**­Supplementary Online Content**

**Predicting states of elevated negative affect in adolescents from smartphone sensors:**

**A novel personalized machine learning approach**

**Supplemental Introduction**

In contrast to the substantial body of research indicating increased frequency of negative emotions in adolescence, there is surprisingly little research investigating the duration of these emotional states. Existing studies, primarily conducted in adult samples, indicate that negative emotional states commonly last hours (Fan et al., 2019; Thornton & Tamir, 2017; Verduyn et al., 2009, 2011, 2012, 2015; Verduyn & Lavrijsen, 2015), with certain negative emotions (e.g., sadness) lasting longer than others (e.g., fear and disgust), on average (Verduyn & Lavrijsen, 2015). In summary, a growing body of research indicates that as children transition into the adolescent years, they are confronted with an increasing frequency of negative emotional states, which commonly last hours and, over time, may increase their risk of developing an emotional disorder. Accordingly, there is an acute need to develop data-driven approaches to reliably predict and ultimately interrupt states of markedly elevated negative emotions as they occur in the daily lives of teens. In addition to the immediate benefits of alleviating acute states of affective distress, reducing the frequency and duration of episodes of high negative affect may serve to reduce the risk of future onset of emotional disorders.

**Supplemental Methods**

 **Participants.** History or current diagnosis of any of the following DSM-5 psychiatric illnesses were exclusionary for the anhedonic (AH) group: all psychotic disorders, bipolar disorder, anorexia nervosa or bulimia nervosa, obsessive-compulsive disorder, substance (including alcohol) use disorder within the past 12 months or lifetime severe substance use disorder, or chronic depression (current episode > 2 years). Anxiety disorders were allowed. For typically developing (TD) participants, additional exclusion criteria included a history of any DSM-5 psychiatric or substance-related disorder, first-degree relative diagnosed with MDD, bipolar disorder, or a psychotic disorder, and current use of any psychiatric medications.

**Snaith-Hamilton Pleasure Scale.** The Snaith-Hamilton Pleasure Scale (SHAPS)(Snaith et al., 1995) is a 14-item self-report measure assessing anhedonia within several domains (e.g., “I would find pleasure in my hobbies and past-times”). Participants rated the extent to which they agreed with each statement on a 4-point scale ranging from 1 (strongly agree) to 4 (strongly disagree). A dimensional scoring approach was used for analyses (possible range, 14-56), with a higher score indicating a higher level of anhedonia.

**Center for Epidemiologic Studies Depression Scale.** The Center for Epidemiologic Studies Depression Scale (CES-D)(Radloff, 1977) is a 20-item self-report measure assessing depressive symptom severity over the past week on a 4-point scale ranging from 0 (rarely or none of the time to <1 day) to 3 (most or all of the time to 5-7 days). Total scores range from 0 to 60. A higher score indicates greater severity of depressive symptoms, with four items being reverse scored.

**Schedule for Affective Disorders and Schizophrenia for School-Age Children.** The Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS-PL)(Kaufman et al., 1997) is a semi-structured clinical interview that assesses current and past psychiatric disorders in accordance with the DSM-5. Postdoctoral fellows or research assistants of bachelor’s degree level conducted the interviews under supervision and after receiving at least 40 hours of training.

**Ecological Momentary Assessment.** The EMA data collection period began after an in-person assessment visit to our lab. During this assessment session, the Metricwire app was installed on participant’s phones to collect the EMA data. Following this in-person assessment session, EMA surveys were delivered 2-3 times per day (outside of school hours), Thursday through Monday, every other week during the course of the study. Thursday through Monday was selected to sample adolescent affect on both weekdays and the weekend (for similar EMA designs in adolescents, see e.g., Forbes et al., 2009, 2012; Webb, Israel, et al., 2021). Surveys were triggered using a time-stratified random sampling strategy (i.e., teens were signaled once at a random time during two timeslots on weekdays [4pm to 6:30pm and 6:30pm to 9:00pm], and three timeslots on weekends [11am and 4pm, 4pm to 6:30pm and 6:30pm to 9:00pm]. The mean number of observations per subject is 52.1 with a standard deviation of 34.8 (range 7-126).

