DIGITAL HEALTH
Volume 10: 1–17
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241256730
journals.sagepub.com/home/dhj

# Multiclassification of the symptom severity of social anxiety disorder using digital phenotypes and feature representation learning

Hyoungshin Choi[1,2], Yesol Cho[3], Choongki Min[1], Kyungnam Kim[1], Eunji Kim[3], Seungmin Lee[3] and Jae-Jin Kim[3,4]

## Abstract

**Objective:** Social anxiety disorder (SAD) is characterized by heightened sensitivity to social interactions or settings, which disrupts daily activities and social relationships. This study aimed to explore the feasibility of utilizing digital phenotypes for predicting the severity of these symptoms and to elucidate how the main predictive digital phenotypes differed depending on the symptom severity.

**Method:** We collected 511 behavioral and physiological data over 7 to 13 weeks from 27 SAD and 31 healthy individuals using smartphones and smartbands, from which we extracted 76 digital phenotype features. To reduce data dimensionality, we employed an autoencoder, an unsupervised machine learning model that transformed these features into low-dimensional latent representations. Symptom severity was assessed with three social anxiety-specific and nine additional psychological scales. For each symptom, we developed individual classifiers to predict the severity and applied integrated gradients to identify critical predictive features.

**Results:** Classifiers targeting social anxiety symptoms outperformed baseline accuracy, achieving mean accuracy and F1 scores of 87% (with both metrics in the range 84–90%). For secondary psychological symptoms, classifiers demonstrated mean accuracy and F1 scores of 85%. Application of integrated gradients revealed key digital phenotypes with substantial influence on the predictive models, differentiated by symptom types and levels of severity.

**Conclusions:** Leveraging digital phenotypes through feature representation learning could effectively classify symptom severities in SAD. It identifies distinct digital phenotypes associated with the cognitive, emotional, and behavioral dimensions of SAD, thereby advancing the understanding of SAD. These findings underscore the potential utility of digital phenotypes in informing clinical management.

## Keywords

Digital phenotype, social anxiety disorder, autoencoder, representation learning, integrated gradient, multiclassification

Submission date: 23 November 2023; Acceptance date: 7 May 2024

[1]AI Medtech R&D, Waycen Inc, Seoul, Republic of Korea
[2]Department of Electrical and Computer Engineering, Sungkyunkwan University and Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, Republic of Korea
[3]Institute of Behavioral Sciences in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea
[4]Department of Psychiatry, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

HC and YC contributed equally to this work.

**Corresponding author:**
Jae-Jin Kim, Department of Psychiatry, Gangnam Severance Hospital, Yonsei University College of Medicine, 211 Eonju-ro, Gangnam-gu, Seoul, 06273, Republic of Korea.
Email: jaejkim@yonsei.ac.kr

## Introduction

Social anxiety disorder (SAD) is characterized by intense fear and avoidance of social situations, often accompanied by physical symptoms such as rapid heartbeat, sweating, and trembling.[1,2] This anxiety often triggers negative thought patterns, resulting in self-doubt and compromised self-esteem during social interactions, which contribute to a perpetuating cycle of anxiety and avoidance.[3] Such patterns can severely disrupt daily activities and social relationships, leading to decreased quality of life, delayed medical treatment, and even exacerbation in social, academic, and occupational areas.[4–6] Moreover, if left untreated, the risk for comorbid psychiatric conditions like depression and other anxiety disorders also increases.[5,6] Given that SAD significantly impacts daily activities, maladaptive behaviors such as the avoidance of social settings and increased time spent at home can be quantitatively assessed through continuous monitoring.

In this context, digital phenotypes have emerged as valuable metrics for capturing one's behavioral and physiological patterns. Derived from data gathered through smartphones and wearable sensors, these digital phenotypes offer insights into the emotional and mental states experienced in daily life.[7,8] In existing studies investigating the relationship between digital phenotypes and mental health, there are attempts to investigate linear relationships between digital phenotypes and mental states.[9–11] However, the constraints of a linear approach can lead to an oversimplification of the intricate interactions between digital phenotypes and the levels of symptoms associated with mental health.

To overcome these limitations, machine learning techniques offer a powerful toolset for analyzing complex relationships. In particular, machine learning methods have been proposed for the analysis of mental disorders, and some studies have employed these methods to classify individuals based on digital phenotypes.[12–15] Identifying latent features of digital phenotypes, low-dimensional representations using dimensionality reduction technique, which can simplify the complex dataset to facilitate better understanding. For using latent features, one of the techniques identifying the characteristic of data is latent class analysis, which focuses on categorical data.[16] Digital phenotypes include continuous values and are so complicated that dimensionality reduction techniques are inevitable for analysis. Autoencoders, a type of deep learning-based neural network, have shown promise in employing nonlinear compression and reconstruction to generate latent features that effectively represent the original data.[17,18] This technique has been successfully applied in various medical domains, such as differentiating autism spectrum disorder through magnetic resonance imaging data,[19] depression through speech analysis,[20] and anxiety disorder through sleep-related digital phenotypes.[21]

We identified two previous studies that applied machine learning techniques using digital phenotypes to predict the severity of SAD symptoms with validated psychological scales. One study utilized GPS data to predict levels of social anxiety,[22] while another study applied accelerometer and phone/text interactions for a similar objective.[23] However, these studies were limited by their reliance on nonclinical samples, specifically college students, and focused solely on the Social Interaction Anxiety Scale, which assesses primarily emotional aspects of social interaction anxiety.[24] Additionally, these analyses included only a select few behavioral traits as captured by digital phenotypes.

To address these issues, the present study conducts a comprehensive analysis encompassing cognitive, behavioral, and emotional symptoms related to SAD with a clinical sample. It leverages a series of machine learning models, each one designed to process and learn from a unique combination of digital phenotypes, such as GPS activity, phone call logs, and app usage metrics. Correspondingly, individual models are each designated to predict a particular symptom measurement, including levels of fear and avoidance of social situations, fear of negative evaluation by others, depression, and life satisfaction. By employing these predictive models on a dataset comprising individuals diagnosed with SAD as well as those from a healthy control group, enabling a comparative analysis across the severity groups.

With these in mind, this study postulates that autoencoders and feature representation learning techniques could provide a novel methodological framework for uncovering the latent characteristics of SAD. This enhanced understanding of symptom severity through digital phenotyping could, in turn, contribute to the development of more precise diagnostic and therapeutic protocols for SAD.

## Methods

### *Participants and demographic characteristics*

During the recruitment period from October 2021 to May 2023, a total of 88 Koreans, consisting of 49 patients with SAD and 39 healthy control (HC) participants living in Seoul or nearby areas, voluntarily applied to participate in the study through an outpatient clinic or online platforms. Eligibility for the SAD group required a previous diagnosis of SAD, and this diagnosis was confirmed using Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) criteria and a Mini-International Neuropsychiatric Interview (MINI) conducted by a clinical psychologist. Exclusion criteria for SAD group included the presence of substance and alcohol use disorders, brain injuries or neurological diseases, intellectual disabilities, left-handedness, or the presence of metallic implants in the body. HC participants had neither a psychiatric history nor current symptoms according to DSM-V criteria.

Among the recruited participants, four withdrew their consent to participate in the study, and 10 were disqualified due to measurement duration less than 2 weeks. Data were collected from the remaining participants, but data from 16 of these were excluded from analysis due to incomplete data across one or more types. Consequently, our study included a total of 58 participants, with 27 patients with SAD and 31 HC participants, with ages ranging from 19 to 49 years. Demographically, the SAD group had a mean age of 28.37 (SD 6.57) years, with 52% female, while the HC group had a mean age of 28.41 (SD 7.34) years, with 58% female (detailed demographics in Table 1). Statistical tests revealed no significant differences in age ($t_{56} = 0.03$, $P = .98$) and sex ($\chi^2_1 = 0.04$, $P = .83$) between the two groups; however, the effects of age and sex were controlled for in the following analysis.

**Table 1.** Participant demographics (N = 58).

| Variable | Group | |
|---|---|---|
| | SAD (n = 27) | HC (n = 31) |
| Sex, n (%) | | |
| Male | 13 (48%) | 13 (42%) |
| Female | 14 (52%) | 18 (58%) |
| Age, mean (standard deviation) | | |
| Years | 28.4 (6.6) | 28.4 (7.3) |
| Education, n (%) | | |
| <High school | 10 (37%) | 10 (32%) |
| >College | 17 (63%) | 21 (68%) |
| Occupation, n (%) | | |
| Employed | 25 (93%) | 29 (94%) |
| Unemployed | 2 (7%) | 2 (6%) |
| Comorbidity, n (%) | | |
| PDD and MDD | 10 (37%) | N/A |
| OCD | 3 (11%) | N/A |
| GAD | 1 (4%) | N/A |
| PD | 4 (15%) | N/A |

GAD: Generalized Anxiety Disorder; HC: healthy control; MDD: major depressive disorder; N/A: not applicable; OCD: obsessive compulsive disorder; PD: panic disorder; PDD: persistent depressive disorder; SAD: social anxiety disorder.

