Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health

Taewan Kim
KAIST
Daejeon, Republic of Korea
taewan@kaist.ac.kr

Haesoo Kim
KAIST
Daejeon, Republic of Korea
haesoo1108@gmail.com

Ha Yeon Lee
Seoul National University
Seoul, Republic of Korea
raon0172@snu.ac.kr

Hwarang Goh
Inha University
Incheon, Republic of Korea
hrgoh@nsl.inha.ac.kr

Shakhboz Abdigapporov
Inha University
Incheon, Republic of Korea
mike@nsl.inha.ac.kr

Mingon Jeong
Hanyang University
Seoul, Republic of Korea
mingon21@ajou.ac.kr

Hyunsung Cho
Carnegie Mellon University
Pittsburgh, United States
hyunsung@cs.cmu.edu

Kyungsik Han
Hanyang University
Seoul, Republic of Korea
kyungsikhan@hanyang.ac.kr

Youngtae Noh
KENTECH
Naju, Republic of Korea
ytnoh@kentech.ac.kr

Sung-Ju Lee
KAIST
Daejeon, Republic of Korea
profsj@kaist.ac.kr

Hwajung Hong
KAIST
Daejeon, Republic of Korea
hwajung@kaist.ac.kr

ABSTRACT

Reflecting on stress-related data is critical in addressing one's mental health. Personal Informatics (PI) systems augmented by algorithms and sensors have become popular ways to help users collect and reflect on data about stress. While prediction algorithms in the PI systems are mainly for diagnostic purposes, few studies examine how the explainability of algorithmic prediction can support user-driven self-insight. To this end, we developed MindScope, an algorithm-assisted stress management system that determines user stress levels and explains how the stress level was computed based on the user's everyday activities captured by a smartphone. In a 25-day field study conducted with 36 college students, the prediction and explanation supported self-reflection, a process to re-establish preconceptions about stress by identifying stress patterns and recalling past stress levels and patterns that led to coping planning. We discuss the implications of exploiting prediction algorithms that facilitate user-driven retrospection in PI systems.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

personal informatics, mental wellbeing, stres management, algorithm experience, explainability

ACM Reference Format:
Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3491102.3517701

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517701

1 INTRODUCTION

Many college students report experiencing varying levels of stress throughout their college lives. Furthermore, they are especially vulnerable because they are in the life stage which mental health-related problems first appear [33]. Stress can interfere with students' productivity and adversely affect both physical and mental health [48, 53]. Exposure to numerous stressors, and the resulting reactions to stress can lead to decreased well-being. However, learning about one's stress reactions can be a challenge. Discovering how to change stress-related behavior requires empowering views of reality with insights about the current situation [71]. Thus, self-reflection—a process of examining and paying attention to one's cognitive, emotional, and behavioral status—plays a key role in managing one's stress reactions and dealing with situations that cause stress, as it can help to calibrate the individual's actions in the future based on data-driven self-insights [79].

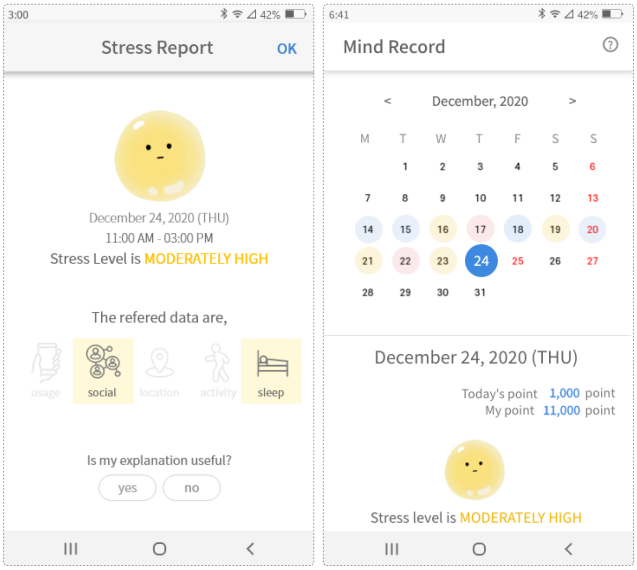

Figure 1: The MindScope Application. Stress Prediction Reports (left) and Stress pattern in calendar Overview (right)

In the human–computer interaction (HCI) field, personal informatics (PI) systems have proven their effectiveness in managing diverse health issues by helping users gain insights from self-reflection on past behaviors and suggesting future actions to deal with identified problems [23]. Earlier research reported that users of PI systems were able to gain insight about personal health by observing peaks, trends, and correlations of the health-related data [16]. For example, the PI system can facilitate introspection of health-related problems by visualizing associations between data obtained from sensors that reflect users' daily lives and self-reported data [28, 40]. The PI system, which aims to collect and reflect personal information to help users achieve well-being and positive behavior changes, is widely used to manage mental health domains (e.g., stress [1], mood [13, 54], and emotion [29]). Moreover, the PI system continues to evolve with the development of intelligent technology. Recent advances in passive sensing technology using smartphones and wearable devices have brought significant progress to the field. Enhanced PI systems enable behavioral data collection at a more detailed level by combining multiple sensors while reducing the user's data input burden [18, 49, 68]. Moreover, machine learning technologies can be incorporated to support users' health behavior changes by extracting insights and patterns from personal data and suggesting proper actions in the future [25].

However, previous studies on the existing intelligent systems augmented by prediction algorithms have mainly focused on accurately detecting and reporting a user's state instead of assisting a user's self-reflection, which the user can subsequently employ to change behaviors [59]. Moreover, the current intelligent PI system has limitations that make it difficult for users to interact with the algorithm or understand its functioning or decisions. In other words, users are mainly provided with algorithmic output from the system without a full understanding of its decisions, barring users from gaining further insights.

Recent human–AI interaction research has emphasized that interaction opportunities such as Explainable AI (XAI) [20, 44] and user feedback [3] positively influence the user's experience, as well as improve the model's performance [61]. In addition, it has been reported that explainability can help users gain insights from data and facilitate learning [41, 63] by allowing users to understand the functioning and decision of an algorithm [20].

Nevertheless, incorporating algorithm and explainability into the PI system requires careful consideration because of the possibility that they might not work as intended (e.g., override personal interpretation and trust algorithmic output [29], decreases positive impression of the system [46], distracts users' attention, and violates expectation [76]) We argue that the algorithm-incorporated PI system should assist users' self-reflection by allowing them to understand the algorithmic decisions through the appropriate level of explainability. Thus, we suggest a research prototype that allows us to examine the design spaces of explainability in an algorithm-incorporated PI system for stress management.

To this end, we developed *MindScope*, an algorithm-assisted stress management system that determines user stress levels and explains how the stress level was computed based on the user's smartphone use and activity, to investigate the use and value of a prediction algorithm for user-driven stress self-reflection. The data collected through smartphones include GPS information, ambient noise, app usage, screen status, and accelerometer readings. This data is then organized into five categories: social, location, activity, sleep, and phone usages, inferring a stress level. The MindScope system includes (1) a 10-day data collection phase that requires a user's input about his/her current stress levels four times a day to establish a personalized stress ML model, and (2) a 15-day reflection phase that provides a report about the user's current stress level as determined by the ML model and an explanation of several predictors of stress levels four times a day. In examining how the explainability of the algorithm affects the user's understanding of stress as well as perception toward the algorithm, we designed and evaluated the following three different visualization of providing prediction information to understand how the level of explainability affects the self-reflection experience of the PI system. In explanation *Type 1*, only predicted stress level was provided. In explanation *Type 2*, we supplemented the explanation by highlighting data categories (phone usage, social activity, movement, physical action, and sleep) that had significant weight in calculating the prediction. In explanation *Type 3*, we provided a more detailed and granularized explanation, such as major deviations from the norm in each datapoint (e.g., using an app more/less than usual) as presented in Figure 4.

This paper reports on how prediction algorithms support college students' stress self-reflection based on the 25-day field deployment with 36 college students. In particular, we investigated how explanation of the results of prediction affect users' perception of the system and self-reflection experience by providing different levels of prediction explainability. Our results indicate that the algorithm-assisted approach helped users understand detailed stress patterns and plan stress intervention. Stress prediction with the explanation proved useful by reducing ambiguity in the process of recalling

past stress levels and related events. While most users rated *Type 3*, the most detailed explanation type, highly for concrete understandings of their stressors and patterns, they often focused on specific activity details that were inconsistent with their memories, which lowered the perceived accuracy of the system. The explanations provided at the category level for stress prediction provided an opportunity to understand stressors in a user-led way.

Our contributions are as follows: We developed MindScope, a mobile app that predicts user stress levels with personalized stress prediction algorithms based on smartphone data, allowing users to understand and cope with stress. MindScope also generates prediction and explanation in three different ways using varying levels of explanation to explore the impact of explainability in the PI domain. We validate the effectiveness of MindScope in stress reduction through empirical findings on how participants use the algorithmic stress prediction and explanation for their stress management. More importantly, we report on the ways the explainability of stress prediction affected users' self-reflection and algorithmic perception. We discuss the design consideration when creating the visualization of predictive information in the PI system to promote self-reflection. Furthermore, we summarize the design suggestions that a designer should consider to improve the user experience of the PI System using the prediction algorithm.

