared to a baseline mixed-effects model without ensemble model predictions (Table 1).

4.3   Novelty and Innovation

This study is novel in its use of time lags to derive independent datasets for parallel model construction. Instead of building an ensemble model consisting of several different machine learning algorithms applied to the same data points, this research applies the ensemble to interrogate lag structure and capitalize on the potential predictive affordances of feature-outcome associations at varying time scales. Specifically, the work applies a Random Forest model to five distinct lagged datasets that capture one-hour windows of app use behavior and link them to fatigue and boredom measures up to five hours in the future. Indeed, the resulting predictions of the lower-level models used to inform the ensemble model are a reflection of lagged relationships, rather than algorithmic differences and idiosyncrasies. In addition to the usual benefit of predictive scope in an ensemble paradigm, notions of variable importance become linked to the potential relative significance of lagged relationships and may help guide future exploration into temporal associations on more seemingly disconnected scales. For example, this study generally found that app use four to five hours prior is more important to prediction of fatigue than app use one to two hours prior. While the original analysis of this data investigated twenty minutes before and after fatigue self-report, the current investigation beyond this time frame, and the simultaneous analysis of the data across multiple ranges of time, uncovered a potential trend that warrants future investigation.

By extension, the implementation of lagged machine learning models allows for the comparison of variable importance both within and between different time scales. Consistencies and discrepancies between lags may also be elucidating. As previously mentioned, the results indicated consistent influence of communication-based app use, echoing previous studies (Smith, Aaron, 2015). Additionally, the consistently high importance of RMSSD-based features suggests a focus toward quantification of app use behavioral *patterns* in addition to absolute magnitude of use over time. The variable importance scores in this study are based on models whose performance indicated an ability to account for a moderate proportion of the variance, thus the observed trends in what was driving these models, whether these are lower-level app use predictors or higher-level ensemble lags, may signal qualities of the exposure-outcome relationship that have yet to be further explored in the literature.

4.4   Limitations

Despite the mentioned strengths of this study, there are a number of limitations that should be considered both for evaluating the current results, and for future implementation of similar methodologies. Most notably, the study sample consists solely of employed PhD students that own a smartphone, suggesting a well-educated and financially stable group of participants. Due to the homogeneity of a student-specific study sample, issues of generalizability to the public have been found to be problematic (Peterson, 2001; Peterson & Merunka, 2014). Factors such as age, income, and type of working environment may impact

prediction of the models, and personal perceptions of boredom and fatigue may manifest differently as a consequence of such factors. Further, student self-reports on personal and attitudinal variables do not generalize well to the public (Hanel & Vione, 2016), which may be extended to the self-reported fatigue and boredom metrics considered in the present study. An additional limitation is the relatively small sample size and inconsistent availability of fatigue and boredom scores both across participants and within participants across days. It has been shown that an insufficient sample size may affect result interpretations, and subsequent clinical decisions (Faber & Fonseca, 2014). In general, analysis on a larger cohort in future would statistically enable use of a true held-out test set (e.g., 80–20 training-test split instead of sole reliance on cross-validation during training) to evaluate performance. The limited sample size is also a concern for external validity to PhD students of varying years of completion or degree focus as these factors may influence smartphone usage. Taken together, further research and application of the proposed lagged-ensemble predictive framework to more diverse non-student and/or older adult samples with differing demographic and work profiles is required to more fully assess generalizability and utility. Of final analytical note, when considering the variable importance of each model, we were not able to infer the predictor's directionality of impact on the outcome.

5. Conclusion

This study sought to leverage a nested-LOSO ensemble machine learning approach to predict workplace boredom and fatigue from passively collected and derived smartphone usage features. The novelty of this work stems from the interrogation of lag-specific models to assess their relative contribution to a final outcome prediction. Despite several limitations, the results appear promising. Future research will aim to utilize a more robust and representative study population, along with a stringent outcome-reporting protocol to more fully evaluate the utility of the explained methodology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Al-Khlaiwi T, & Meo SA (2004). Association of mobile phone radiation with fatigue, headache, dizziness, tension and sleep disturbance in Saudi population. Saudi Medical Journal, 25(6), 732–736. [PubMed: 15195201]

Bates D, Mächler M, Bolker B, & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1). 10.18637/jss.v067.i01