All participants submitted written informed consent to participate in the study before installing a custom-developed app on their smartphones. They were informed that they could withdraw from the study at any point, and that data collection could be terminated at their discretion by uninstalling the app. This study protocol was approved by the institutional review board of Yonsei University Gangnam Severance hospital.

## Data collection

Various types of behavioral and physiological data were collected from participants over a period ranging from 7 to 13 weeks (mean, 8.5; SD, 1.8). Digital phenotype data were automatically collected through a noncommercial app ('HAID', Waysen Co., Ltd), developed by our research team and installed on the participants' own Android smartphone and a commercial smart band ('DOFIT', Medi Plus Solution Co., Ltd) provided to them. All participants stayed in South Korea during the data collection period, and thus there was no disruption to data collection.

To assess the severity of symptoms in SAD, participants reported on self-report psychological scales once every 2 weeks, facilitating an in-depth analysis of the utility of digital phenotypes. As social anxiety-specific scales, the Liebowitz Social Anxiety Scale (LSAS)[25] with fear and avoidance subfactors was used to assess the emotional and behavioral aspects of social anxiety, and the Brief Fear of Negative Evaluation Scale (BFNE)[26] was used to measure the cognitive dimension, specifically the fear of negative evaluation by others. Additionally, secondary psychiatric symptom measurements included the Hospital Anxiety and Depression Scale,[27] Generalized Anxiety Disorder-7,[28] Panic Disorder Severity Scale (PDSS),[29] Maudsley Obsessive Compulsive Inventory,[30] Penn State Worry Questionnaire (PSWQ),[31] State-Trait Anxiety Inventory,[32] and Satisfaction with Life Scale (SWLS).[33] We confirmed the copyright of all tools and psychological scales used in the study.

## Extraction of digital phenotype data

As shown in Table 2, there were eight distinct types of digital phenotype data, from which 76 features were extracted to represent the participants' underlying behavioral and physiological patterns. These phenotypes were extracted at 2-week intervals, with a 1-week overlap between consecutive segments (e.g. weeks 1–2, weeks 2–3, etc.) to effectively capture transitions in behavioral and physiological patterns over time. Nonlinear imputation called K-Nearest Neighbors Algorithm[34] was applied to

**Table 2.** Data and digital phenotype descriptions.

| Data | Digital phenotypes | Descriptions |
|---|---|---|
| App log | Sum of usage | Extracted data by categorizing 10 types of apps (total, camera, communication (SNS, messaging), entertainment (OTT, streaming), music, trip (reservation, airport), shopping, culture, religion, and game) |
| | Mean of usage | |
| | Number of usage | |
| Phone log | Sum of on-time | |
| | Mean of on-time | |
| | Number of on-time | |
| | Daily number of on-time | |
| Call log | Call duration | Extracted in 3 types (total, incoming, and outgoing) |
| | Call max at one time | |
| | Number of people | |
| | Number of call | |
| | Missed ratio | |
| | In-out ratio | |
| | Call entropy | |
| GPS | Local variance | Location variance $= \log(\sigma_{\text{lat}}^2 + \sigma_{\text{long}}^2)$[b] |
| | Circadian movement | Circadian movement $= \log(E_{\text{lat}} + E_{\text{long}}\ )\ E = \frac{1}{i_u - i_L} \sum_{i=i_L}^{i_u} psd(\ f_i\ )$[c] |
| | Speed | Speed $= \sqrt{\left(\frac{\text{lat}_i - \text{lat}_{i-1}}{t_i - t_{i-1}}\right)^2 + \left(\frac{\text{long}_i - \text{long}_{i-1}}{t_i - t_{i-1}}\right)^2})$[d] |
| | Distance | Distance $= \sum_i \sqrt{(\text{lat}_i - \text{lat}_{i-1})^2 + (\text{long}_i - \text{long}_{i-1})^2}$ |
| | Cluster number | |
| | Location entropy | Entropy $= -\sum_{i=1}^{N} p_i \log\ (\ p_i)$[e] |
| | Home stay | |
| | Transition | |
| Noise log | Ambient noise level | |
| Lux log | Ambient brightness level | Extracted in 3 types (total, daytime, and nighttime)[e] |
| Heart rate | Mean heart rate | |
| Gyroscope | Mean sleep duration | |
| | Mean sleep latency | |

SNS: Social Network Service; OTT: over-the-top media service; GPS: Global Positioning System.

[a] lat: latitude; long: longitude; σ: variance of latitude and longitude values.

[b] *f*: least-squares spectral analysis (the Lomb-Scargle method); psd: the power spectral density.

[c] *t*: time point.

[d] *p*: the percentage of time spent at location *i* (%); *N*: cluster number.

[e] daytime: sunrise to sunset; nighttime: sunset to sunrise.

fill in the remaining missing data. Finally, the analysis included 511 overlapping 2-week segments. Table S1 in Supplemental Appendix presents the number of samples in each participant.

(A) App usage: We quantified app usage based on logs, calculating metrics over a 2-week span including the total duration of app usage (i.e. sum of usage), the average time per session (i.e. mean of usage), and the frequency of app sessions (i.e. number of usage). Data were classified into 10 categories based on their primary function to investigate their impact on symptom severity.

(B) Phone usage: Data on interactions with the phone were extracted from phone logs. Metrics such as total phone usage duration over 2 weeks (i.e. sum of on-time), average time per interaction (i.e. mean of on-time), and the frequency of phone interactions (i.e., number of on-time) were calculated.

(C) Call patterns: Call logs were analyzed to understand interactions with other individuals. Metrics including total call duration, maximum duration of a single call (i.e. call max at one time), the number of unique individuals called (i.e. number of people), frequency of calls (i.e. number of call), missed call ratio, incoming-to-outgoing call ratio (i.e. in–out ratio), and variability in call duration (i.e. call entropy) were calculated. These metrics were gathered for both incoming and outgoing calls.

(D) Movement patterns: Spatial behavior was ascertained via GPS data. Metrics such as the diversity of locations visited (i.e. local variance), regularity of movement (i.e. circadian movement), speed, total distance covered (i.e. distance), the number of unique locations visited (i.e. cluster number), diversity in time spent at locations (i.e. location entropy), duration spent at home (i.e. home stay ratio), and transition times (i.e. transition ratio) were computed.

(E) Environmental patterns: Given the correlation between anxiety disorders and environmental factors like noise and light,[35,36] we measured ambient noise and brightness levels. The 2-week average levels of ambient noise and brightness were calculated. Recognizing the differential impact of light during daytime and nighttime, we also computed the average light levels during these respective periods (i.e. daytime and nighttime).

(F) Physiological patterns: Physiological metrics such as the heart rate, sleep duration, and sleep latency were measured. The average heart rate in beats per minute over a 2-week span was calculated. Sleep metrics were derived from gyroscope and lux log data; specifically, sleep onset and wake times were inferred from sharp changes in light levels and movements.[37] Sleep latency was calculated as the time lapse between the moment the light was turned off and the onset of sleep. The 2-week averages for sleep duration and latency were computed.

## Model development and evaluation

We designed and evaluated models to classify the levels of clinical symptoms by leveraging characteristics identified within digital phenotypes over a 2-week period. In these models, the digital phenotype data were paired with symptom measurement scores in a cross-sectional manner for the purpose of model training. The models comprised three primary components: an autoencoder tasked with capturing the underlying structure of digital phenotypes (Figure 1a), symptom severity classifiers that utilize latent features as inputs extracted by the autoencoders (Figure 1b), and integrated gradients (IGs) that assess the contributions of digital phenotypes to the prediction of symptom severity (Figure 2a).