2 RELATED WORK

2.1 Personal Informatics for Mental Health

Technologies that collect and reflect personal information for well-being and behavioral changes continue to proliferate [23]. Li et al. [43] defined these as PI systems that "help people collect personally relevant information for the purposeof self-reflection and gaining self-knowledge.", and suggested five stages of PI systems: 1) preparing to collect data, 2) collecting data, 3) integrating, 4) reflecting, and 5) acting. Later, Epstein et al. [24] extended this model to better reflect the diverse motivations and use of the PI system in terms of "Lived Informatics". In this model, collecting, integrating, and reflecting are not separated stages, but are considered as practices that can occur simultaneously in the process of "Tracking and Acting". The scope of PI has broadened, covering numerous domains such as physical activity, chronic condition [25], sleep [15], and productivity [38]. Among these domains, mental health is a significant social issue facing many people in contemporary society [34]. As a result, there have been growing studies of digital technology to understand and address mental health issues. Early HCI research for mental health focused on replicating traditional therapeutic strategies (e.g., workshop sessions provided in electronic formats [4]) to support limited capacity and availability of treatment services [78]. Recent research has been expanded to the field of mental well-being promotion, and various works have been conducted on the system for a strong sense of self [73], mindfulness [88], and social well-being [81]. From the point of view of PI research, it is worth noting that: technology has offered people the possibility to facilitate understanding, observation, and reflection about themselves based on their recollection of the past. For instance, Bardram et al. presented a personal monitoring system MONARCA, which enabled monitoring with feedback through data visualization and triggers to support the treatment of bipolar [5]. As the use of mobile phones and wearable technologies provides potential of continuous tracking and personalized intervention,the development and assessment of PI systems in mental health has accelerated [6, 12]. However, accurate tracking of mental states has proved much more challenging than physical health studies because of mood instability and variations between individuals [54, 75]. By virtue of technological advances in passive sensing with smartphones and wearable devices, many efforts have been made to track mental aspects through an algorithmic approach. Morshed et al. useda user's self-reporting system, Ecologically Momentary Assessments (EMA), with passive sensing to predict mood instabilities [54]. Sarker et al. designed a system that detect stress episode using physiological, GPS, and activity data [66]. Many PI research in the mental health domain aim to detect and report on the user's status, thereby putting effort into effectively collecting tremendous amounts of emotional data to improve the accuracy of predictions. However, there is a potential risk that users override personal interpretation and instead trust algorithmic output [29, 72, 84]. Hollis et al. revealed that some participants defer to system feedback trusting affect detection algorithms to be more accurate than their own intuitions [29]. In addition, publications have discussed the importance of reflection stages and subsequent links to practical action. For example, research has identified that the performance of stress interventions after indication and prediction of stressors can lead to significant stress reductions [42]. Thus, we need to examine the design of the intelligent PI system that promotes users' self-understanding and behavior change rather than unconditionally receiving algorithm output.

2.2 Intelligent Computing in Personal Informatics

Intelligent computing aims to bring the elements of intelligence, reasoning, analysis, and information gathering to systems [9]. Recently, in the PI field, a system that helps people understand personal data by utilizing intelligent computing has been proposed for physical [25, 46] and mental well-being [28, 59]. Furthermore, due to the recent advances in artificial intelligence (AI) technologies such as machine learning, the system goes beyond assisting in automating repetitive tasks to even collaborating with users in complex tasks such as collaborative drawing [58] and writing [32]. Ohlin and Olsson proposed the concept of cooperative action orchestration in which intelligent computing systems and humans repeatedly influence each other [59]. This approach emphasizes that users and computers do not unconditionally accept each other's decisions but rather work together and influence each other to attain the goal of personal reflection. For example, EmotiCal [73], a PI system that accurately forecasts individual mood, reported that user participation and evaluation was able to significantly improve the system's predictive power. In addition, users perceived the benefit of intelligent computing in the PI system, which supplements users' subjective and intuition-based judgment by utilizing long-term objective data [50, 72]. For example, EmVive [29], which predicts personal stress based on EDA sensor data, was evaluated as a useful system to gain self-insights from users who considered their self-awareness level insufficient. Previous studies show that intelligent computing can make the user's self-reflection process more concrete and

persuasive by using objective and quantified data about users with the algorithms. Further, computing systems and people can build cooperative relationships that can positively influence each other and enable better personal reflection [36, 59]. We were inspired by these results that reveal the benefits of intelligent computing for mental health.

2.3 Explainability of Intelligent Computing and PI Systems

Explainable AI (XAI) refers to an intelligent system that allows users to understand a system's functions and decisions. The goal of explainability is to enhance a system's user experience by increasing users' trust in the algorithm's decisions [3, 20]. Furthermore, it has been reported that explainability can help users gain insights from data and facilitate learning [41, 44, 63]. However, the effects of explainability are not always positive [44]. Studies have found that transparency lowers user trust by inducing users to question the system and that complex descriptions increase the user's excessive cognitive load, ultimately negatively affecting the perception of algorithms [46, 76]. When the system behaves appropriately but presents a low certainty, intelligibility notably reduces the user's impression of the system [46]. In addition, more transparency may distract a user's attention or violate a user's expectations by emphasizing the relatively unimportant or specific elements of explanation [76]. In particular, unintended consequences can be a critical problem in systems dealing with health issues (e.g., affecting users' emotions negatively [67] or unintentionally encouraging negative behavior [22]). Therefore, possible risks should be carefully examined early in the design process by rethinking potential assumptions and unintended consequences [39]. In the PI domain, Woźniak et al. [85] revealed that providing both the fitness goal generated by the algorithm and explaining how this goal was calculated improved users' trust in the suggested goal. This increased trust led to enhanced goal commitment. However, in the study of the E-meter [75], which rates the emotion of written text that contains the user's emotional experience, excessive explanations negatively affected the user's view of the system's reliability and his/her satisfaction. The above studies collectively suggest that the effects of explainability are complex and depend on the setting and purpose of the interactions. Furthermore, there is a high possibility that explainability can work differently than the designer intended. This is because, so far, the development and application of XAI have been based on a technology-oriented solutionism approach rather than on the situated needs of the intended user of the system [20, 27].

In this study, we first examine the user experience and perception of intelligent PI systems that predict a user's stress levels and investigate how this algorithmic output affects the user's self-reflection. In particular, we focus on investigating whether explaining how stress predictions are computed by the algorithm can affect users in their self-reflection and stress management, which have not yet been sufficiently addressed in an intelligent PI system for mental health. To this end, we developed a stress management PI system—MindScope—that can determine a user's stress level with a personalized prediction algorithm. The system was also designed to provide stress prediction in three ways with three different levels of explanation. Through a field study, we wanted to understand how the level of explainability affects users' understanding of stressors and patterns, intervention planning, and overall perception of the PI system.

3 SYSTEM DESIGN

We developed MindScope, an app that predicts users' stress levels based on smartphone sensors and usage data. The use of the MindScope app is designed to operate in two phases: a 10-day modeling phase and a 15-day prediction phase (see Figure 2). While the main functionality regarding stress prediction and algorithmic interaction is centered on the prediction phase, we wanted to ensure that a sufficient volume of data was collected to produce more reliable predictions. Thus, we emphasized the importance of the modeling phase and the creation of a personalized stress prediction model for each individual participant based on the correlation between the two types of data [82].

3.1 Iterative Design Process

We used an iterative design process that examined multiple prototype versions and conducted a pilot study with 30 college students to identify core usability issues and interesting usage patterns worthy of further investigation. The initial concept was designed by referencing earlier works on the model of the PI system [23, 43]. Our system supports the data collection by utilizing passive mobile sensor data collection, along with the user's self-reported stress levels. For the data integration, and reflection, the prediction algorithm was used to provide users with data-driven insight concerning understanding and managing stress by delivering stress level prediction and explanation. We selected the data category based on earlier works that revealed the relationship between passive sensing data from mobile phones, including social activity, movement, physical action, sleep, and phone usage (app usage) with mental health outcomes such as depression, stress, and loneliness [8, 51, 65, 82, 83]. Additionally, the system was designed to supports users' stress relief based on microtask suggestions, which are known to be effective in stress management [42]. Based on this initial concept, we developed the system through an iterative design process. In this section, we focused on the design process of predictive information visualization, which is the main contribution of this study.

3.1.1 Designing Visualizations of Predictive Information. We initiated the design work of the predictive information interface by applying XAI grounded on earlier findings that explainability can help a user gain insight from the data [41, 44, 63]. Our iterative design process repeats prototyping and conducting internal validation within the research team regarding technical feasibility and user experience. We designed our visualization styles with the following considerations based on earlier research on human-centered XAI [19, 44] and algorithm-mediated self-reflection system [7, 25, 28]: (a) whether visualization can provide insight into self-reflection and actionable behavior change to increase well-being [7, 25], (b) whether the design of predictive information is appropriate regarding the amount of information and cognitive burden given the usage context of MindScope, which is frequently used in daily life [75],

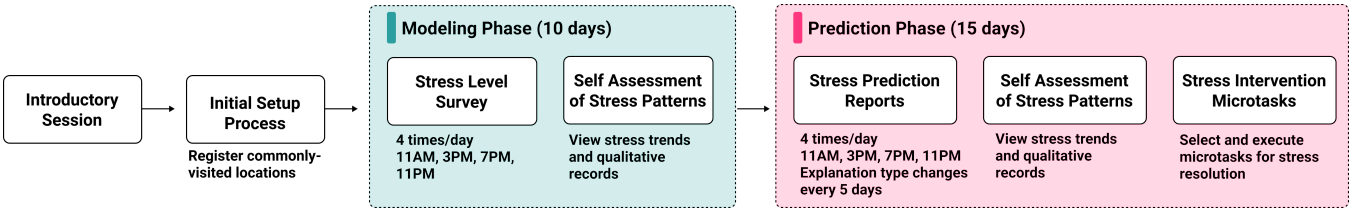

Figure 2: System user interaction flow of MindScope

and (c) whether the visualization contains any confusing or misleading expressions so that people without background knowledge in AI or data science can easily understand [44].