Biolcati R, Mancini G, & Trombini E. (2018). Proneness to Boredom and Risk Behaviors During Adolescents' Free Time. Psychological Reports, 121(2), 303–323. 10.1177/0033294117724447 [PubMed: 28776483]

Boksem MAS, Meijman TF, & Lorist MM (2006). Mental fatigue, motivation and action monitoring. Biological Psychology, 72(2), 123–132. 10.1016/j.biopsycho.2005.08.007 [PubMed: 16288951]

Breiman L. (2001). Random forests. Machine Learning, 45(1), 5–32. 10.1023/A:1010933404324

Chittaranjan G, Blom J, & Gatica-Perez D. (2013). Mining large-scale smartphone data for personality studies. Personal and Ubiquitous Computing, 17(3), 433–450. 10.1007/s00779-011-0490-1

Cohen J. (1988). Statistical power analysis for the behavioral sciences (et al.). L. Erlbaum Associates.

Daniel Makowski, Indrajeet Patil D, & Waggoner P. (2020). easystats/performance: Performance 0.4.7 (0.4.7) [Computer software]. Zenodo. 10.5281/ZENODO.3952174

David P, Kim J-H, Brickman JS, Ran W, & Curtis CM (2015). Mobile phone distraction while studying. New Media & Society, 17(10), 1661–1679. 10.1177/1461444814531692

Demirci K, Akgönül M, & Akpinar A. (2015). Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. Journal of Behavioral Addictions, 4(2), 85–92. 10.1556/2006.4.2015.010 [PubMed: 26132913]

Diefenbach S, & Borrmann K. (2019). The Smartphone as a Pacifier and its Consequences: Young adults' smartphone usage in moments of solitude and correlations to self-reflection. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 1–14. 10.1145/3290605.3300536

Dora J, van Hooff M, Geurts S, Kompier MAJ, & Bijleveld E. (2020). Fatigue, boredom, and objectively-measured smartphone use at work [Preprint]. PsyArXiv. 10.31234/osf.io/uy8rs

Faber J, & Fonseca LM (2014). How sample size influences research outcomes. Dental Press Journal of Orthodontics, 19(4), 27–29. 10.1590/2176-9451.19.4.027-029.ebo

Fawagreh K, Gaber MM, & Elyan E. (2014). Random forests: From early developments to recent advancements. Systems Science & Control Engineering, 2(1), 602–609. 10.1080/21642583.2014.956265

Gantois P, Caputo Ferreira ME, Lima-Junior D. de, Nakamura FY, Batista GR, Fonseca FS, & Fortes L. de S. (2020). Effects of mental fatigue on passing decision-making performance in professional soccer athletes. European Journal of Sport Science, 20(4), 534–543. 10.1080/17461391.2019.1656781 [PubMed: 31424354]

Hagberg M. (1981). Muscular endurance and surface electromyogram in isometric and dynamic exercise. Journal of Applied Physiology, 51(1), 1–7. 10.1152/jappl.1981.51.1.1 [PubMed: 7263402]

Hanel PHP, & Vione KC (2016). Do Student Samples Provide an Accurate Estimate of the General Public? PLOS ONE, 11(12), e0168354. 10.1371/journal.pone.0168354

Hemphill JF (2003). Interpreting the magnitudes of correlation coefficients. American Psychologist, 58(1), 78. 10.1037/0003-066X.58.1.78

Jacobson NC, & Chung YJ (2020). Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. Sensors (Basel, Switzerland), 20(12). 10.3390/s20123572

Jacobson NC, Yom-Tov E, Lekkas D, Heinz M, Liu L, & Barr PJ (2020). Impact of online mental health screening tools on help-seeking, care receipt, and suicidal ideation and suicidal intent: Evidence from internet search behavior in a large U.S. cohort. Journal of Psychiatric Research. 10.1016/j.jpsychires.2020.11.010

Kim S-Y, & Koo S-J (2016). Effect of duration of smartphone use on muscle fatigue and pain caused by forward head posture in adults. Journal of Physical Therapy Science, 28(6), 1669–1672. 10.1589/jpts.28.1669 [PubMed: 27390391]

Kossek EE, & Lautsch BA (2012). Work–family boundary management styles in organizations: A cross-level model. Organizational Psychology Review, 2(2), 152–171. 10.1177/2041386611436264

Kuhn M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. [PubMed: 27774042]