## Architecture of the autoencoder model

Using an autoencoder, latent features were generated from digital phenotype. The autoencoder model is defined as follows:

$$h = e\,(U) = \mathrm{LeakyReLU}(WU + b),$$

$$U' = d\,(h) = \mathrm{LeakyReLU}(Wh + b),$$

$$L = \sum \frac{|U - U'|^2}{n}$$

where,$e$  is the encoder, which takes an input vector $U$ of size $n$ and maps it to a lower-dimensional representation called the hidden or bottleneck layer $h$. This hidden layer is then used by the decoder $d$  to reconstruct the input vector $U$ and generate $U$’ using nonlinear activation functions. The autoencoder is trained by minimizing the sum of mean square errors $L$ between the input and output vectors, where the input vector $U$ is compared to the reconstructed output vector $U$’. The architecture of the autoencoder model consists of a total of five layers, including two encoder layers, two decoder layers, and one latent feature layer. In all layers, we utilized the Leaky rectified linear unit (LeakyReLU) activation function, and we applied dropout in the input layer.[38] The model was optimized using the AdamW optimizer with a learning rate of 1e,[3] batch size of 15,500 epochs.[39] We calculated linear correlations between $U$ and $U$^’ to assess the reconstruction process of the autoencoder model. Additionally, we generated latent features by slightly modifying the architecture: (i) by subtracting one layer from both the encoder and decoder (i.e. extra model 1) and (ii) by adding one layer to both the encoder and decoder (i.e. extra model 2), to compare the performance of different model architectures. To evaluate all datasets, we used five-fold cross-validation.

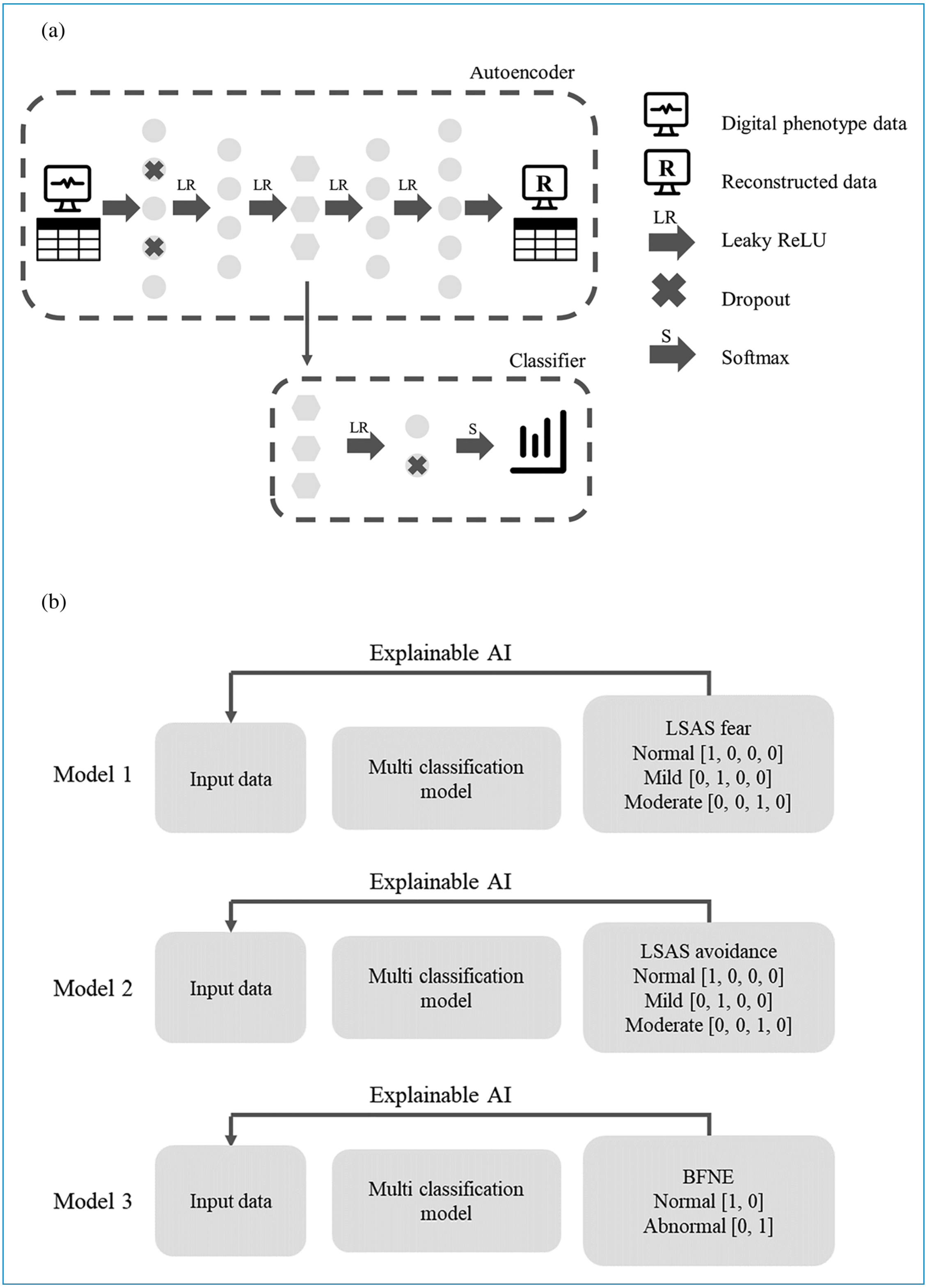

**Figure 1.** Symptom severity classification using the autoencoder-based feature representation. (a) Schematic of the autoencoder model. The autoencoder model learns latent features of the digital phenotype, and the classifier predicts symptom severity using these latent features. (b) The schema of the classification model for each questionnaire. The predicted output of each model is interpreted using an explainable AI technique.

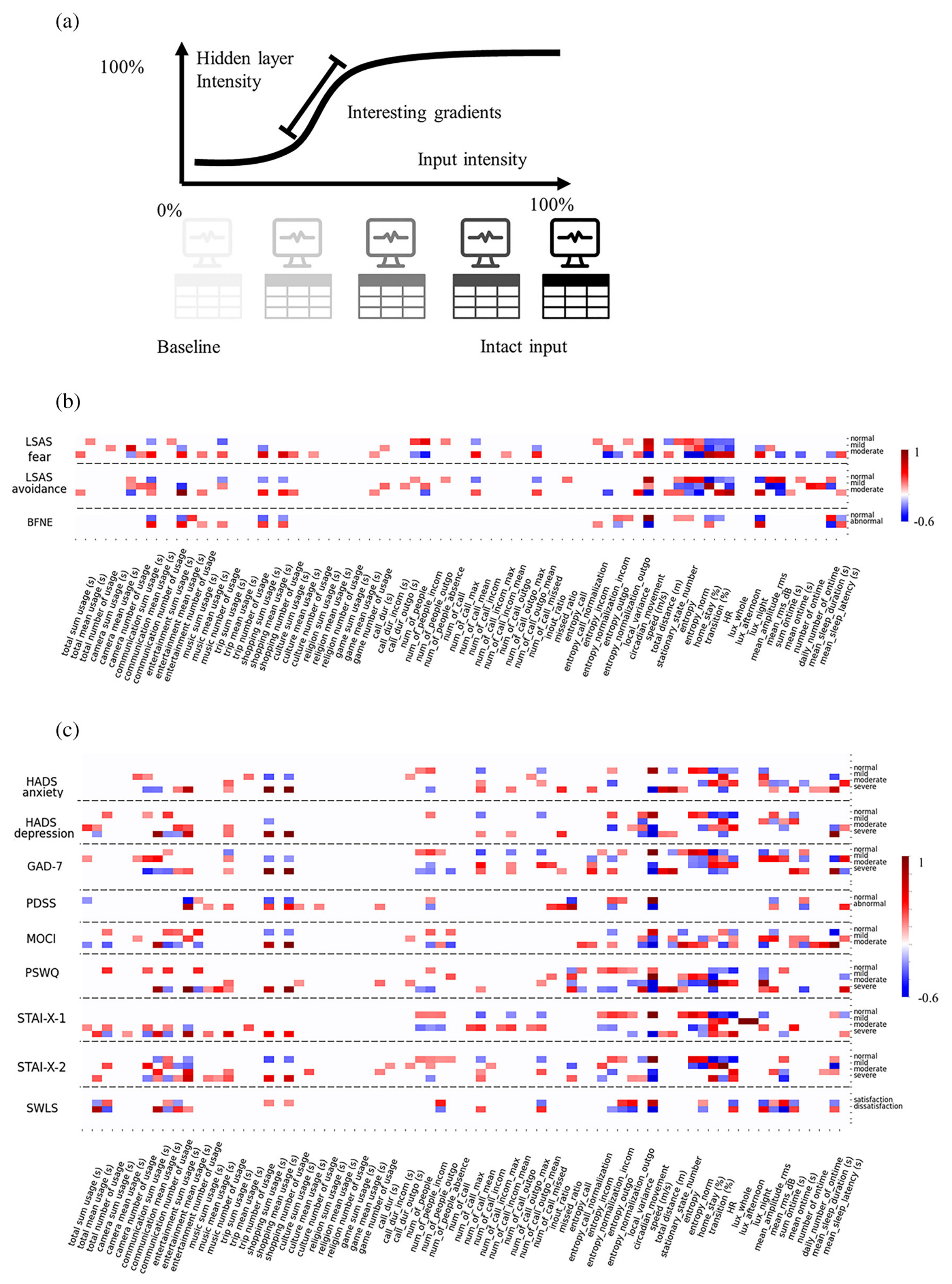

Figure 2. The attribution of digital phenotype using the integrated gradient technique. (a) The integrated gradient technique computes the attribution of the input in predicting the output. It achieves this by averaging the contributions as the intensities of the input changes. (b) The attribution of digital phenotypes by social anxiety disorder symptom measurements using integrated gradients. Effects estimated by symptom severity class based on the social anxiety disorder symptom measurements models. (c) Comparative plots showing the attribution of digital phenotypes to secondary psychiatric symptom measurements.