The design space we first considered was the *needs for explainability*. We referenced Liao's XAI Questioning Bank, which provides an overview of the type of explainability needs [44]. Through investigating this design space, we could made a decision on "what to explain" (*i.e.,* content of an explanation) [21]. Our system was designed to provide a stress level prediction within a certain period (*e.g.,* stress level during 1 ~ 5 pm) to help users' reflect on stress-related information. In particular, our system needs to go beyond simply providing users with accurate stress predictions and provide them with opportunities to reflect on stress prediction results. Therefore, providing explanations on why this prediction was made was needed, such as "what data (feature) led the system to make this prediction?" Based on this consideration, we decided to provide an explanation of a single prediction (*i.e.,* local) to provide insight into users' self-reflection.

The design space also considered *user type* [21]. When designing explainability, users' AI knowledge and experience should be considered. We were able to set the direction for "how to explain" a problem based on investigating this design space [21]. One of the factors we mainly considered while iterating the prototype was the solutions for explanation (e.g., UI element, graph, and textual explanation). Our initial two prototypes were designed to visualize the SHapley Additive exPlanations (SHAP) values, a XAI technique that calculates the influence of each feature on prediction results [47], using radar and bar charts (see Figure 3). Regarding the visualization medium, we found quantitatively visualizing features' importance in graphs could confuse and mislead users because it requires advanced experience and knowledge in a related field (e.g., AI or data science) [19]. Moreover, quantitatively presented explanations are expected to be inappropriate for systems that aim at actionable behavior change through reflection. For example, the following questions were reported during our internal validation: *"What does it mean to have an activity feature located on the outer edge of a radar chart? Does it means I walked too much or I have to walk longer?", "At what level does feature importance become meaningful?", and "On what basis can we determine that a particular feature is important but another is not important?"*. Furthermore, earlier work on algorithm technology reflection systems also utilized natural language to deliver the algorithmic output and positively reported their efficacy [7, 25]. Therefore, we delved deeper into how natural language such as words or sentences can be used to explain predictions for ease of understanding.

We further considered the *level of detail in explanation* [75]. The explanation in our system was used to support the user to examine and reflect on information related to his or her stress level based on the data. However, relatively few studies have investigated the level of explainability related to self-reflection, which is also our motivation for the study. We first considered the context of using MindScope, which is 'everyday' PI system [21]. Because users frequently encounter our system more than four times a day, we tried to minimize the cognitive burden that excessive information can induce [75]. In addition, earlier work on the algorithm reflection system reported that providing insight that can support users' behavior change in an actionable way is important [25, 28]. Based on the above points of view, we developed an explanation method providing a limited number of behavior-level explanations (*Type 3*). In designing the prototype of *Type 2*, we considered the earlier findings that providing a detailed level of analysis may lead a user to ignore their interpretation of stress prediction [29]. From this point of view, we developed the explanation *Type 2* with a lower level of detail than was used *Type 3*. Through an iterative design process, we finalized three prediction visualization methods that vary in the degree of detail of the explanations. Detailed stress prediction visualizations are described in Section 3.3.1.

3.1.2 Pilot Study and System Design Iteration. We conducted a pilot study with 30 college students using the initial version of the prototype. The initial prototype provided a single view of stress prediction visualization, which is the most detailed and granular explanation, such as major deviations from the norm in each data point (e.g., using an app more/less than usual). We did this to conduct an initial evaluation of MindScope's core feature while reducing the complexity of our pilot study. Further, this decision was made to verify the technical feasibility and usability of the system using the most complex explanation type, which generates the most diverse explanation outputs that have high potential for unexpected output or errors. As the previous work reported that output complexity is a major design challenge for AI products [86], we wanted to ensure that the prediction information and explanatory properties are adequately generated in real-world deployment. The rest of the system's primary functionality was consistent with the complete system while minor details were slightly different. We will further explain the modification made after the pilot study in the next section. During the pilot study, the researchers used a chat service to communicate with the participants at any time to collect reports on bugs, errors, questions, and concerns related to the study. We also conducted follow-up interviews with ten participants.

Our pilot study revealed several usability and logistics issues. For instance, the pilot system provided the last questionnaire to

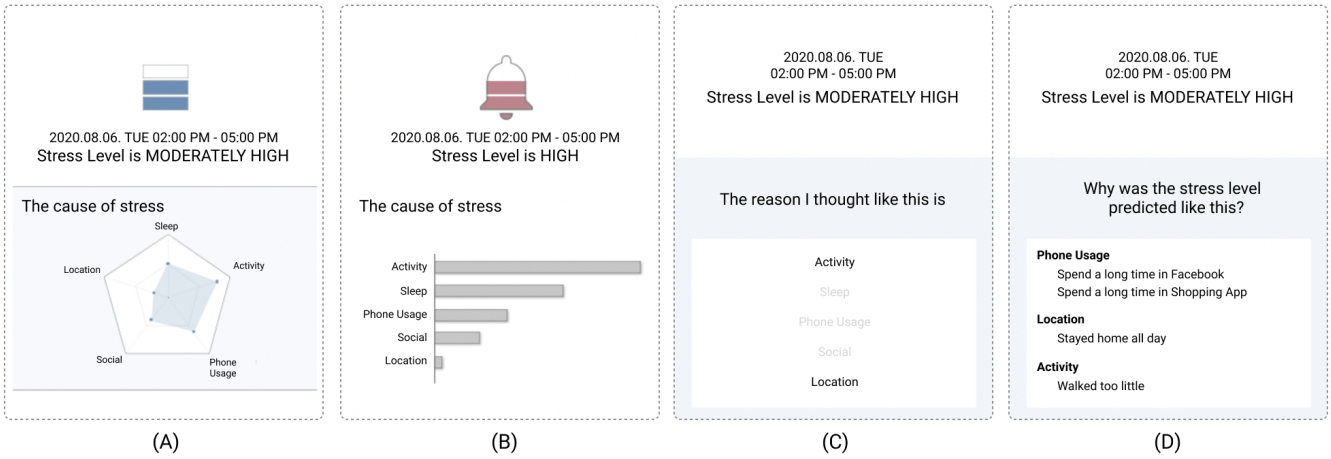

Figure 3: Prototypes created during the design process.

participants at 11 pm, and we limited the response time to two hours. However, it was difficult for those who went to bed earlier than 11 o'clock to answer the questionnaire. Based on this problem, we changed the final system design to have participants answer the last questionnaire by 7 am the following day. In addition, we found the 2-hour time limit, which is the available time to answer the questionnaire when they receive the stress report, put a burden on users to participate in the study. In our final system, participants could check the report and input their stress level up to 30 minutes before the following report is provided, thereby reducing the restriction of the reporting activity. Participants complained of boredom during the EMA process during the initial 10-day modeling phase, which could negatively affect the user experience and data collection process for model building. In the final system, we added self-reflection features, such as a calendar view and a journaling feature to support users review and reflect on their past stress, in the modeling phase to improve the user experience during the first phase, which would naturally connect with the self-assessment–related features of the second, prediction phase. These measures were important for designing the user experience of MindScope, a data-driven algorithmic PI system where the data collection process is critical.

We also found that the way prediction results are explained could significantly influence users' perception of MindScope. For example, some participants' reliability and evaluations were lowered due to discrepancies between specific stress-explanation content and their perceptions. Overall, consistent with earlier findings [46, 76], participants reported that the contents of the explanation had a significant influence on their perception and expectation of MindScope. However, few studies examine the effect that the level of explanatory detail can have on self-reflection in an intelligent PI system. Therefore, we felt the need to observe whether the users' experience of self-reflection varied depending on the level of analysis. Based on the pilot study, we expanded the final study design to encompass diverse levels of explainability to investigate further the effects of explainability on user experience during self-reflection. We note that the design of this app was intended to facilitate and observe the way people might interact with an algorithmic system in terms of their perceived utility or emotional reactions. Therefore, our primary task was not improving the accuracy of predictions or technical precision when building the prediction algorithm. The section below details the complete MindScope system by explaining the modeling phase and prediction phase.

3.2 Modeling Phase

3.2.1 Setup for Data Collection. During the modeling phase, we collected extensive records of each participant's mobile phone usage for stress prediction. The collected data categories included activity information based on accelerometer and GPS data, app usage, app type, and noise levels in the surrounding environment. When users first sign up for MindScope, users are asked to register commonly visited locations (e.g., home, workplace, school, etc.) to add context information to the collected data and improve the model performance.