Kushki A, Anagnostou E, Hammill C, Duez P, Brian J, Iaboni A, Schachar R, Crosbie J, Arnold P, & Lerch JP (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. Translational Psychiatry, 9(1), 318. 10.1038/s41398-019-0631-2 [PubMed: 31772171]

Lanaj K, Johnson RE, & Barnes CM (2014). Beginning the workday yet already depleted? Consequences of late-night smartphone use and sleep. Organizational Behavior and Human Decision Processes, 124(1), 11–23. 10.1016/j.obhdp.2014.01.001

Lee S, Choi Y-H, & Kim J. (2017). Effects of the cervical flexion angle during smartphone use on muscle fatigue and pain in the cervical erector spinae and upper trapezius in normal adults in their 20s. Journal of Physical Therapy Science, 29(5), 921–923. 10.1589/jpts.29.921 [PubMed: 28603372]

Leung Louis. (2007). Leung, L. (2007). Leisure Boredom, Sensation Seeking, Self-esteem, Addiction Symptoms and Patterns of Mobile Phone Use.

Li L, & Lin TTC (2018). Examining how dependence on smartphones at work relates to Chinese employees' workplace social capital, job performance, and smartphone addiction. Information Development, 34(5), 489–503. 10.1177/0266666917721735

Liaw A, & Wiener M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22.

Lin Y-H, Lin Y-C, Lee Y-H, Lin P-H, Lin S-H, Chang L-R, Tseng H-W, Yen L-Y, Yang CCH, & Kuo TBJ (2015). Time distortion associated with smartphone addiction: Identifying smartphone addiction via a mobile application (App). Journal of Psychiatric Research, 65, 139–145. 10.1016/j.jpsychires.2015.04.003 [PubMed: 25935253]

Liu D. (2019). Play Scraper. https://github.com/danieliu/play-scraper

Mellem MS, Liu Y, Gonzalez H, Kollada M, Martin WJ, & Ahammad P. (2020). Machine Learning Models Identify Multimodal Measurements Highly Predictive of Transdiagnostic Symptom Severity for Mood, Anhedonia, and Anxiety. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5(1), 56–67. 10.1016/j.bpsc.2019.07.007 [PubMed: 31543457]

Middleton CA (2007). Illusions of Balance and Control in an Always-on Environment: A Case Study of BlackBerry Users. Continuum, 21(2), 165–178. 10.1080/10304310701268695

Mikulas WM, & Vodanovich SJ (1993). The essence of boredom. The Psychological Record, 43(1), 3.

Mizuno K, Tanaka M, Yamaguti K, Kajimoto O, Kuratsune H, & Watanabe Y. (2011). Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. Behavioral and Brain Functions, 7(1), 17. 10.1186/1744-9081-7-17 [PubMed: 21605411]

OECD. (2017). PISA 2015 Results (Volume III): Students' Well-Being. OECD. 10.1787/9789264273856-en

Orrù G, Monaro M, Conversano C, Gemignani A, & Sartori G. (2020). Machine Learning in Psychometrics and Psychological Research. Frontiers in Psychology, 10, 2970. 10.3389/fpsyg.2019.02970 [PubMed: 31998200]

Oulasvirta A, Rattenbury T, Ma L, & Raita E. (2012). Habits make smartphone use more pervasive. Personal and Ubiquitous Computing, 16(1), 105–114. 10.1007/s00779-011-0412-2

Papini S, Pisner D, Shumake J, Powers MB, Beevers CG, Rainey EE, Smits JAJ, & Warren AM (2018). Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. Journal of Anxiety Disorders, 60, 35–42. 10.1016/j.janxdis.2018.10.004 [PubMed: 30419537]

Patalay P, & Gage SH (2019). Changes in millennial adolescent mental health and health-related behaviours over 10 years: A population cohort comparison study. International Journal of Epidemiology, 48(5), 1650–1664. 10.1093/ije/dyz006 [PubMed: 30815691]

Pearson R, Pisner D, Meyer B, Shumake J, & Beevers CG (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. Psychological Medicine, 49(14), 2330–2341. 10.1017/S003329171800315X [PubMed: 30392475]

Peterson RA (2001). On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-analysis. Journal of Consumer Research, 28(3), 450–461. 10.1086/323732

Peterson RA, & Merunka DR (2014). Convenience samples of college students and research reproducibility. Journal of Business Research, 67(5), 1035–1041. 10.1016/j.jbusres.2013.08.010