**Table 3.** Symptom measurement scales.

| Scale | Description | Class |
|---|---|---|
| *Social anxiety disorder symptom measurements* | | |
| LSAS fear | Measures the level of fear experienced by individuals in various social situations | Normal [0~14], Mild [15~29], Moderate [30~] |
| LSAS avoidance | Measures the extent to which individuals avoid social situations due to anxiety or fear | Normal [0~14], Mild [15~29], Moderate [30~] |
| BFNE | Measures the level of anxiety associated with the anticipation of negative evaluation from others | Normal [0~47], Abnormal [48~] |
| *Secondary psychiatric symptom measurements* | | |
| HADS anxiety | Assesses the degree of anxiety symptoms | Normal [0~7], Mild [8~10], Moderate [11~14], Severe [15~] |
| HADS depression | Assesses the severity of depressive symptoms | Normal [0~7], Mild [8~10], Moderate [11~14], Severe [15~] |
| GAD-7 | Measures the severity of generalized anxiety disorder symptoms | Normal [0~4], Mild [5~9], Moderate [10~14], Severe [15~21] |
| PDSS | Measures the severity and frequency of panic disorder symptoms | Normal [0~8], Abnormal [9~] |
| MOCI | Measures the severity of obsessive-compulsive symptoms | Normal [30~43], Mild [44~48], Moderate [49~] |
| PSWQ | Measures the severity and frequency of worry symptoms | Normal [0~29], Mild [30~52], Moderate [53~65], Severe [66~] |
| STAI-X-1 | Measures temporary, situational anxiety levels | Normal [0~51], Mild [52~56], Moderate [57~61], Severe [62~] |
| STAI-X-2 | Measures long-term, generalized anxiety as a personality trait | Normal [0~51], Mild [52~56], Moderate [57~61], Severe [62~] |
| SWLS | Evaluates subjective wellbeing by measuring individuals' satisfaction with their lifestyle | Dissatisfaction [0~19], Satisfaction [20~] |

BFNE: Brief Fear of Negative Evaluation Scale; GAD-7: Generalized Anxiety Disorder-7; HADS: Hospital Anxiety and Depression Scale; LSAS: Liebowitz Social Anxiety Scale; MOCI: Maudsley Obsessive Compulsive Inventory; PDSS: Panic Disorder Severity Scale; PSWQ: Penn State Worry Questionnaire; STAI: State-Trait Anxiety Inventory; SWLS: Satisfaction with Life Scale.

## Architecture of the classifier model

The scores of the self-report scales were categorized based on established cut-off values (Table 3). These categorical classifications of symptom severity were implemented to perform an analysis that takes into account the current clinical framework used to recommend specific therapies.[40,41] See Table S2 in Supplemental Appendix for detailed descriptions of the severity levels.

The architecture of the classifiers was designed to predict multiple severity classes, as delineated by various questionnaires that a range of two to four classes (see Figure 1b for the schema of models and Tables S3 in Supplemental Appendix for score distributions). In this model, latent features in the hidden representation layer $h$ were used to classify symptom severity. The classifier model is defined as follows:

$$y = c(h) = \mathrm{softmax}(\mathrm{LeakyReLU}(Wh + b) + b),$$

$$L = -\sum_{i}^{\mathrm{label}} y_i \log \widehat{y_i}$$

where $i$ is the number of classes and c the classifier, which takes a layer $h$ from the autoencoder model, to predict symptom severity class. The classifier model has one linear layer and is trained by cross-entropy loss $L$ between true class $y_i$ and predicted class $\widehat{y_i}$. Each symptom severity score has its own classifier model. The model was optimized using the AdamW optimizer with a learning rate of $1e^3$, batch size of 15,300 epochs. For comparison, we generated predicted classes by (i) subtracting one layer from the classifier (i.e. extra model 3) or (ii) adding one layer to the classifier (i.e. extra model 4) and (iii) removing the dropout layer (i.e. extra model 5) to assess the performance of classifiers across different model architectures. Also, we applied the random forest classifier to digital phenotype instead of using an autoencoder model in order to evaluate the effect of latent features on profiling clinical and behavioral traits.

### IGs for model explanation

To elucidate the role of each digital phenotype in predicting symptom severity, we employed the IGs method for interpretability.[42] This technique quantifies the contribution of individual input variables—in our case, digital phenotypes—to the model's predictions (i.e. symptom severity classes). In particular, the IG for the $i^{th}$ neuron is mathematically expressed as:

$$\mathrm{IG}_i(x) = (x_i - z_i) * \int_{a=0}^{1} \frac{\delta f(z + a * (x - z))}{\delta x_i} da$$

where $x$ is the input data, $z$ the baseline, and $\alpha$ the interpolation constant. The path integral can be approximated as follows:

$$\mathrm{IG}_i(x) = (x_i - z_i) * \frac{1}{M} \sum_{M=1}^{M} \frac{\delta f\left(z + \frac{m}{M} * (x - z)\right)}{\delta x_i}$$

where $m$ and M are the number of steps in the scaled feature perturbation constant and Riemann sum approximation of the integral, respectively. If the attribution of the input data was high, we assumed that the output would have significantly changed and vice versa. Thus, it was possible to evaluate the digital phenotype that played a significant role in the feature representation learning processes of the input data.

### Assessment of digital phenotypes and statistical analysis

In pursuit of a comprehensive understanding of the digital phenotypes and their cognitive, emotional, behavioral, and physiological relationships through diverse scales for the target groups, our analysis was guided by three primary objectives. Initially, we assessed the values of IG to understand how digital phenotypes contribute to predictions across diverse symptoms. Subsequently, we identified key variables that characterize each severity level, focusing on variables with the largest absolute IG values, as outlined in previous reports.[43–45] Within this framework, we further concentrated on comparing the top three variables in each model, placing a particular emphasis on SAD symptom measurements. Finally, we validated the discriminatory power of the predicted results through statistical tests (t-tests or analysis of variance (ANOVA), as appropriate) for IG values across severity levels for each digital phenotype. To correct for multiple comparisons and control the false discovery rate (FDR), we employed the Benjamini-Hochberg procedure.[46]

## Results

### Evaluation of classification model performance

We generated latent features of digital phenotype from an autoencoder model, which demonstrated a significant correlation between the original and reconstructed digital phenotypes across the participants (mean $r = 0.98$, $\mathrm{SD} = 0.02$; $P < .001$; Figure S1 in Supplemental Appendix). The predictive models exhibited substantial accuracy in classifying severity of SAD symptom measurements, with results notably exceeding the baseline predictive accuracies within a range from 84% to 90% (Table 4). Furthermore, models for secondary psychiatric symptom measurements also yielded notable accuracies. For instance, the PSWQ model demonstrated an accuracy approximately 34% higher than baseline, and the PDSS model achieved the highest classification accuracy at 93% (Table 5). Moreover, when comparing the main model with five neural network models and one random forest classifier, the main model stood out, demonstrating superior performance across a variety of questionnaires (Tables S4–S9 in Supplemental Appendix).