3.2.2 Self-Assessment and Logging of Stress. In addition, we collected user stress data through ecological momentary assessment (EMA) surveys conducted four times during the day at four-hour intervals. Participants would receive push notifications at 11am, 3pm, 7pm, and 11pm to nudge them toward answering. We adopted this high-granularity approach to account for the daily and regular life routines, which are highly dependent on the time of day and day of the week. In each survey, participants responded with their perceived stress level among *low*, *moderately high*, and *high* (see Figure 4-(D)).

If there was any qualitative information that they wanted to record in correlation to their stress levels, they were able to record a set of hashtags to provide more context regarding their current situation. Their recorded stress levels, as well as the logged data, were organized into a calendar view where the users could review their stress trends and events on specific days. The use of calendar view was intended to provides users with the benefit of being effective in understandability in an familiar format [40]. In the calendar view, the average stress level of each day was provided as a summary, and the dates were color-coded for the visualization.

3.3 Prediction Phase

After the modeling phase, participants moved on to the prediction phase. In this phase, we provided two main features to aid in stress management: 1) personalized stress reports and 2) intervention through microtasks.

3.3.1 Personalized Stress Prediction Reports. In the prediction phase, we considered how presenting the results of predictions would be the most helpful to users when providing algorithm-based prediction services in the field of mental health. In addition, our pilot study results confirmed that explainability of prediction and its visulization seemed to have a significant impact on user perception and reliability of MindScope system. Accordingly, we felt the need to observe whether the user response varied depending on the level of analysis when the model's accuracy was imperfect (see Figure 4-Upper row).

We divided the interface into three different types according to the level of explanation: 1) *Type 1*: No explanation, 2) *Type 2*: Categorical explanation, and 3) *Type 3*: Detailed explanation. In all types, the user received a summarized stress prediction of either low, moderately high, or high. In *Type 1*, only this information was provided (see Figure 4-(A)). In *Type 2*, we organized the data into five categories (phone usage, social activity, movement, physical action, and sleep) and highlighted data categories that had significant weight in calculating the final prediction (see Figure 4-(B)). In *Type 3*, we provided more detailed and granularized context such as major deviations from the norm at each datapoint (e.g., using an app more/less than usual). We selected the five data points that were considered the most significant (those given the highest weight in the final prediction calculation) and showed them to the user (see Figure 4-(C)).

As in the modeling phase, users received four reports each day and were able to retrospectively check the patterns and data in the calendar view (see Figure 1-(Right)). After receiving each report, the user was asked to confirm whether the prediction was correct by logging their actual perceived stress level. In Type 2 and 3, detailed information was provided procedurally after confirming the prediction result. They were also asked whether they found the explanation useful. To collect the causes of the prediction error, every three times a user responded that the explanation was not useful, they were given a survey asking them to identify the issue(s) they had with the explanation.

3.3.2 Stress Interventions through Microtasks. To promote user responses to perceived high-stress situations, we also implemented an intervention system. MindScope employs a stress intervention scheme in which users can configure one microtask to relieve stress and the system alerts a user to perform their microtask at opportune moments [31]. Microtasks offer the advantage of requiring little time and effort to perform while also providing the user with a sense of agency (see Figure 4-(E)). The system detects opportune moments to perform the microtask and sends a push notification to suggest the intervention's execution at that moment (see Figure 4-(F)). In MindScope, we define an 'opportune moment' as a short period of time when a user is available for the execution of a microtask intervention. We used habitual phone usage as a proxy for detecting availability. Our rule for detecting habitual phone usage is based on the short duration, isolated, and reward-based (SIRB) concept introduced by Oulasvirta et al. [60]. Based on these findings, we sent a notification when the user would use their phone for a short period of time (<30 seconds) with an interval of >10 minutes from the previous session.

We populated an initial set of 147 microtasks with responses from a survey deployed to online university forums which is also our target group. Respondents of the survey were asked to suggest microtasks that they like to use in everyday life to resolve or mitigate stress. The survey responses were iteratively examined by four researchers to determine inclusion of the microtasks in our system regarding its appropriateness.

When the push notification arrives, the user can either perform a stress intervention that they have set up or receive a new suggestion from the intervention list until the method the user wants pops up by pressing "Do something else". The user also can defer the performing the intervention (see Figure 4-(F)). The system also allows users to search for stress interventions from the list on the intervention setting page (see Figure 4-(E)). We included these features since a previous work found that the users valued viewing interventions created by others when it is challenging to find one suitable [42]. Users could log their completion of a microtask either in the app or through the push notification, and the history of completed microtasks was shown with timestamps (Figure 4-(E)).

3.4 Implementation Details

MindScope builds a personalized stress prediction model through the modeling phase and continuously updates it based on user feedback during the prediction phase. Figure 5 shows the MindScope system constructing a personalized stress prediction model, generating prediction information, and updating the model. Below, the technical details are explained in terms of data collection and processing, and machine learning for stress prediction algorithm.

3.4.1 Data Collection for Stress Prediction. Recent studies report on the relationship between passive sensing behaviors from mobile phones including social activity, movement, physical action, sleep and phone usage (app usages) with mental health outcomes such as depression, stress, loneliness, and flourishing [8, 51, 65, 82, 83]. To gauge social activity, MindScope collects the number of incoming and outgoing phone calls, call duration, and missed calls, based on timestamp. The app also records audio loudness (ambient noise) in decibels every 20 minutes for five seconds where the silence threshold is set to -65 dB. For the movement, MindScope periodically checks if the user has changed location by more than a certain number of meters. We set a 5-minute periodic check-up and 10 meters for the threshold. For sleep, MindScope tracks `SCREEN_ON_OFF` state data. From 6:00 pm to 10:00 am (*i.e.,* potential hours of sleep), the app collects screen off durations and then the maximum duration is considered the amount of sleep. For physical action, MindScope collects data from Google's Activity Recognition and Transition API[1], detects changes in the user activity (`STILL`, `WALKING`, `RUNNING`, riding a `BICYCLE` and on a `VEHICLE` activity), and records the duration of each activity in seconds. Finally, for phone usage, MindScope checks the user's screen status and collects the duration of screen

[1] https://developer.android.com/guide/topics/location/transitions

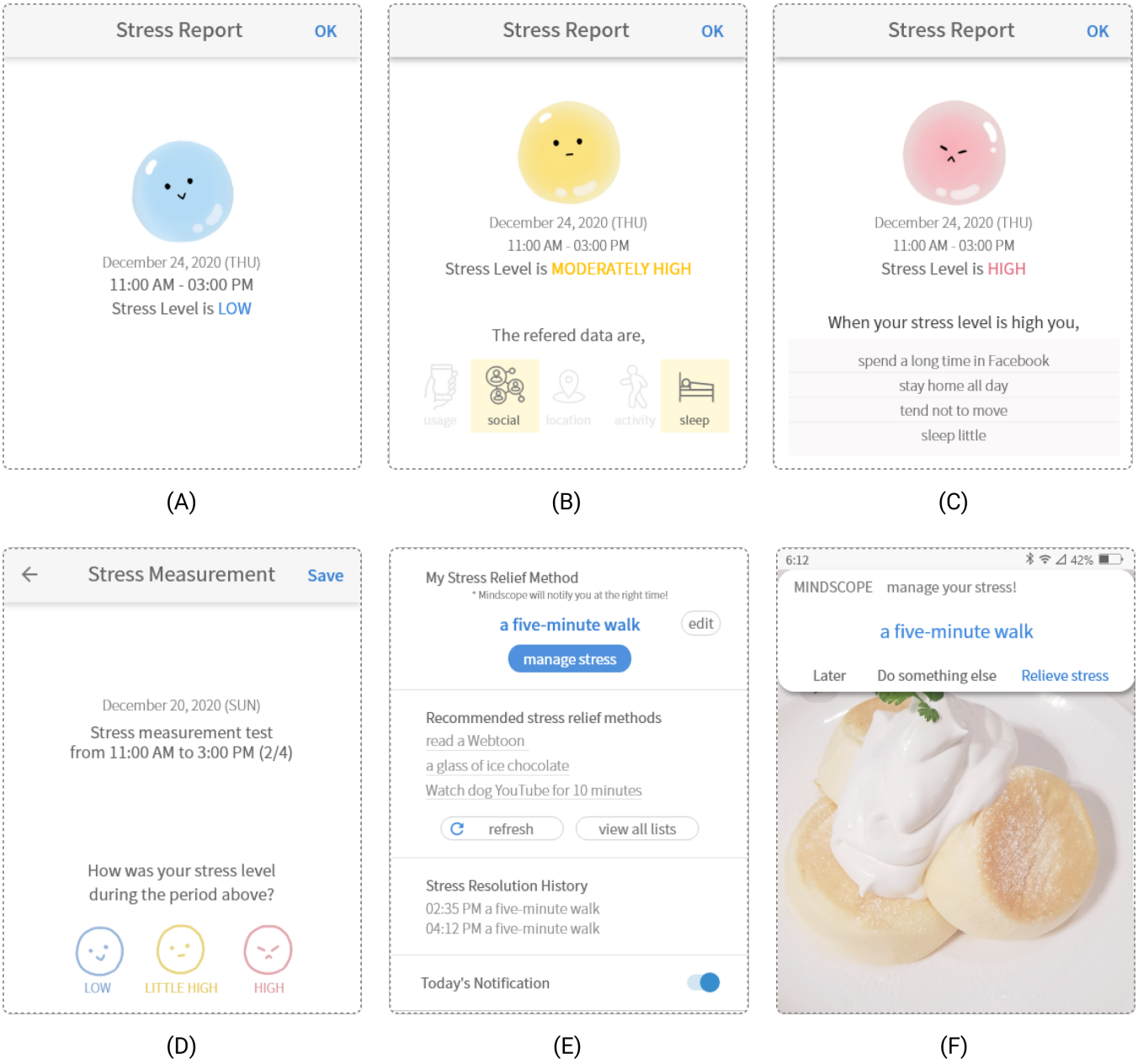

Figure 4: Upper row: A personalized stress prediction reports: (A) *Type 1*-No explanation, (B) *Type 2*-Categorical explanation, (C) *Type 3*-Detailed explanation, Lower row: (D) Stress Assessment (EMA) during Modeling phase, (E) Planning stress intervention, (C) Intervention notification.

unlocked. When the screen is unlocked, MindScope records the usage frequency and duration of other apps that are categorized under 12 groups (e.g., Entertainments/Music, Games/Comics, Social/Communication, Health/Wellness, Education), defined based on the Google Play Store. Overall, MindScope collects a total of 29 data features. MindScope sends the collected data to the gRPC server every 60 seconds, and the server operates modeling. Using gRPC, MindScope can directly call a method on a server, and lightweight and fast data transmission is made between the mobile app and the server.