Porter G. (2010). Alleviating the "dark side" of smart phone use. 2010 IEEE International Symposium on Technology and Society, 435–440. 10.1109/ISTAS.2010.5514609

Seo J, Laine TH, & Sohn K-A (2019). An Exploration of Machine Learning Methods for Robust Boredom Classification Using EEG and GSR Data. Sensors, 19(20), 4561. 10.3390/s19204561

Skues J, Williams B, Oldmeadow J, & Wise L. (2016). The Effects of Boredom, Loneliness, and Distress Tolerance on Problem Internet Use Among University Students. International Journal of Mental Health and Addiction, 14(2), 167–180. 10.1007/s11469-015-9568-8

Smith Aaron. (2015, 4 1). U.S. Smartphone Use in 2015. Pew Research Center. https://www.pewresearch.org/internet/2015/04/01/us-smartphone-use-in-2015/

Spaeth M, Weichold K, & Silbereisen RK (2015). The development of leisure boredom in early adolescence: Predictors and longitudinal associations with delinquency and depression. Developmental Psychology, 51(10), 1380–1394. 10.1037/a0039480 [PubMed: 26214227]

Srividya M, Mohanavalli S, & Bhalaji N. (2018). Behavioral Modeling for Mental Health using Machine Learning Algorithms. Journal of Medical Systems, 42(5), 88. 10.1007/s10916-018-0934-5 [PubMed: 29610979]

Stieger S, & Lewetz D. (2018). A Week Without Using Social Media: Results from an Ecological Momentary Intervention Study Using Smartphones. Cyberpsychology, Behavior, and Social Networking, 21(10), 618–624. 10.1089/cyber.2018.0070

Sultana M, Al-Jefri M, & Lee J. (2020). Using Machine Learning and Smartphone and Smartwatch Data to Detect Emotional States and Transitions: Exploratory Study. JMIR MHealth and UHealth, 8(9), e17818. 10.2196/17818 [PubMed: 32990638]

Van Hooff MLM, Geurts SAE, Kompier MAJ, & Taris TW (2007). "How Fatigued Do You Currently Feel?" Convergent and Discriminant Validity of a Single- Item Fatigue Measure. Journal of Occupational Health, 49(3), 224–234. 10.1539/joh.49.224 [PubMed: 17575403]

Wang S, Pathak J, & Zhang Y. (2019). Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. Studies in Health Technology and Informatics, 264, 888–892. 10.3233/SHTI190351 [PubMed: 31438052]

Whelan E, Najmul Islam AKM, & Brooks S. (2020). Is boredom proneness related to social media overload and fatigue? A stress–strain–outcome approach. Internet Research, 30(3), 869–887. 10.1108/INTR-03-2019-0112

Zuckerman M, & Neeb M. (1979). Sensation seeking and psychopathology. Psychiatry Research, 1(3), 255–264. 10.1016/0165-1781(79)90007-6 [PubMed: 298353]