### Outlining digital phenotypes across severity levels using IG

We employed the IG technique to identify digital phenotypes that affect the predictive outcomes in the hidden layer (Figure 2a). As a result, opposite directional attributions were observed depending on symptom severity. Specifically, in the context of questionnaires related to social anxiety symptom, we observed that lower levels of symptom severity were primarily associated with negative or near-zero IG values for variables such as communication, entertainment, and shopping app usage, along with home stay duration, transition duration, average heart rate, and night ambient brightness (i.e. lux night). Conversely, circadian movement, number of places visited (i.e. stationary state number), location entropy, and sleep duration showed higher positive values. Importantly, these

Table 4. Prediction results of social anxiety symptom severity class.

| Scale | Baseline accuracy[a] | Accuracy[b] | F1 score | Recall | Precision | Specificity |
|---|---|---|---|---|---|---|
| | | | | | | |
| LSAS fear | 0.35 | 0.87 | 0.87 | 0.87 | 0.87 | 0.93 |
| LSAS avoidance | 0.43 | 0.84 | 0.84 | 0.84 | 0.84 | 0.91 |
| BFNE | 0.82 | 0.90 | 0.90 | 0.90 | 0.91 | 0.83 |

[a]The baseline accuracy is a prediction only of the majority class.
[b]The accuracy is an average of cross-validation results.
BFNE: Brief Fear of Negative Evaluation Scale; LSAS: Liebowitz Social Anxiety Scale.

Table 5. Prediction results of secondary psychiatric symptom measurements severity class.

| Scale | Baseline accuracy[a] | Accuracy[a] | F1 score | Recall | Precision | Specificity |
|---|---|---|---|---|---|---|
| | | | | | | |
| HADS anxiety | 0.63 | 0.82 | 0.82 | 0.82 | 0.82 | 0.86 |
| HADS depression | 0.60 | 0.75 | 0.75 | 0.75 | 0.75 | 0.85 |
| GAD-7 | 0.67 | 0.86 | 0.86 | 0.86 | 0.86 | 0.89 |
| PDSS | 0.88 | 0.93 | 0.93 | 0.93 | 0.93 | 0.76 |
| MOCI | 0.81 | 0.92 | 0.92 | 0.92 | 0.92 | 0.85 |
| PSWQ | 0.44 | 0.78 | 0.78 | 0.78 | 0.78 | 0.90 |
| STAI-X-1 | 0.67 | 0.84 | 0.83 | 0.84 | 0.82 | 0.85 |
| STAI-X-2 | 0.72 | 0.86 | 0.86 | 0.86 | 0.86 | 0.84 |
| SWLS | 0.75 | 0.88 | 0.88 | 0.88 | 0.89 | 0.85 |

[a]The baseline accuracy is a prediction only of the majority class.
[b]The accuracy is an average of cross-validation results.
.
GAD-7: Generalized Anxiety Disorder-7; HADS: Hospital Anxiety and Depression Scale; MOCI: Maudsley Obsessive Compulsive Inventory; PDSS: Panic Disorder Severity Scale; PSWQ: Penn State Worry Questionnaire; STAI: State-Trait Anxiety Inventory; SWLS: Satisfaction with Life Scale.

directions are often reversed at high symptom severity (Figure 2b).

Furthermore, similar trends were observed in IG values of digital phenotypes to severity predictions across secondary psychiatric symptom measurements. When lower symptom severities, negative IG values predominantly were associated with missed call ratio, home stay duration, transition duration, average heart rate, and lux night. Conversely, positive IG values were linked to the distinct number of incoming calls, circadian movement, and the number of stationary states. Simultaneously, under conditions of higher symptom severity, a tendency was observed for the direction of the IG values to be opposite to those observed at lower symptom severity levels (Figure 2c).

## Ranking key digital phenotypes by absolute IG within severity levels

We evaluated the most influential digital phenotypes in predicting symptom severity by focusing on the top three features with the largest absolute values of IG. For the LSAS-fear scale that includes normal, mild, and moderate classes, home stay duration was a major contributor in both moderate and mild classes, while exhibiting opposite directional attribution. Notably, normalized location entropy was a major contributor to the prediction of moderate class. Unique to the mild class was camera number of usage, while the distinct number of incoming calls and location entropy were notable for the normal class. Across all severity levels, circadian movement was a consistent predictor (Table 6).

In the predictive model for the LSAS-avoidance scale, which includes normal, mild, and moderate classes, the duration of entertainment app usage and heart rate were crucial factors in predicting moderate class. For both mild and normal classes, time spent at home and ambient noise level were key variables, while indicating opposing attribution. Unique to the mild class was frequency of smartphone usage. Circadian movement served as a main predictor for both normal and moderate classes (Table 7).

In the predictive model for the BFNE scale, which consist of normal and abnormal classes, circadian movement and ambient brightness during nighttime emerged as consistently important features for both classes; however, these variables exhibited opposite directional attribution depending on the class. In addition, while mean sleep duration served as an important predictor in the normal class, the duration of communication app usage emerged pivotal for predicting the abnormal class (Table 8).

### *Validating discriminative power of digital phenotypes across severity classes*

We investigated the discriminatory power of digital phenotypes, using IG maps, to differentiate among levels of symptom severity by conducting $t$ tests and ANOVA. For questionnaires categorizing symptom severity into two classes such as BFNE, PDSS, and SWLS, two-tailed $t$ tests were applied. In contrast, for scales with multiple classes, ANOVA was employed. The analysis, following FDR correction, revealed statistically significant differences (adjusted $P < .001$) among the top 10 digital phenotypes across symptom severity classes, including key digital phenotypes, within SAD symptom measurements (Figure 3a). These phenotypes include app usage (i.e. shopping categories), call patterns (i.e. distinct number of incoming calls and number of call mean), movement patterns (i.e. circadian movement, entropy, duration of home stay, and transition duration), and physiological indicators (i.e. average heart rate). Furthermore, these distinctions were consistent across secondary psychiatric symptom measurements, thereby reinforcing the validity and efficacy of these digital markers for severity classification in mental health conditions (Figure 3b).

## Discussion

Our study advances the utility of digital phenotyping as a predictive tool for assessing symptom severity in SAD. Utilizing digital phenotypes, we were able to classify diverse symptom severities based on autoencoders for feature representation, thus demonstrating the digital phenotypes' discriminative ability. This aligns with existing research that showed the potential of digital phenotyping in healthcare,[12] specifically in understanding complex behaviors within clinical settings.[47,48] Crucially, our models demonstrated high classification accuracy not only for symptoms of social anxiety but also across a spectrum of co-occurring mental states, such as depression, general

**Table 6.** Comparison of key features' integrated gradient values across severity levels for the LSAS-fear scale (the italicized integrated gradient values are top three features indicative of a specific severity level).

| Feature | Normal | Mild | Moderate |
|---|---|---|---|
| Home stay duration | −0.32 | *−0.42* | *0.80* |
| Normalized location entropy | 0.27 | 0.41 | *−0.79* |
| Camera number of usage | −0.11 | *0.64* | −0.21 |
| Distinct number of incoming calls | *0.58* | 0.00 | −0.45 |
| Location entropy | *0.41* | 0.22 | −0.61 |
| Circadian movement | *0.73* | *0.64* | *−1.33* |

LSAS: Liebowitz Social Anxiety Scale.

**Table 7.** Comparison of key features' integrated gradient values across severity levels for the LSAS-avoidance scale (the italicized integrated gradient values are top three features indicative of a specific severity level).

| Feature | Normal | Mild | Moderate |
|---|---|---|---|
| Mean usage duration of entertainment app | 0.12 | −0.41 | *0.96* |
| Heart rate | −0.21 | −0.40 | *0.79* |
| Home stay duration | *−0.73* | *0.87* | −0.09 |
| Smartphone usage frequency | −0.13 | *0.53* | −0.13 |
| Mean ambient noise level (amplitude) | 0.54 | *−0.63* | 0.15 |
| Mean ambient noise level (dB) | *0.64* | −0.48 | −0.42 |
| Circadian movement | *0.80* | 0.22 | *−0.78* |

LSAS: Liebowitz Social Anxiety Scale.

anxiety or worry, panic attacks, obsession, and life satisfaction, thereby emphasizing the expansive potential of digital phenotyping in the characterization of various mental health conditions.

Addressing the challenge of model interpretability inherent in neural networks,[49,50] we employed IGs. This technique helped us identify key contributing digital phenotypes for predicting symptom severity, solving the model's intricate combinations of nonlinearities.[42,51–54] By discerning variations in digital phenotypes that are predictive of differing symptom levels and severities, we extended existing studies that have demonstrated the importance of behavioral, physiological, and social patterns as indicators of symptom severity.[55–58] A crucial consideration in these predictions concerns the temporal dynamics between feature extraction for digital phenotypes and symptom measurement. Our model design assumed no time lag between symptoms and phenotypes and posited that digital phenotypes could reflect symptom levels within a 2-week time window. However, digital phenotype data were to encompass the 2-week period prior to symptom assessment of 2-week intervals. A 50% overlap

**Table 8.** Comparison of key features' integrated gradient values across severity levels for the BFNE scale (the italicized ig values are top three features indicative of a specific severity level).

| Feature name | Normal | Abnormal |
|---|---|---|
| | | |
| Circadian movement | *0.98* | *−0.71* |
| Lux during night | *−0.43* | *0.64* |
| Mean sleep duration | *0.48* | −0.45 |
| Mean usage duration of communication app (SNS) | −0.30 | *0.53* |

BFNE: Brief Fear of Negative Evaluation Scale.