3.4.2 Machine Learning for Stress Prediction Algorithm. With the data received from the smartphone, the gRPC server starts data preprocessing, including data manipulation (e.g., remove missing rows, duplicates), synchronization (by timestamp), and normalization (min-max normalization was used). We used XGBoost [14],

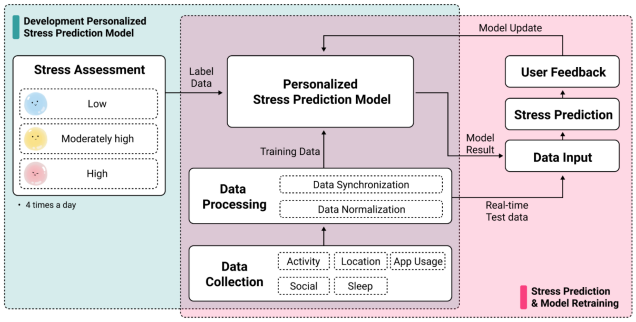

Figure 5: A technical overview of the prediction algorithm used in MindScope

an ensemble algorithm that combines multiple decision trees, for our model development. Our model yielded reasonable performance (68% f1-score) comparable to previous studies that used other machine-learning algorithms, such as logistic regression and decision tree [26, 56]. The model returns a stress prediction result (low, moderately high, and high) with feature importance.

To measure feature importance, we used SHapley Additive exPlanations (SHAP) [47], one of the XAI techniques that use the independence between Shapley values and features. The SHAP values can be expressed numerically by how much each feature contributed to the overall performance. The contribution of each feature can be expressed as the degree of change in overall performance when that contribution is excluded. Based on this concept, the SHAP values for each feature are obtained from the model result, which can be either positive or negative. A positive value means that the feature has a positive effect on the prediction, and vice versa. We considered only the feature in the case of a positive value. Through these processes, MindScope delivers information to a user during the prediction phase. It presents one of the three types of stress prediction reports (Figure 4-(A-C)). For *Type 1*, it displays only a model prediction result without any further explanation. For *Type 2*, it displays a model prediction result and a category that consists of the features with the positive SHAP value. For *Type 3*, it displays a model prediction result with detailed explanation of the five most significant features which correspond to top five SHAP values. After the user checks the report, he or she can give feedback to the model by confirming or adjusting the stress level. The stress level determined by the user will be used as the updated test data for model retraining. As this process continues, the model becomes customized to the user.

4 METHOD

4.1 Field Study

We conducted a 25-day field deployment study with 36 participants using MindScope to examine how personalized stress prediction algorithms affect stress management. In particular, in this experiment, we further investigated how model explainability affects stress management. To this end, we designed the experiment so that all participants could test all three types of explanations. We set the order in which the explanation was presented in three different ways, from the perspective of randomized assignment to the app use. Each explanation type lasted for five days. This experimental design allowed participants to better evaluate each design's features, pros, and cons by comparing them with the others. The study was composed of three parts: (1) a 30-minute introductory session to provide the background for the study and introduce MindScope; (2) a 25-day MindScope usage study in the field; and (3) a 40-minute follow-up interview to elicit user feedback on the perceived impact of stress management using MindScope. The 25-day app usage period consisted of a modeling phase for the first 10 days and a prediction phase for the subsequent 15 days (Details are documented in the 3. System Design section). Our study was approved by a university's IRB, and informed consent was obtained from each participant. Since our study required careful measures in that our system collects participants' personal information, we provided detailed information on data collection and used it in the informed consent provided before participation in the study, based on the advice from the IRB. In the introductory session, we informed the following contents in detail: 1) the types of collected data, 2) the granularity of the collected data, and 3) the data management method. We also provided a Q&A session for participants with questions. After we distributed the app, the researchers used a chat system where participants could ask questions or concerns about the system. Before analyzing the data, we removed personally identifiable information from the collected data and used anonymized code names to preserve the participants' privacy.

4.2 Recruitment

We recruited university undergraduate and graduate student participants who use Android smartphones. For conducting 25-day field study, we have specified the following requirements and restrictions for participation: 1) Those who cannot use their smartphones for more than two days during the experimental period forcibly, 2) Those who have difficulty participating in the required experimental process (online introductory session, pre-and post-survey), 3) Those who cannot install MindScope and, who plan to change their smartphone model during the study. Recruitment announcements were posted across six Korean universities' online communities or via batch email. All experiment processes including recruitment, orientation, and post-interview were conducted by video meetings to comply with necessary safety guidelines in the COVID-19 pandemic. Notices and inquiries were handled through an online group chat. We compensated each volunteer $25 USD for their participation, which entailed an introductory session and 25 days of usage. Additional $20 USD compensation was provided to people who were willing to attend a follow-up interview. To induce users to participate continuously, the accumulation system was used for the honorarium. $0.25 USD was set aside for each survey completed for the modeling phase, and for each stress report in the prediction phase.

4.3 Data Collected

4.3.1 Perceived Stress Scale (PSS). To measure the impact of MindScope on stress, we used PSS—a widely used psychological instrument—to measure stress perception [17]. PSS consists of ten items,

each of which is scored on a scale of 0 to 4. The result is calculated as the sum of the ten items' scores where higher results indicate greater stress. In the case of PSS, evaluation was conducted before the experiment (Pre), after the 10-day modeling phase (Mid), and at the end of the study (Post).

4.3.2 User Experience Questionnaire. To evaluate the user experience with MindScope, we selected 15 UX-related items from user experience and usability studies [2, 58]: 1) useful, 2) easy to use, 3) easy to learn, 4) effective, 5) efficient, 6) comfortable, 7) friendly, 9) consistent, 10) fulfilling, and 11) fun. Users evaluated each item using a seven-point Likert scale ranging from strongly disagree to strongly agree. For the User experience and Algorithm perception questionnaire, participants were asked to complete a questionnaire when the explanation type changed (there were three questionnaires over 15 days in total with five-day intervals).

4.3.3 Algorithm Perception Questionnaire. To measure the user's perception on the algorithm of MindScope, we included additional questionnaires. Six items were designed for assessing level of trust, perceived accuracy, and explanation meaningfulness [57, 87]. Level of trust was measured by using a seven-point Likert scale (1: "I didn't trust it at all", 7: "I fully trust it"). Perceived accuracy was measured in two ways, using questions asking a numerical estimate of the system's accuracy on the scale of 0 to 100 percent and using a seven-point Likert scale (1: "not accurate at all", 7: "very accurate"). To understand the evaluation of explanation, we used seven-point scales asking the usefulness, convincingness, and sufficiency of the prediction explanation.

4.3.4 Application Usage Log. To measure the user behavior in the MindScope app, we embedded a tracking code as part of the development process. In this study, we analyzed and reported log data on the individual user's self-report, and stress interventions for understanding user engagement of the system. In addition, we collected both stress levels from the algorithm initially predicted and user-entered stress levels for further analysis on users' agreement on system output.

4.3.5 Follow-up Interview. After participants completed the 25-day deployment field study, we conducted a semi-structured interview to understand users' detailed experiences on stress management using MindScope. The interview lasted between 40 minutes and one hour and was conducted via a Zoom-video meeting. The interview mainly covered the following four themes: 1) Understanding participants' motivation to participate in the experiment and existing stress-management skills, 2) How MindScope affected participants' understanding and perception of stress, 3) How participants accepted and understood MindScope's algorithm, 4) Overall usability and user experience evaluation.

4.4 Analysis

In quantitative analysis, we first aimed to investigate MindScope's stress management effect through PSS scores. Then, we investigated how the different types of explanations affected the user experience and algorithm perception. We analyzed the results from questionnaires using one-way repeated measures ANOVA (RM-ANOVA) with Greenhouse-Geisser correction. Tukey's HSD was used as a post-hoc analysis to provide further understanding by pairwise comparison. The test was performed using GraphPad Prism version 9.0 [2]. In qualitative analysis, we analyzed the qualitative data from interviews by conducting open coding with thematic analysis [11]. All interviews were transcribed for analysis. Three researchers individually read the interview transcripts and generated open codes. The open codes were discussed among research team members to resolve disagreements and identify patterns and we then generated themes from these open codes. We identified statements that revealed how MindScope supported users' stress management behavior and how users perceived the MindScope system providing stress prediction information and then structured the themes around understanding 1) effect of algorithmic stress prediction on stress management 2) user perceptions and reactions on the algorithmic stress prediction, and 3) effect of explainability on algorithm perception and stress understanding.