Zuñiga JA, Harrison ML, Henneghan A, García AA, & Kesler S. (2020). Biomarkers panels can predict fatigue, depression and pain in persons living with HIV: A pilot study. Applied Nursing Research, 52, 151224. 10.1016/j.apnr.2019.151224 [PubMed: 31987713]
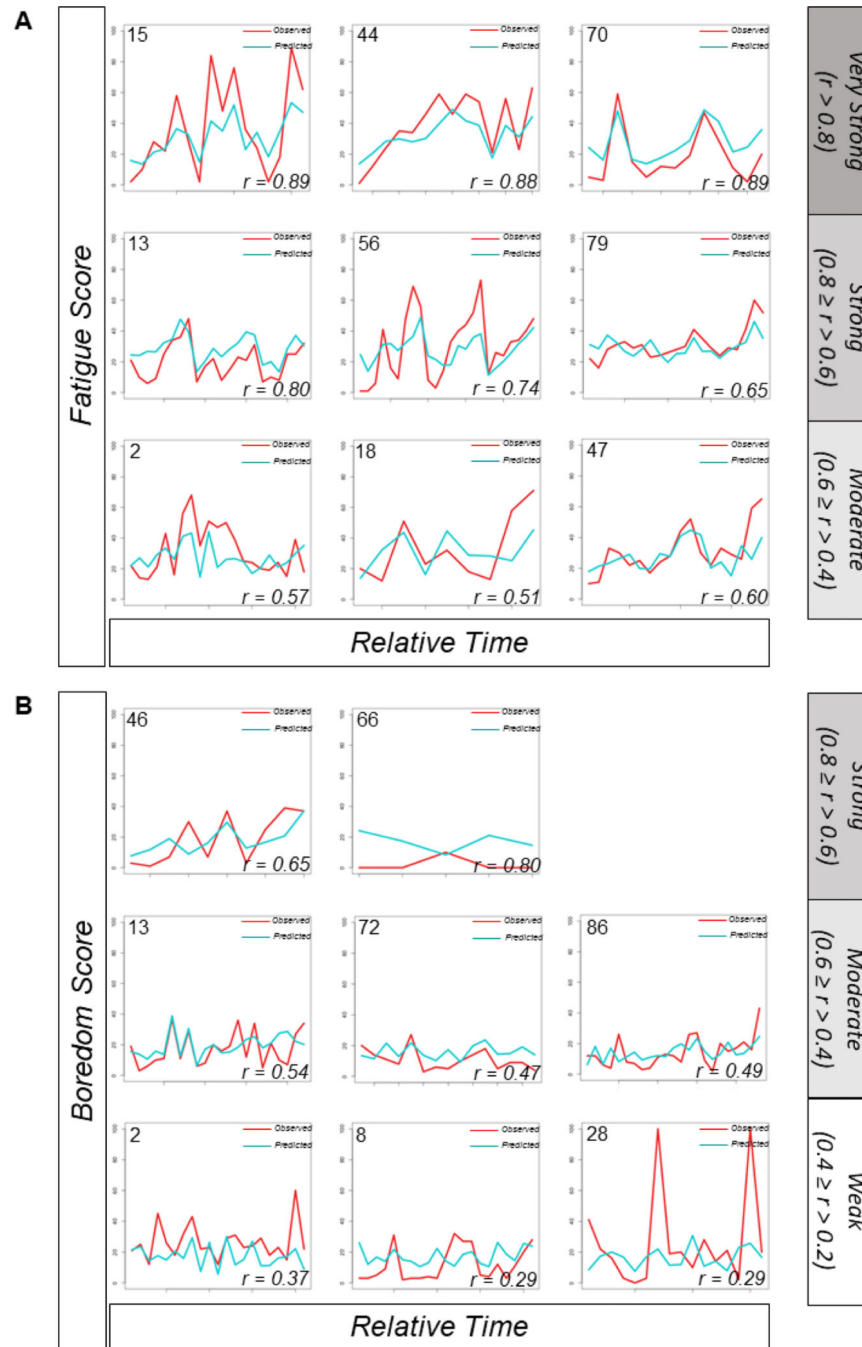
Highlights

- A lag-specific ensemble machine learning paradigm offers promise for prediction of fatigue and boredom

- Duration, communication, and patterns of app use frequency are among the most important features for prediction across lags