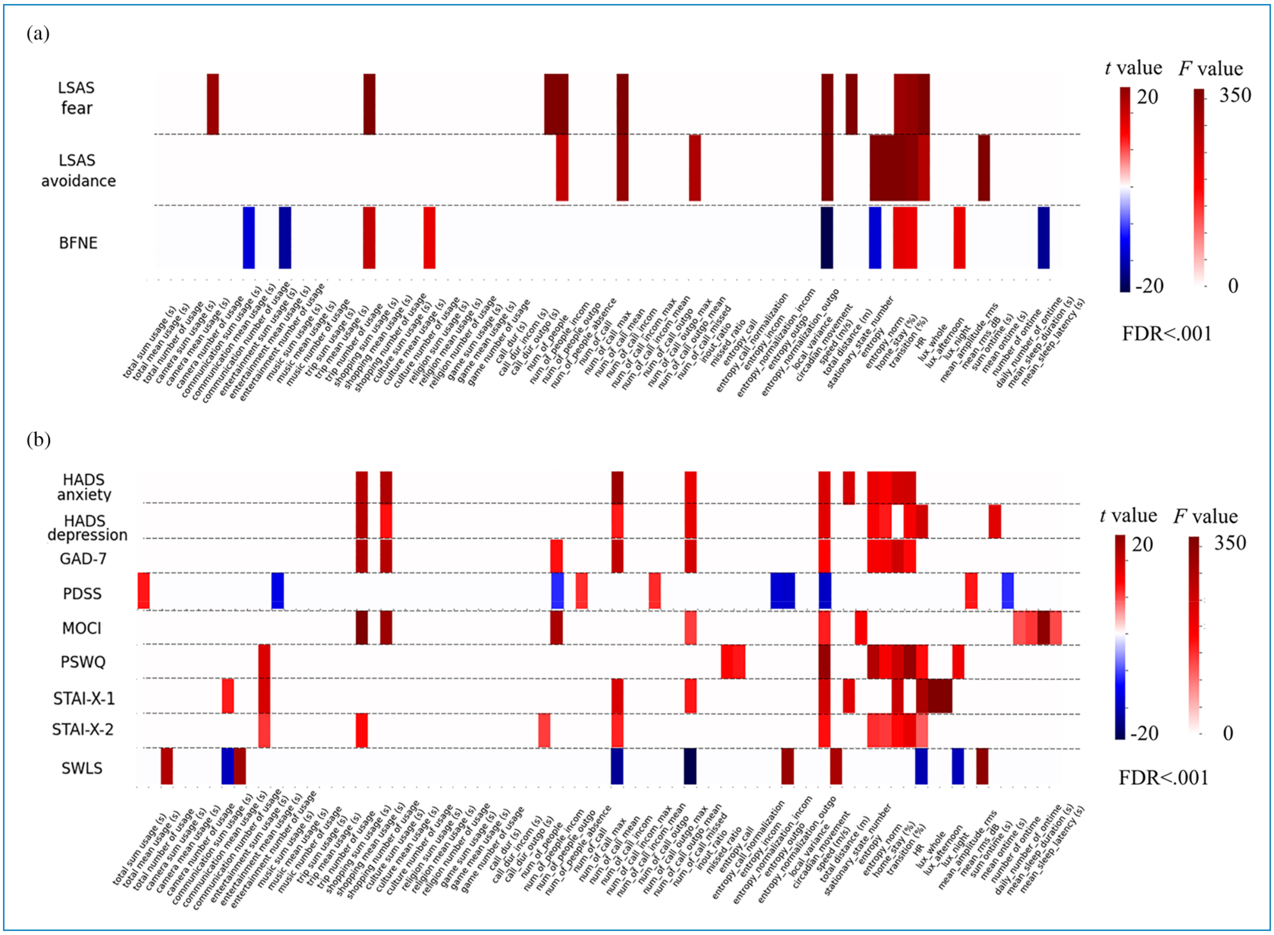

**Figure 3.** Differences in integrated gradients of digital phenotypes across symptom severity levels. (a) Differences in the effects from integrated gradients of digital phenotypes across social anxiety symptom severity, plotted with FDR corrected multiple comparisons ($P < .001$). (b) Comparative plots of integrated gradients across secondary psychiatric symptom severity. Multiple comparisons were corrected using FDR ($P < .001$). FDR: false discovery rate.

of segments in the extraction of digital phenotype features resulted in a discrepancy in frequency between assessment and symptoms. To bridge this gap, we applied an interpolation method to symptom scores, but there was still a difference in frequency. Additionally, different symptoms and behaviors have different timescales. For example, heart rate and limited mobility are expected to have different temporal correlations with anxiety. Therefore, it is inevitable that fixed time lags and windows across all phenotypes have a potentially negative impact on model performance. Even taking this inevitability into account, there is evidence that a 2-week period is within the acceptable range for the relationship between digital phenotypes and psychiatric symptoms. As an example, Stamatis et al. in a recent study found an optimal temporal relationship between circadian movement and depressive symptoms, which was specifically within a 2-week time window and with no time lag.[59] Although the temporal dimensions were not incorporated into the model due to the cross-sectional design of our study, designing a model that accounts for these dimensions by predicting future symptoms from features identified at a specific time[60,61] can be a useful approach to demonstrating the model's actual predictive ability.

Autoencoders extract representation features, which successfully create interpretability by capturing the essence of the original data, and low-dimensional representation techniques using autoencoders make it possible to simplify the understanding of complex datasets.[62,63] The use of these techniques allowed us to observe specific patterns in digital phenotypes that related to symptom severity. For example, individuals with higher severity levels in overall trends tended to spend more time without contact with others compared to lower severity levels. In addition, circadian movement, calculated as movement regularity over a 2-week period, was a key predictor across multiple severity levels and various scales, ranging from anxiety and depression to life satisfaction. This finding is consistent with several previous reports that have supported the role of circadian movement as a significant indicator of mental health states. For instance, Velten et al.[64] found a relationship between the regularity of daily activities such as sleep and eating and mental states including life satisfaction, depression, and anxiety. Saeb et al.[65] discovered the relationship between the regularity of GPS movement patterns and depression. Difrancesco et al.[66] showed that circadian rhythm and physical activity based on sleep time and activity levels were correlated not only with depression but also anxiety.

Focusing on the emotional dimensions of social anxiety, it is evident that limited mobility patterns play a meaningful role in predicting high levels of social anxiety. An increase in time spent at home and limited geographic range of activities could serve as behavioral indicators of fear related to social exposure. Notably, normalized location entropy stood out as a prominent factor solely in the LSAS-fear model when predicting moderate class, suggesting its potential utility in reflecting behavioral characteristics associated with social anxiety. In contrast, active interaction patterns through phone calls and geographical diversity in location visits were crucial in predicting normal levels of social anxiety.

Regarding avoidance behaviors in social interactions, particularly for the moderate severity class, substantial contributory factors included elevated heart rate and increased use of entertainment apps, as revealed by the analysis. Supporting a previous study that highlighted a correlation between heart rate and levels of social anxiety in real-world settings,[67] our findings suggest that heart rate could serve as a more accurate marker for assessing high levels of social anxiety, particularly its avoidance behaviors. Similarly, the increased use of entertainment apps could signify an avoidance strategy to cope with elevated anxiety levels. Prior research indicates that individuals with high social anxiety tend to use the internet more to avoid social interactions and cope with loneliness.[68] Intriguingly, avoidance behaviors associated with social anxiety can manifest not only through the avoidance of physical social situations but also through increased use of apps,[69,70] such as those for entertainment, which are not primarily motivated by the purpose of social interaction but rather facilitate content consumption in a unidirectional manner.[71] Furthermore, for the mild severity class, characterized by partial avoidance tendencies, key predictors included longer home stays, reduced social exposure, and frequent smartphone usage—collectively serving as markers of social isolation. Conversely, for individuals with normal severity levels, frequent exposure to diverse social settings and increased durations outside the home predicted more adaptive behavioral patterns in response to social situations.