5 RESULT

In this section, we first summarize the descriptive statistics results to illustrate how participants engaged in MindScope. Then, we present statistical findings showing the effect of using MindScope for stress management. In the discussion of the qualitative findings, the effects of MindScope in stress management, user perception of the stress prediction algorithm, and the effect of explanation type on stress management are summarized.

5.1 Descriptive Statistics

Thirty-nine students across six universities signed up as potential candidates during the recruitment period, but three of them dropped out at the introductory session. We finally selected 36 participants who fulfilled our requirements. Among them, 16 participants identified as male and 20 as female. The age range of the participants was 19–33 years (mean = 25.3). There were no cases of participants dropping out during the course of the experiment. Follow-up interviews were held by voluntary participation, and 34 participants ultimately completed an interview.

5.1.1 User Engagement. During the modeling phase, participants were asked to answer a stress-level EMA survey, four times a day, for 10 days. Participants answered a total of 1,123 out of 1,440 surveys, an average of 31.19 times per user (Min = 2, Max = 39) at an average response rate of about 78%. For the prediction phase, users were provided stress reports where they could either confirm or correct the prediction made by the stress prediction model through self reports. They submitted 1,842 self-reports out of 2,160 possible instances, averaging at 51.16 submissions per user (Min = 19, Max = 58) and at a 85% average response rate. We note the users consistently managed a high rate of engagement despite the repetitiveness and demanding quality of the tasks. We will revisit the possible explanations for participants' engagement later in the qualitative result section.

For the stress intervention feature in the prediction phase, 31 users (86% of all users) used this feature at least once, either through registering a microtask or actually executing it. Users performed stress intervention tasks a total of 751 times (Min = 2, Max = 56),

[2] https://www.graphpad.com/scientific-software/prism/

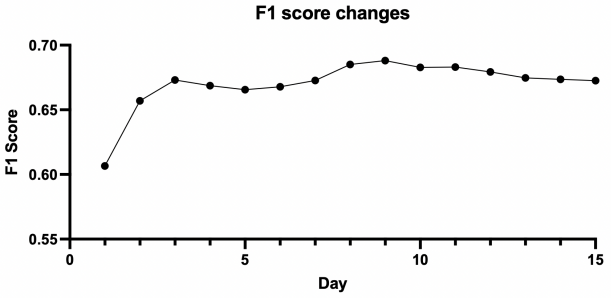

Figure 6: The stress prediction model performance changes over 15 days of the prediction phase.

with each unique intervention being used an average of 4.94 times. In addition, users committed to 152 unique interventions including examples like 'stretching', 'watch a short Youtube clip', 'text a friend', and 'organize my schedule'. Of the 751 cases, 146 of them were in response to a push notification from the app, and 605 of them were logged independently, implying that the users preferred to execute stress interventions of their own volition rather than being prompted by the app. Users also spent a lot of time searching for other things to do, evidenced by the fact they clicked on the 'Do something else' button a total of 2,563 times when choosing a microtask for their stress intervention, which was more than any other type of engagement in the stress intervention feature group. Consistent with earlier findings [42], users found browsing a list of microtask interventions created by others valuable because it provides a new perspective on stress relief. Overall, we found that our participants were highly engaged in the day-to-day stress intervention execution in their own way.

5.1.2 Stress Level Changes Recorded on MindScope. Stress levels were recorded on MindScope using a three-level scale. When responding to a stress level assessment question, users were asked to choose an appropriate stress level between Low Stress, Moderate Stress, and High Stress. We organized the aggregation of stress levels that our participants have reported through the study period. The majority of the self-reported stress levels were low in both phases, increasing slightly in the Prediction phase (56.3% in the Modeling phase, and 61.5% in the Prediction phase). Similarly, the total ratio of reports where participants reported having high stress levels decreased slightly across the two phases, from 8.5% in the Modeling phase and 5.8% in the Prediction phase. While receiving system reporting predictions with explanation *Type 2* and *Type 3*, participants were also given the chance to review the accuracy of the analysis provided. When a participant would respond that it was not accurate, they were prompted to complete a survey every third report asking why it was inaccurate. We received 118 responses in total, with duplicate responses recorded. The majority of the cases reported that the analysis was incorrect (100 responses, 84.75%), or that the analyzed result was unrelated to their perception of stress (89 responses, 75.42%). Other cases mentioned that the results were unclear (40 reponses, 33.90%) or that they were too trivial (5 responses, 4.24%).

5.1.3 Changes in Prediction Model Performance. In the Prediction phase, participants were provided an algorithm-generated stress report of their stress levels, and were given the chance to confirm or adjust the system's prediction by responding with their own perceived stress level. Through the log data analysis, we were able to check whether the prediction was correct or not by comparing the initially reported value by the system with the user's self-reported stress level. Out of the 1,842 cases, 1,177 cases (63.89%) were consistent with the users' self-reported stress level, and 665 cases (36.1%) were adjusted by the user. Meanwhile, the mean score of perceived accuracy was 59.97% (SD = 17.7). This result was measured using a questionnaire asking, "How accurate do you think the system is?". Out of the 665 cases which user corrected the stress level, 506 cases (27.47%) predicted a lower stress level than the participant actually experienced, and 159 cases (8.63%) predicted a higher stress level than the participant's self assessment.

To supplement this data, we further investigated how the performance of our model changed throughout the study period (See figure 6). First, we noted that during the Prediction phase, we found an increase in the average model performance from 61% to 68%. Within that 15-day phase, the model performance would usually rise above 68% on the eighth day, but would then be saturated and not improve further. The algorithm conducted a retraining process based on the confirmed stress level we received from the users. As retraining continued, the average performance converged to the overall average (68%) stated above, which can be considered the effect of stabilizing the stress prediction models based on user feedback. There were no significant group-wise differences based on the explanation type.

5.2 Statistical Findings

In this section, we summarized findings from the statistical analysis conducted on the questionnaire responses. First, we investigated the effect of using MindScope for stress management. The result shows significant reductions in stress after a 25-day field study. From the analysis of questionnaires on user experience and algorithm perception (i.e., user trust and perceived accuracy), we first found that algorithm perception positively correlated with the system's usefulness and effectiveness. We then found that providing explanations were useful but found no differences between categorical explanations and detailed explanations.

5.2.1 Stress Significantly Reduced first, then Persisted. To investigate the stress management effect of MindScope, the perceived stress level was measured before (Pre), during (Mid), and after the experiment (Post). 'Pre' was measured before participants started using MindScope, 'Mid' was measured after participants completed the 10-day Modeling phase, and 'Post' was measured after the 25-day field study was completed. We first confirmed a significant difference in the participants' stress levels at the three different time points through the RM-ANOVA test ($F(1.968, 68.86) = 10.00$, $p = .0002$) (See figure 7-(A)). Using Tukey's HSD as a post-hoc test, pairwise comparison was performed. We found significant stress reduction in Pre vs. Mid ($p = .0034$, 95% C.I. = [0.9051, 5.039]), and Pre vs. Post ($p = .0010$, 95% C.I. = [1.388, 5.890]). There was no significant change in perceived stress level in Mid vs. Post ($p =$

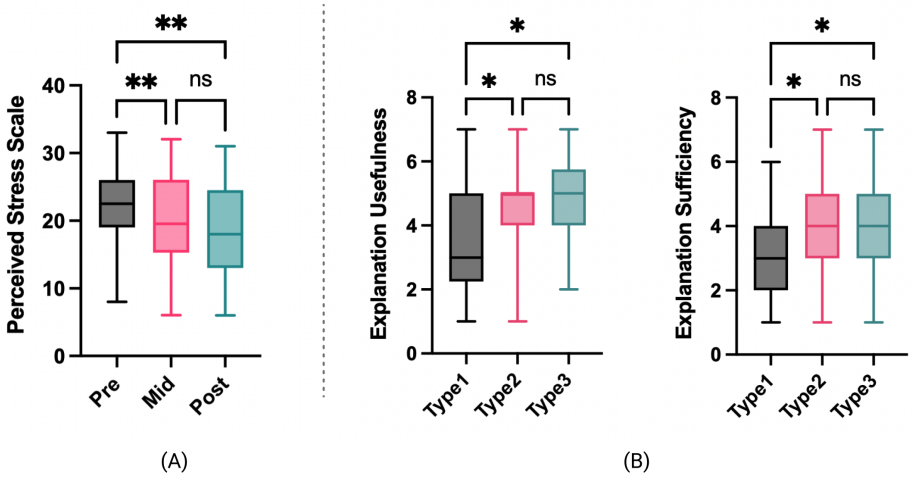

Figure 7: (A) Changes in Perceived Stress. 'Pre' was measured before participants started using MindScope, 'Mid' was measured after participants completed the 10-day Modeling phase, and 'Post' was measured after the 25-day field study was completed. (B) Evaluation of explanation usefulness and sufficiency depending on explanation type. Statistically significant results are reported as *<0.05, **<0.01, *<0.001**

.7047). This result indicates that reduced stress during the early stage of Mindscope persisted until the later stage.