**Computing Sleep Episodes from Passive Data.** Sleep episodes were calculated by running 150-minute forward and backward moving windows over the phone use data to find epochs of continuous low signal. When the data is missing, the sleep episode was limited to the average low phone use epoch of the participant over the entire study (Staples et al., 2017)

**Supplemental Results and Figures**

For both approaches (GLMER and PEM), we first performed a principal component analysis (PCA) of the 14 smartphone-derived predictors in the merged dataset of all subjects and kept the top five principal components (PCs) that together explain 90% of the variability. We then used the selected PCs as the predictors. For each emotion, we only considered subjects with at least four HNA states so that the training data in each cross-validation step contains sufficient positive and negative observations. In Figure S2, we plot the loadings of each principal component (PC) to illustrate how each of the original 14 smartphone variables contribute to a particular PC.

In Figure S3, we plot the occurrence of high anger, sadness and nervousness states for all participants separately over the full study period along with the predictions from the random forest (RF) personalized ensemble model (PEM-RF). We focus on the relatively simpler PEM-RF given that the PDEM (i.e., ensemble of each algorithm) yielded similar predictive performance. The optimal cut-off value is used to determine the predicted status (HNA state: yes or no). We label high negative affect (HNA) states in red and non-HNA states in green. Incorrect predictions are marked with a cross.

In Figure S4, we illustrate the feature importance of different variables in the random forest idiosyncratic models (IM-RFs). The results suggest that activity level (“accelerometer score”) was the strongest predictor of HNA for all three emotions. However, there was substantial heterogeneity in the directionality of effects (see red vs. blue coloring which reflects a positive vs. negative association, respectively).

In Figure S5, we plot the combination weights $\hat{w}^{k}$ of IM-RFs for each PEM-RF. Notice that if PEM-RF does not borrow information from other subjects, all weights (red) would be assigned to the squares along the diagonal dotted line. The figures confirm that most of the PEM-RFs use information from the models of both the target subject and other subjects (see main text for additional detail).

In the main text, we used regular 10-fold cross-validation (CV) to evaluate the performance of the personalized prediction model even though the data for each subject is a time series. The reason for this choice is that our models assume that the observations are independent. We report the average prediction accuracy across all folds. We also implemented a time-series CV that uses all data before time $t$ to train the personalized prediction models and evaluates their performance using data at time $t$. In Figure S6, we show the Receiver Operating Characteristic curves of PEMs when time-series CV is used instead of regular 10-fold CV. We use every time point from the last quarter of the subject-specific time series as the evaluation set and calculate the average time-point-specific prediction accuracies over this set. For each time point $t$ in the evaluation set, the models are trained using all observations before time $t$ and evaluated using data at time $t$. We then calculate the average time-point-specific prediction accuracies over this set to characterize the performance of a model. As seen in Table S1, most of the relative predictive performances of different models remain the same in the time-series CV (relative to the results presented in the main text), while PDEM tends to have slightly worse performance than PEM-RF for HNA predictions in sadness and nervousness. Compared with the results from standard CV, time-series CV tends to produce smaller AUCs for each model. This is likely because that the average number of observations used for training in each iteration of time-series CV is substantially smaller than that in the standard CV, which takes advantage of the entire dataset.

In Figure S7, we visualize the proportion of predicted HNA states for each emotion when the threshold probabilities (*p\**) vary from 0 to 1. In the regions where PEM-RF/PDEM outperform all other methods in the decision curve analysis presented in the main text ($p^{\*}≈0.2-0.4$), the proportions of HNA states identified are at reasonable levels.

In Figure S8 and Table S2, we illustrate the ROC curves and AUC under the ROC curves from a sensitivity analysis where the definition of HNA stats is based on subject-specific quantiles: an HNA state is reached for a subject when the current emotion score is higher than 75% quantile of all their observed emotion scores. We repeated the analyses in the manuscript for anger, sadness, and nervousness with this new definition of HNA, which yielded similar results.

References

Fan, R., Varol, O., Varamesh, A., Barron, A., van de Leemput, I. A., Scheffer, M., & Bollen, J. (2019). The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nature Human Behaviour*, *3*(1), 92–100. https://doi.org/10.1038/s41562-018-0490-5

Forbes, E. E., Hariri, A. R., Martin, S. L., Silk, J. S., Moyles, D. L., Fisher, P. M., … Dahl, R. E. (2009). Altered Striatal Activation Predicting Real-World Positive Affect in Adolescent Major Depressive Disorder. *American Journal of Psychiatry*, *166*(1), 64–73. https://doi.org/10.1176/appi.ajp.2008.07081336