Transitioning to the cognitive aspects of social anxiety, specifically the fear of negative evaluation in social interactions, circadian movement and nighttime ambient brightness emerged as noteworthy indicators for both normal and abnormal severity classes. This observation aligns with the study of Lyall et al.,[72] who demonstrated that disruptions in circadian rhythms could lead to various adverse mental health outcomes. Similarly, findings from our study reveal that nighttime ambient brightness may be associated with fear of negative evaluation in social situations. Intriguingly, the duration of communication app usage emerged as a pivotal predictor specifically when the severity of symptoms was pronounced. Previous studies have shown that communication skills are important in predicting mental illnesses.[73,74] Our observation on the use of communication apps may be related to the cognitive model proposed by Clark and Wells.[75] According to this model, cognitive characteristics of social anxiety encompass excessive self-monitoring and adherence oneself to perfectionist standards during social performances, such as the need to be positively evaluated by everyone in social situations. This model was empirically supported not only in face-to-face interaction situations related to social anxiety, but also in online interaction settings.[70] Importantly, these cognitive factors, measured by the BFNE, are recognized to prolong task performance time.[76] Therefore, this finding suggests that the extended duration of communication apps, especially those involving interaction with others, may be associated with high levels of cognitive symptoms related to social anxiety. Furthermore, the duration of communication app usage was identified as a key feature in predicting cases categorized as moderate obsession/compulsion and dissatisfaction (Tables S10 and S11 in Supplemental Appendix). This aligns with a previous study that identified a negative relationship between SNS usage duration and life satisfaction.[77]

In the broader context of this study, the findings hold significance in identifying neurotic traits and digital phenotypes that are specific to various levels of symptom severity in SAD. In addition, this suggests their potential utility in diagnosis and management. The prevailing diagnostic frameworks for psychiatric disorders largely rely on clinician-conducted interviews and self-report scales. Complementing this, the incorporation of behavioral and physiological metrics obtained from digital phenotypes could offer an enriched contextual basis for diagnostic assessments in SAD. From a therapeutic perspective, our results may give insights for targeted interventions in key digital phenotypes depending on the severity of symptoms. Intervention can be customized by focusing on key variables that highly contribute to predicting symptom severity, including strategies for behavior modification. For instance, traditional exposure therapy for avoidance behaviors that contribute to maintaining[78,79] and exacerbating[68] social anxiety symptoms has focused on direct interpersonal situations, but incorporating behavior modification strategies for the use of entertainment apps as a coping mechanism may offer a valuable treatment approach for social anxiety.

### *Limitations*

Despite the contributions of this study to the field of digital phenotyping in psychiatric disorders, several limitations warrant mention. First, it must be acknowledged that the final analyzed data of 58 participants was a rather small sample size in digital phenotype analysis. More statistical power could have been obtained if the analysis had been conducted by collecting data from a larger number of participants based on a method recently presented in the literature on determination of appropriate sample size in smartphone-based digital phenotyping studies such as ours.[80] In addition, the combining of data from both SAD and HC groups, with unequal sample sizes across severity classes, may make findings not fully representative of characteristics of either group. While the inclusion of a control group enhances the generalizability of the study, it also underscores the necessity for more targeted future research. Second, although the categorical classifications of symptom severity were implemented to perform the analysis reflecting the current clinical frameworks used to recommend specific therapies, transforming continuous symptom scores into categorical level might oversimplify the subtle variations in symptom severity. Third, since our research was based on technologies that utilize artificial intelligence, inherent biases such as training data bias, algorithmic bias, and cognitive bias might affect the results. The representation learning approach only assesses indirect data contributions, necessitating future research on direct influences from latent features. Additionally, although the predictive models are tested through sensitivity analysis, no such analysis was conducted for the feature importance scores. Consequently, the reported importance of these features is primarily exploratory and necessitates additional validation. Given that there are separate datasets to train and test the model, a subset of the available datasets may be affected by bias, and thus we adopted cross-validation to mitigate bias affecting the results, but the existence of overfitting still remains a problem. Finally, the scope of comparative analysis is restricted to three variables for interpretability, potentially omitting less critical yet still meaningful predictors.

## Conclusion

In this study, we have demonstrated the efficacy of digital phenotypes for classifying symptom severities in SAD by leveraging feature representation learning. We uncovered distinct digital phenotypes associated with the cognitive, emotional, and behavioral dimensions of the disorder. These findings, which emerged from an analysis of digital phenotypes, lay the groundwork for a more comprehensive understanding of SAD and open up potential applications for managing symptoms through digital behavioral modifications. Building upon these insights, further randomized controlled trials are warranted to establish the causal relationships between key digital phenotypes and symptomatology of SAD, which holds the potential to significantly improve intervention strategies.

**Acknowledgements**: We thank Bohyun Park and Soomin Kim for their partial participation in data collection.

**Contributorship:** HC and YC drafted this article. YC, EK, and SL supervised the data collection. CM and KK provided supervision for the entire project and helped draft the manuscript. HC and CM provided support for data analysis. JJK was principal investigator for the project. All authors read and approved the final version of the manuscript.

**Declaration of conflicting interests:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethical approval:** Ethics approval was obtained from the institutional review board of Yonsei University Gangnam Severance hospital (3-2021-0267).

**Funding:** The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (grant 2021M3E5D9025019).

**Guarantor:** JJK.

**ORCID iDs:** Yesol Cho https://orcid.org/0000-0002-5121-2797
Jae-Jin Kim https://orcid.org/0000-0002-1395-4562

**Supplemental material:** Supplemental material for this article is available online.

## References

1. Association AP. *Diagnostic and statistical manual of mental disorders*. Washington, DC: DSM-5: American Psychiatric Association, 2013.
2. Stein MB and Stein DJ. Social anxiety disorder. *Lancet* 2008; 371: 1115–1125.
3. Hirsch C, Meynen T and Clark D. Negative self-imagery in social anxiety contaminates social interactions. *Memory* 2004; 12: 496–506.
4. Ruscio AM, Brown TA, Chiu WT, et al. Social fears and social phobia in the USA: results from the National Comorbidity Survey Replication. *Psychol Med* 2008; 38: 15–28.
5. Lecrubier Y, Wittchen H-U, Faravelli C, et al. A European perspective on social anxiety disorder. *Eur Psychiatry* 2000; 15: 5–16.
6. Belzer K and Schneier FR. Comorbidity of anxiety and depressive disorders: issues in conceptualization, assessment, and treatment. *J Psychiatr Pract* 2004; 10: 296–306.
7. Jain SH, Powers BW, Hawkins JB, et al. The digital phenotype. *Nat Biotechnol* 2015; 33: 462–463.
8. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA* 2017; 318: 1215–1216.
9. Boukhechba M, Daros AR, Fua K, et al. Demonicsalmon: monitoring mental health and social interactions of college students using smartphones. *Smart Health* 2018; 9: 192–203.
10. Moshe I, Terhorst Y, Opoku Asare K, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Front Psychiatry* 2021; 12: 625247.
11. Melcher J, Lavoie J, Hays R, et al. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. *J Am Coll Health* 2023; 71: 736–748.
12. Dlima SD, Shevade S, Menezes SR, et al. Digital phenotyping in health using machine learning approaches: scoping review. *JMIR Bioinform Biotech* 2022; 3: e39618.
13. Bai R, Xiao L, Guo Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. *JMIR mHealth uHealth* 2021; 9: e24365.
14. Carreiro S, Chintha KK, Shrestha S, et al. Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: a mixed methods pilot study. *Drug Alcohol Depend* 2020; 209: 107929.
15. Jacobson NC and Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. *Behav Res Ther* 2022; 149: 104013.
16. Weller BE, Bowen NK and Faubert SJ. Latent class analysis: a guide to best practice. *J Black Psychol* 2020; 46: 287–311.
17. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006; 313: 504–507.
18. Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res* 2010; 11: 3371–3408.
19. Heinsfeld AS, Franco AR, Craddock RC, et al. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin* 2018; 17: 16–23.
20. Sardari S, Nakisa B, Rastgoo MN, et al. Audio based depression detection using Convolutional Autoencoder. *Expert Syst Appl* 2022; 189: 116076.
21. Jacobson NC, Lekkas D, Huang R, et al. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. *J Affect Disord* 2021; 282: 104–111.
22. Boukhechba M, Chow P, Fua K, et al. Predicting social anxiety from global positioning system traces of college students: feasibility study. *JMIR Ment Health* 2018; 5: e10101.
23. Jacobson NC, Summers B and Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. *J Med Internet Res* 2020; 22: e16875.
24. Hedman E, Ström P, Stünkel A, et al. Shame and guilt in social anxiety disorder: effects of cognitive behavior therapy and association with social anxiety and depressive symptoms. *PLoS ONE* 2013; 8: e61713.
25. Liebowitz MR. Social phobia. Modern problems of pharmacopsychiatry. *Mod Probl Pharmacopsychiatry* 1987; 22: 141–173.
26. Leary MR. A brief version of the Fear of Negative Evaluation Scale. *Pers Soc Psychol Bull* 1983; 9: 371–375.
27. Zigmond AS and Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983; 67: 361–370.
28. Spitzer RL, Kroenke K, Williams JB, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006; 166: 1092–1097.
29. Shear MK, Brown TA, Barlow DH, et al. Multicenter collaborative panic disorder severity scale. *Am J Psychiatry* 1997; 154: 1571–1575.
30. Hodgson RJ and Rachman S. Obsessional-compulsive complaints. *Behav Res Ther* 1977; 15: 389–395.
31. Meyer TJ, Miller ML, Metzger RL, et al. Development and validation of the Penn State Worry Questionnaire. *Behav Res Ther* 1990; 28: 487–495.
32. Spielberger CD. *State-trait anxiety inventory for adults*. APA PsycTests, 1983. doi: 10.1037/t06496-000.
33. Diener E, Emmons RA, Larsen RJ, et al. The satisfaction with life scale. *J Pers Assess* 1985; 49: 71–75.