5.2.2 The More Accurate It Is Perceived, the More Useful It Is. Since user perception factors such as trust are a critical factor in the adoption of systems and their outcomes [64], we aimed to investigate the relationship between the effectiveness of MindScope and users' algorithm perception (i.e., user trust and perceived accuracy). To this end, we analyzed the correlation between user trust and perceived accuracy vs. user experience scores. We adjusted the p-values for multiple inference using Holm's method [30]. The results showed that users' overall algorithm perception positively correlated with the usefulness and effectiveness of the MindScope system (See figure 8). We found a moderate correlation on user trust vs. useful ($r = .681$, $p = .0007$). and efficient ($r = .690$, $p = .0005$). and showed a strong correlation between user trust and effective ($r = .799$, $p \leq .0001$). Users' perceived accuracy also showed a similar correlation (accuracy vs. useful ($r = .674$, $p = .0010$); accuracy vs. efficient: ($r = .626$, $p = .0003$); accuracy vs. effective: ($r = .699$, $p = .0064$). However, no significant correlation was found in other user experience indicators. These results imply that the more users perceive an algorithm to be accurate and trustworthy, the more likely they will appreciate the system's usefulness, effectiveness, and efficiency.

5.2.3 It Is Better Than No Explanation, but More Doesn't Make It Useful. Then, we performed further analysis to see whether there was a significant difference in user experience and algorithm perception according to the explanation type. We confirmed that there was a significant difference between the three types when evaluating the usefulness and sufficiency of the prediction explanation (usefulness: ($F(1.724, 60.35) = 5.738$, $p = .007$), sufficiency: ($F(1.779, 62.25) = 6.664$, $p = .003$) (See figure 7-(B)). As a result of observing the difference from a pairwise point of view through the post-hoc test using Tukey's HSD, both evaluation of the usefulness and sufficiency showed a significant difference between *Type 1* vs. *Type 2* (usefulness: $p = .042$, 95% C.I. = [-1.476, -0.02445], sufficiency: $p = .017$, 95% C.I. = [-1.433, -0.1229]), and *Type 1* vs. *Type 3* (usefulness: $p = .023$, 95% C.I. = [-1.830, -0.1140], sufficiency: $p = .013$, 95% C.I. = [-1.864, -0.1920]). However, no significant difference was found for *Type 2* vs. *Type 3* (usefulness: $p = .642$, sufficiency: $p = .618$). This result implies that the participants evaluated the *Type 2* and *Type 3* as more useful and effective than *Type 1*, but there was no significant difference between *Type 2* and *Type 3*. No significant differences were found in other questionnaires on user experience and algorithm perception. In our qualitative finding section, we described how the participants evaluated the pros and cons of each explanation type and how explanation type affected their self-reflection.

5.3 Qualitative Findings

5.3.1 The Impact of Algorithmic Stress Prediction and Explainability on Stress Management Practices. Our participants reported that the algorithm-assisted approach helped them understand detailed stress patterns and plan interventions to cope with stress. In addition, they responded that the stress prediction information lowered the ambiguity in the process of recalling their past stress levels. Overall, MindScope's method of analyzing one's stress level received positive feedback from participants. Many of our participants reported that being regularly provided with predictive information about oneself was the most interesting part of the MindScope experience. Further, The intimacy established through interacting with MindScope allowed the users to engage in the system actively.

Identifying Stress Patterns through a Data-Driven Approach: Overall, our participants reported that MindScope enabled a concrete understanding of stressors and patterns that they could not clearly understand in the past because they thought that the system

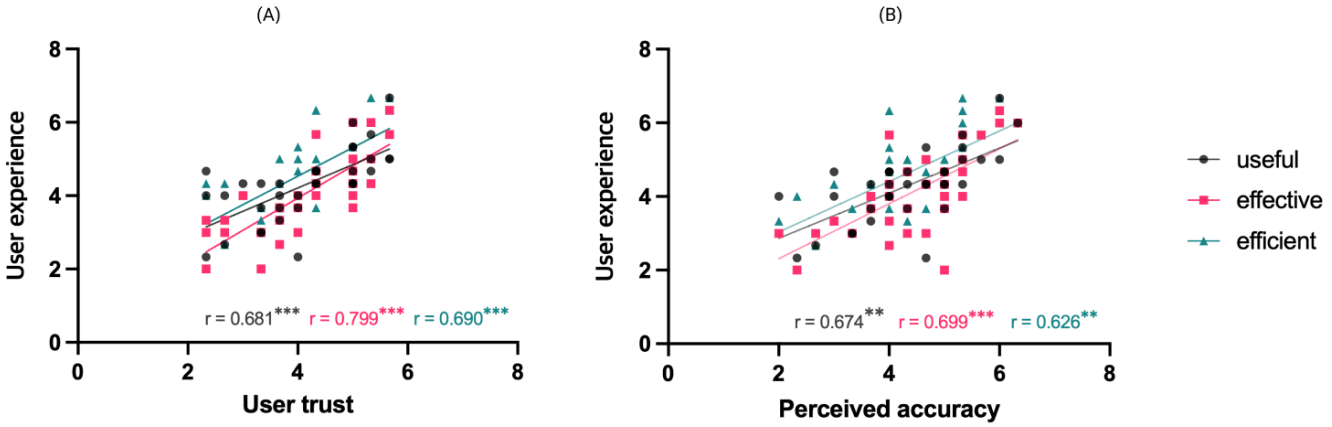

Figure 8: (A) Scatter plots representing the correlation between the user trust and the three user experience score (useful, effectiveness, efficient) (B) Scatter plots representing the correlation between the perceived accuracy and the three user experience score (useful, effectiveness, efficient) 'r' = Pearson's correlation coefficient. Statistically significant results are reported as *<0.05, **<0.01, *<0.001**

Table 1: The Multiple Comparisons Table of the *Explanation Usefulness* and *Explanation Sufficiency* Score.

	Pair	Mean Difference	Std. Error	Adjusted P-Value	95% Confidence Interval	
					Lower Bound	Upper Bound
Tukey HSD	***Explanation Sufficiency***					
	Type1 vs Type2	-0.7778*	0.2676	0.017	-1.433	-0.1229
	Type1 vs Type3	-1.028*	0.3415	0.013	-1.864	-0.1920
	Type2 vs Type3	-0.2500	0.2654	0.618	-0.8995	0.3995
	Explanation Usefulness					
	Type1 vs Type2	-0.7500*	0.2965	0.042	-1.476	-0.02445
	Type1 vs Type3	-0.9722*	0.3507	0.023	-1.830	-0.1140
	Type2 vs Type3	-0.2222	0.2460	0.642	-0.8241	0.3797

objectively determined one's stress level using mobile sensor data. *"In my case, I found that I get really stressed when I don't follow my daily routines like, exercising in the morning or having my meals at the right time. I also found that I get stressed when I chat with my friends" (P2A).* In addition, MindScope's stress report provided an opportunity to re-establish users' preconceptions about stress. Some participants responded that they already had a rigid stereotype about how their daily activities correlated with their stress levels (e.g., physical activity would help reduce stress). However, participants mentioned that they would re-establish their mental model about stress when they received a stress report contradicting their previous assumption. *"My stress level was lower on days I didn't walk much compared to the days I walked a lot. I used to think that I'd become less stressed by outdoor activities, but now I've learned that I rather become less stressed when I rest in my room doing my things then busily walking around" (P28C).*

Supporting Reminiscence with Data about the Past: The stress prediction information helped participants recall their past stress levels and related events. Our participants reported their stress levels throughout the study, without any prediction information for the first 10-day modeling phase and with the received stress prediction for the next 15 days. By using a MindScope stress report as a reference, participants were able to think about what they actually did, and what their stress was like during the last three hours. Participants reported that this process helped them recall stress levels and related events with less ambiguity. *"The first (Modeling phase) was done by prediction without any standards, and was subjective. Things went ambiguous then but the app started predicting from the mid-late period. The prediction made me wonder if I really was at the state when, let's say when I got a rating of high. I wondered and came to agree that I might have been highly stressed" (P4A).*

Planning Actionable Stress Intervention to Offset Identified Stressors: MindScope's stress report with explanation about the context of the current stress level encouraged participants to establish detailed, actionable stress interventions. Our participants utilized the explanation describing possible factors affecting the current stress level (e.g., decreased social activity, increasingly sedentary) to plan an intervention specifically designed to remedy the problem prescribed by MindScope. A participant said that he scrutinized MindScope's stress reports especially when his stress level was rated low, then he strategically selected an activity he actually

showed in the report as his own stress-reduction method. *"I began to think about things that relieve my stress. It showed some activities and I thought about the things that make my stress rates lower, and it was when I listen to music a lot. And when I move from place to place frequently, I guessed that walking while listening to music relieved my stress" (P15B)..* Contrary to the example above, some participants found activities they performed in the past when stress was high were not working to relieve stress. They then would hypothesize that they might be able to reduce stress if they did the opposite. *"Then, if I change my pattern of behavior, like, I thought I could find a way to relieve my stress through this (device). If you see here, when it says you become highly stressed when you spend a lot of time walking, I think about experimenting myself to see if my stress rates are lower when I only move by car." (P28C).*

5.3.2 User Perceptions Toward Stress Prediction Algorithms. We found that the participants' background knowledge of AI-related subjects impacted their perception and interaction patterns with prediction algorithms. We also found that the time period of interacting with the algorithm (i.e., introductory session, initial setup phase, prediction phase) and the level of model explainability (i.e., *Type 1*: no explanation, *Type 2*: categorical explanation, *Type 3*: detailed explanation) had an impact on users' awareness of data collection and privacy concerns. In addition, because the characteristics of stress provide highly personal data, participants wanted to provide feedback on the stress prediction model.