- Future research will benefit from evaluating associations on densely collected data across longer time scales
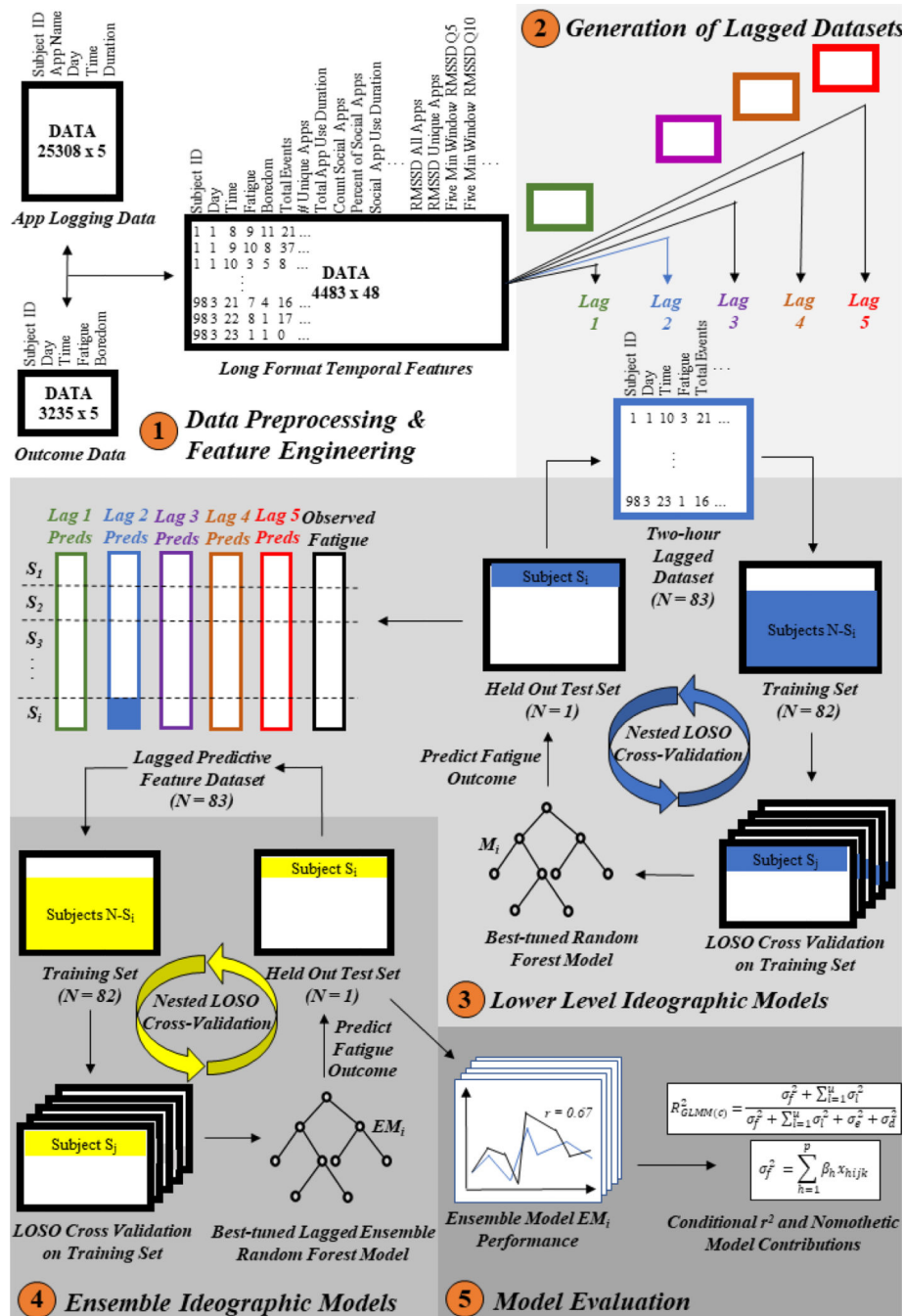
Figure 1.

Analysis Pipeline for the Prediction of Fatigue and Boredom

Note. (1) App use logs and EMA outcome datasets were combined to feature engineer time-based temporal features of app use. (2) The resulting long format dataset was lagged to create five distinct derivative datasets linking fatigue and boredom outcome measures with app use at distinct relative intervals of time. (3) For each of the five lagged datasets, a nested LOSO cross-validation framework was implemented to train, validate, and test a best-tuned Random Forest model on participant-level prediction of fatigue/boredom outcome. Lag-

specific predictions for each time point across participants are saved and used as features for an ensemble Random Forest model. (4) Nested LOSO cross-validation is performed on this ensemble predictor space. As before, a unique model is trained, validated, and tested for each participant. (5) The performance of each of these final lagged-ensemble models is evaluated idiographically and general performance of the ensemble framework is assessed to contrast nomothetic capabilities.

Figure 2.

Example Idiographic Ensemble Model Performance.

Note. Observed (red) and predicted (teal) (A) fatigue and (B) boredom trajectories across sample individuals highlighting heterogeneity in model performance across the cohort.

Table 1.

Nomothetic Ensemble Model Performance: Conditional R-Squared and Model Contribution for Fatigue and Boredom

	Conditional R^2	Model Contribution
Fatigue Models		
lmer(fatigue ~ (1|Subject), data = fatigueData))	0.369	-
lmer(fatigue ~ model_preds + (1|Subject), data = fatigueData)	0.427	+ 5.8%
lmer(fatigue ~ model_preds + (model_preds|Subject), data = fatigueData)	0.453	+ 2.6%
Boredom Models		
lmer(boredom ~ (1|Subject), data = boredomData)	0.306	-
lmer(boredom ~ model_preds + (1|Subject), data = boredomData)	0.311	+ 0.5%
lmer(boredom ~ model_preds + (model_preds|Subject), data = boredomData)	0.309	− 0.2%

Note. Conditional R^2 and relative model contributions of ensemble fatigue and boredom-based model predictions. Fixed and random effects are calculated relative to models where prediction-based information is not included.