34. Peterson LE. K-nearest neighbor. *Scholarpedia* 2009; 4: 1883.
35. Beutel ME, Jünger C, Klein EM, et al. Noise annoyance is associated with depression and anxiety in the general population-the contribution of aircraft noise. *PLoS ONE* 2016; 11: e0155357.
36. Bossini L, Martinucci M, Paolini K, et al. Panic-agoraphobic spectrum and light sensitivity in a general population sample in Italy. *Can J Psychiatry* 2005; 50: 39–45.
37. van Hees VT, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. *Sci Rep* 2018; 8: 12975.
38. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929–1958.
39. Loshchilov I and Hutter F. Decoupled Weight Decay Regularization. 7th International Conference on Learning Representations, New Orleans, 6-9 May 2019. https://dblp.org/rec/conf/iclr/LoshchilovH19.html
40. Reddy YJ, Sudhir PM, Manjula M, et al. Clinical practice guidelines for cognitive-behavioral therapies in anxiety disorders and obsessive-compulsive and related disorders. *Indian J Psychiatry* 2020; 62: S230.
41. Fineberg NA, Hollander E, Pallanti S, et al. Clinical advances in obsessive-compulsive disorder: a position statement by the International College of Obsessive-Compulsive Spectrum Disorders. *Int Clin Psychopharmacol* 2020; 35: 173–193.
42. Sundararajan M, Taly A, Yan Q, et al. Axiomatic attribution for deep networks. International Conference on Machine Learning 2017: PMLR.
43. Lyu W, Dong X, Wong R, et al. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. AMIA Annual Symposium Proceedings 2022; 2022: 719–728. PMID: 37128451
44. Dai R, Kannampallil T, Kim S, et al. Detecting mental disorders with wearables: A large cohort study. Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation 2023: 39–51. doi: 10.1145/3576842.3582389.
45. Shrestha I and Srinivasan P. Comparing deep learning and conventional machine learning models for predicting mental illness from history of present illness notations. AMIA Annual Symposium Proceedings 2021; 2021: 1109–1118. PMID:35308915
46. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc, B: Stat Methodol* 1995; 57: 289–300.
47. Liu K and Tao D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Comp Hum Behav* 2022; 127: 107026.
48. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* 2021; 46: 45–54.
49. Castelvecchi D. Can we open the black box of AI? *Nature News* 2016; 538: 20–23.
50. Montavon G, Samek W and Müller K-R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process* 2018; 73: 1–15.
51. Huff DT, Weisman AJ and Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys Med Biol* 2021; 66: 04TR1.
52. Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 2015; 10: e0130140.
53. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2016: 2921–2929. doi: 10.1109/CVPR.2016.319
54. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision* 2017: 618–626. doi: 10.1109/ICCV.2017.74
55. Lima RA, de Barros MVG, Dos Santos MAM, et al. The synergic relationship between social anxiety, depressive symptoms, poor sleep quality and body fatness in adolescents. *J Affect Disord* 2020; 260: 200–205.
56. Walker WH, Walton JC, DeVries AC, et al. Circadian rhythm disruption and mental health. *Transl Psychiatry* 2020; 10: 28.
57. Sparrevohn RM and Rapee RM. Self-disclosure, emotional expression and intimacy within romantic relationships of people with social phobia. *Behav Res Ther* 2009; 47: 1074–1078.
58. Wenzel A, Graff-Dolezal J, Macho M, et al. Communication and social skills in socially anxious and nonanxious individuals in the context of romantic relationships. *Behav Res Ther* 2005; 43: 505–519.
59. Stamatis CA, Meyerhoff J, Meng Y, et al. Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. *NPJ Ment Health Res* 2024; 3: 1.
60. Fried EI, Proppert RK and Rieble CL. Building an early warning system for depression: rationale, objectives, and methods of the WARN-D study. *Clin Psychol Eur* 2023; 5: e10075.
61. Haslbeck JM, Bringmann LF and Waldorp LJ. A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behav Res* 2021; 56: 120–149.
62. Michelucci U. An introduction to autoencoders. 2022; arXiv:2201.03898.
63. Chen S and Guo W. Auto-encoders in deep learning—a review with new perspectives. *Mathematics* 2023; 11: 1777.
64. Velten J, Lavallee KL, Scholten S, et al. Lifestyle choices and mental health: a representative population survey. *BMC Psychol* 2014; 2: 1–11.
65. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res* 2015; 17: e4273.
66. Difrancesco S, Lamers F, Riese H, et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: a 2-week ambulatory assessment study. *Depress Anxiety* 2019; 36: 975–986.
67. Rösler L, Göhring S, Strunz M, et al. Social anxiety is associated with heart rate but not gaze behavior in a real social interaction. *J Behav Ther Exp Psychiatry* 2021; 70: 101600.
68. Hernández C, Ferrada M, Ciarrochi J, et al. The cycle of solitude and avoidance: a daily life evaluation of the relationship between internet addiction and symptoms of social anxiety. *Front Psychol* 2024; 15: 1337834.

69. Prizant-Passal S, Shechner T and Aderka IM. Social anxiety and internet use – a meta-analysis: what do we know? What are we missing? *Comput Human Behav* 2016; 62: 221–229.
70. Hutchins N, Allen A, Curran M, et al. Social anxiety and online social interaction. *Aust Psychol* 2021; 56: 142–153.
71. Elhai JD, Hall BJ, Levine JC, et al. Types of smartphone usage and relations with problematic smartphone behaviors: the role of content consumption vs. social smartphone use. *Cyberpsychology (Brno)* 2017; 11: 3.
72. Lyall LM, Wyse CA, Graham N, et al. Association of disrupted circadian rhythmicity with mood disorders, subjective well-being, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. *Lancet Psychiatry* 2018; 5: 507–514.
73. Parola A, Brasso C, Morese R, et al. Understanding communicative intentions in schizophrenia using an error analysis approach. *NPJ Schizophr* 2021; 7: 12.
74. Dall M, Fellinger J and Holzinger D. The link between social communication and mental health from childhood to young adulthood: a systematic review. *Front Psychiatry* 2022; 13: 944815.
75. Clark DM and Wells A. A cognitive model of social phobia. In: Heimberg G, Liebowitz MRMR, Hope D and Scheier F (eds) *Social phobia: diagnosis, assessment, and treatment*. New York: The Guilford Press, 1995, pp.69–93.
76. Kelly WE. Anxiety and the prediction of task duration: a preliminary analysis. *J Psychol* 2002; 136: 53–58.
77. Stieger S. Facebook usage and life satisfaction. *Front Psychol* 2019; 10: 2711.
78. Martín C S, Jacobs B and Vervliet B. Further characterization of relief dynamics in the conditioning and generalization of avoidance: effects of distress tolerance and intolerance of uncertainty. *Behav Res Ther* 2020; 124: 103526.
79. Hofmann SG. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. *Cogn Behav Ther* 2007; 36: 193–209.
80. Barnett I, Torous J, Reeder HT, et al. Determining sample size and length of follow-up for smartphone-based digital phenotyping studies. *J Am Med Inform Assoc* 2020; 27: 1844–1849.