Forbes, E. E., Stepp, S. D., Dahl, R. E., Ryan, N. D., Whalen, D., Axelson, D. A., … Silk, J. S. (2012). Real-World Affect and Social Context as Predictors of Treatment Response in Child and Adolescent Depression and Anxiety: An Ecological Momentary Assessment Study. *Journal of Child and Adolescent Psychopharmacology*, *22*(1), 37–47. https://doi.org/10.1089/cap.2011.0085

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., … Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(7), 980–988. https://doi.org/10.1097/00004583-199707000-00021

Radloff, L. S. (1977). The CES-D Scale A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *The British Journal of Psychiatry*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99

Staples, P., Torous, J., Barnett, I., Carlson, K., Sandoval, L., Keshavan, M., & Onnela, J.-P. (2017). A comparison of passive and active estimates of sleep in a cohort with schizophrenia. *NPJ Schizophrenia*, *3*, 37. https://doi.org/10.1038/s41537-017-0038-0

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982–5987.

Verduyn, P., Delaveau, P., Rotgé, J.-Y., Fossati, P., & Van Mechelen, I. (2015). Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, *7*(4), 330–335.

Verduyn, P., Delvaux, E., Van Coillie, H., Tuerlinckx, F., & Van Mechelen, I. (2009). Predicting the duration of emotional experience: Two experience sampling studies. *Emotion*, *9*(1), 83.

Verduyn, P., & Lavrijsen, S. (2015). Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion*, *39*(1), 119–127.

Verduyn, P., Van Mechelen, I., Kross, E., Chezzi, C., & Van Bever, F. (2012). The relationship between self-distancing and the duration of negative and positive emotional experiences in daily life. *Emotion*, *12*(6), 1248.

Verduyn, P., Van Mechelen, I., & Tuerlinckx, F. (2011). The relation between event processing and the duration of emotional experience. *Emotion*, *11*(1), 20.

Webb, C. A., Israel, E. S., Belleau, E., Appleman, L., Forbes, E. E., & Pizzagalli, D. A. (2021). Mind-Wandering in Adolescents Predicts Worse Affect and Is Linked to Aberrant Default Mode Network–Salience Network Connectivity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(3), 377–387. https://doi.org/10.1016/j.jaac.2020.03.010

**Table S1**

AUCs of different models when predicting HNA states for each of the three emotions using time-series cross-validation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Anger | Sadness | Nervousness |
| GLMER | 0.60 | 0.55 | 0.52 |
| PEM-ENet | 0.63 | 0.54 | 0.55 |
| PEM-SVM | 0.63 | 0.58 | 0.57 |
| PEM-RF | 0.69 | 0.59 | 0.66 |
| PDEM | 0.70 | 0.57 | 0.62 |

**Table S2**

AUCs of different models when predicting HNA states defined with subject-specific 75% quantiles. Regular subject-level 10-fold cross-validation was used to derive the ROC curves.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Anger | Sadness | Nervousness |
| GLMER | 0.60 | 0.57 | 0.56 |
| PEM-Enet | 0.65 | 0.60 | 0.67 |
| PEM-SVM | 0.59 | 0.57 | 0.65 |
| PEM-RF | 0.74 | 0.64 | 0.71 |
| PDEM | 0.73 | 0.64 | 0.70 |

**Figure Captions**

***Figure S1.*** Workflow of the Personalized Ensemble Model (PEM) training approach via stacking of Idiosyncratic Models (IMs).

***Figure S2.*** Loadings of the top five principal components (PCs) for each emotion.

***Figure S3.*** Per-subject prediction results for the HNA states of anger, sadness, and nervousness. Red indicates actual HNA states and green non-HNA states. Incorrect predictions are marked with crosses.

***Figure S4:***Feature importance of each variable derived from random forest idiosyncratic models (IM-RFs) for each emotion.

***Figure S5.*** Combination weights assigned to each idiosyncratic random forest model (IM-RF) in each participant’s personalized ensemble random forest model (PEM-RF). Each row represents a PEM-RF and each column an IM-RF. The weights are bounded between 0 and 1.

***Figure S6:***Receiver operating characteristic (ROC) curve of different models when time-series cross-validation (CV) is used.

***Figure S7***. Proportion of predicted HNA states across subjects when the threshold (*p*\*) changes.

***Figure S8***. ROC curve of different models when HNA is defined using subject-specific 75% quantiles.

















**