Table 2.

Lag-Specific Feature Importance - Fatigue

FATIGUE									
Lag 1 Ensemble Imp: 26.8		Lag 2 Ensemble Imp: 0.0		Lag 3 Ensemble Imp: 16.8		Lag 4 Ensemble Imp: 44.9		Lag 5 Ensemble Imp: 100.0	
Feature	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*
Hour	100	Hour	100	Hour	100	Hour	100	Hour	100
Total App Duration	38.2	Total App Duration	54.4	Total App Duration	55.3	Total App Duration	55.1	Total App Duration	60.5
Duration Communication	35.3	Duration Communication	40.8	Duration Communication	43.8	Duration Communication	44.9	Duration Communication	43.8
RMSSD Unique Apps	31.7	RMSSD Unique Apps	34.2	RMSSD All Apps	33.1	RMSSD All Apps	34.6	RMSSD All Apps	38.4
RMSSD All Apps	29.6	RMSSD All Apps	32.9	RMSSD Unique Apps	29.5	RMSSD Unique Apps	32.6	% Duration Communication	38.3
% Duration Communication	27.4	% Duration Communication	30.0	% Duration Communication	25.4	% Duration Communication	29.9	RMSSD Unique App	33.9
% Communication	25.4	% Communication	25.2	% Communication	21.5	% Communication	29.5	% Communication	23.4
Five Minute Sliding Window Q95	24.9	Five Minute Sliding Window Q95	24.4	Five Minute Sliding Window Q95	21.5	Total App Events	20.6	Day	21.7
Total App Events	21.3	Total App Events	20.9	Day	21.0	Five Minute Sliding Window Q95	19.0	Total App Events	20.9
Count Communication	20.0	Five Minute Sliding Window Unique Q95	19.7	Total App Events	19.9	Day	18.8	Five Minute Sliding Window Q95	17.8

Note. Top ten most important features driving each of the five lagged random forest model predictions (most important = 100, least important = 0). Ensemble relative importance scores in bold underneath each of the lag column headers represent the importance of that model's predictions to the final ensemble random forest prediction of fatigue.

Table 3.

Lag-Specific Feature Importance - Boredom

BOREDOM									
Lag 1 Ensemble Imp: 76.1		**Lag 2 Ensemble Imp: 0.0**		**Lag 3 Ensemble Imp: 41.1**		**Lag 4 Ensemble Imp: 98.9**		**Lag 5 Ensemble Imp: 62.6**	
Feature	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*	*Feature*	*Scaled Imp*
Hour	99.1	Total App Duration	92.5	Total App Duration	99.4	Total App Duration	99.6	Total App Duration	98.5
Total App Duration	90.5	Hour	90.1	Hour	88.7	Hour	84.6	Hour	93.2
Duration Communication	78.0	Duration Communication	71.2	Duration Communication	68.4	Duration Communication	76.9	Duration Communication	85.6
% Duration Communication	66.4	% Duration Communication	54.2	RMSSD All Apps	52.9	RMSSD All Apps	57.4	% Duration Communication	84.4
RMSSD All Apps	66.2	RMSSD All Apps	54.1	RMSSD Unique Apps	45.1	RMSSD Unique Apps	50.5	RMSSD All Apps	84.2
RMSSD Unique Apps	62.4	RMSSD Unique Apps	51.8	% Duration Communication	44.2	% Duration Communication	45.3	% Communication	76.1
% Communication	55.3	% Communication	48.1	% Communication	37.7	% Communication	37.8	RMSSD Unique Apps	74.7
Duration Social	51.9	Count Communication	42.2	Day	36.4	Day	36.1	Total App Events	61.8
Total App Events	46.8	Five Minute Sliding Window Q95	38.9	Five Minute Sliding Window Q95	34.7	Five Minute Sliding Window Q95	31.3	Five Minute Sliding Window Q95	58.4
Count Communication	45.1	Total App Events	38.1	% Social	28.3	% Social	28.6	Count Communication	52.0

Note. Top ten most important features driving each of the five lagged random forest model predictions (most important = 100, least important = 0). Ensemble importance scores in bold underneath each of the lag column headers represent the importance of that model's predictions to the final ensemble random forest prediction of boredom.