AI Literacy and Algorithm Perception: Our participants ranged from people with no experience (73%) or to those who were relatively knowledgeable about AI (19%) to experts such as those who had worked on AI projects in academia or industry (8%). The level of such background knowledge acted as a factor in determining how participants reported their perceptions of the algorithm. First, regardless of the level of knowledge of AI, most participants were aware that the more data input, the better the prediction accuracy. Such background knowledge served as a motivation for participants to actively engage in data input processes but in order to receive a more accurate prediction. *"I think there was something like 'proper analysis comes from plentiful answers'. I actively participated for more specific analysis" (P15B).* On the other hand, a participant who has considerable knowledge in machine learning, recognized the limitations of the algorithm in detail compared to others. *"I gave rather plain answers because there weren't enough events during the 10-day analysis, which is a shame because it made the analysis difficult for the algorithm. In fact, balanced results of data are needed for an accurate analysis but I never chose 'high' during my experiment, so I think this made the analysis difficult" (P3A).* In the follow-up interview, rather than criticizing the performance of the system, he reported that the fact that he had always inputted a uniform stress level during the modeling phase served as a limitation.

Challenges in Data Collections: The participants reported privacy concerns caused by personal data collection and the limitation that the data used in MindScope were insufficient to reflect users' stress levels. First, in relation to privacy concerns, we found that the duration of usage and the level of explainability affected users' perceptions of data collection and privacy. At the introductory session, we ran a Q&A session to provide a detailed description of data collection and processing methods. This session was important in addressing participants' data-related concerns. For example, one participant expressed concerns about how the system collects and stores "ambient noise" data without continuous recording. The researchers were able to address participants' concerns by showing the captured image of our database to confirm the user's actual data stored and labeled. However, over time, the effect of providing information initially did not last because some participants usually forgot the information provided in the initial session. The initial access permission process during the setup for the app was also a point where participants raised concerns about data collection. *"I was worried when I constantly had to click grant data collection rights" (P5A).* The level of explainability in the algorithm was also been reported as a factor influencing participants' awareness of data collection. *"At first I thought it didn't collect much, but as it went on (providing more explanation) showing what it has analyzed, such as my amount of phone calls or the radius of my movements, I was amazed rather than uncomfortable" (P9A).*

5.3.3 The Level of Explainability and Self-Reflection for Managing Stress. In the section below, we report on how the level of explainability affected users' self-reflection ability, which is critical in managing one's stress. First, we examine how participants evaluated the three types of explainability options offered by MindScope. We then investigate how participants understood and utilized MindScope's explanations.

The more detailed, the more useful, but the more sensitive to the reliability of the prediction: In the follow-up interview, 62% of participants answered that they preferred a *Type 3* report, which provided the most detailed prediction information. They mentioned that it might be useful to understand their stress by providing the most specific information related to stress. In addition, the provision of such detailed predictive information was highly rated in that it could guide users to take a specific action to mitigate stress. *"Since it shows 'you use facebook when you are stressed,' it makes me think about what I should do next." (P6A).* Further, participants reported that they reconstructed their past based on the descriptions of the contextual information provided in detail in *Type 3* reports and used it to infer their stressors.

However, some participants who preferred *Type 3* reports noted that the accuracy should be guaranteed. Because *Type 3* reports almost directly showed feature variables of the stress prediction algorithm, users often recognized that the features composed of sending data from various channels often did not represent what they actually did. These user reactions seem to be related to the findings of previous studies that the provision of excessive information to ensure the transparency of intelligent systems can lead to lower evaluations of user trust [74]. In particular, these reactions were predominant in the group of participants who encountered *Type 3* reports in the beginning of the prediction phase. Due to the cold-start problem, MindScope often showed a message, "`I am not able to generate a stress report due to insufficient data.`" for the users who received *Type 3* reports first. These users also pointed out inaccurate explanations about stressors. The fact that the model's accuracy was inevitably low at the beginning of the experiment and increased as the study progressed made participants

more sensitive to the model's performance and the algorithm's capability for *Type 3* reports, which reveal the performance of the prediction model in more detail than other types of reports.

In contrast, some participants (27%) valued *Type 2*, which provided only the category of data. We found that *Type 2* reports allowed proactive reflection by deeply investigating specific stress-related activities and events individually. *"If it says 'social activity' I start to think about the kinds of social activities I've done, so I think this type of report [Type2] is more useful for the process of metacognition of stress" (P2A). "I think type2 was useful. Type3 might seem most useful, but type 2 gave me the opportunity to make a hypothesis and draw a conclusion by myself" (P32C).* Moreover, in terms of the trust, the participants also rated *Type 2* higher than *Type 3*, because *Type 2* allowed for their own interpretation and understanding often leading to self-experiment. These participants also reported that even a categorical level explanation is useful enough to identify one's own stressors and patterns.

Participants who preferred *Type 2* reported that because the amount of data was smaller than that of *Type 3*, the effort required for analyzing the information was less. Based on this advantage, *Type 2* was effective in situations where a smartphone could be checked only briefly, such as while on the move or in class. About 10% of the participants answered that they wanted to continue using *Type 1*, which only provides predictive values of the stress level. Similar to the preference for *Type 2* reports, the reason for the preference for *Type 1* was that it was easy to analyze the data provided because it contains the least amount of information.

Explainability Uncovered the Reasoning Process: Providing detailed explanations allowed users to observe the reasoning process for how stress is predicted in the system. For example, observing how the system interpreted specific user behavior into stress level while using explanation *Type 3*, users discovered that the algorithm behaves differently from their perception of stress. This disparity negatively affected users' evaluations of the model's trust and reliability. For example, some participants reported that it was difficult to understand that certain behaviors affected stress in certain directions (e.g., lack of sleep or staying in one place for a long time increases stress). *"The relationship between sleep and stress is ambiguous because… people could sleep late doing something fun, or doing their work. So I'm not sure if this could work as a method to predict stress." (P5A).* Other participants further noted that the algorithm's stress reasoning was not convincing, seeing that the same activity items were presented as an explanation even when predicting different stress levels.

6 DISCUSSION

MindScope is a system to assist participants in collecting and reflecting on their stress levels and patterns by utilizing a prediction algorithm that can complement the limitations of the existing PI system [25, 43]. A 25-day deployment of MindScope allowed users to experience the overall stress management process, including the data input process for modeling a personalized stress model, the reflection process through stress prediction and explanation, and the process of intervention planning and execution. We observed how participants used predictive information and its explanation for self-reflection. Consistent with previous works on algorithm-mediated reflection, our qualitative findings revealed increased self-understanding and self-awareness based on the algorithmic approach [7, 28]. The observation of algorithmic output allowed users to build more specific and detailed behavior changes or interventions [25, 28], and users positively evaluated the scientific nature of the data-driven approach [28]. Our study identified additional findings that users employed predictive information to reconstruct past stressful experiences as well as related information by supplementing their subjective reasoning. In particular, designing and deploying a system that generates prediction and explanation in three ways led us to add new findings to the earlier works on technology-mediated reflection. Our study reports the influence of visualization of predictive information for the PI system. While users highly evaluated both categorical and detailed explanations as more useful than presenting only stress prediction level was, we discovered that the use of explanation data and its influence on system perception differed based on the level of explainability.

Our study result can be extended to the system that explainability plays a pivotal role. Consistent with work on user-centric XAI research, our qualitative findings revealed no meaningful differences in users' evaluation of *Type 2* and *Type 3* explanations [75]. The preference for Explanation varied depending on the individuals' needs [52]. Some users preferred the most detailed explanation for gaining specific insight from the data, whereas others preferred the categorical explanation for the ease of checking the explanation. In our study, we identified that the categorical explanation *Type 2* (*i.e.,* a medium level of explanation but not too detailed) can generate a meaningful user experience by creating space to explore algorithmic outcomes in a user-driven way. We also discovered that as the level of explanation increases, the user expectation of the system could be violated [75]. For example, observing how the system interpreted specific user behavior to determine stress levels led users to discover that the algorithm behaves differently from their perception of stress. Drawing upon reflections on our field study, we first propose high-level guidance on how the prediction algorithm should be utilized in the PI system to support self-reflection. Based on our findings, we discuss aspects that should be considered when applying explainability to an algorithm-incorporated PI system for self-reflection in the following section.

6.1 Prediction for Retrospection: Exploiting Algorithms to Facilitate Technology Mediated Reflection

Through the MindScope study, we investigated how the prediction algorithm can be exploited in the PI system for self-reflection. Based on the study result, we identified a design possibility of utilizing the prediction algorithm to support the user's retrospection process, which is different from the primary goal of the current prediction algorithm aimed at performing accurate classification and regression [45]. The core components of PI consist of data collection of personal information and reflection. [43]. Accordingly, we summarize below how an algorithmic stress management system such as MindScope can aid in data collection and reflection through PI systems for mental health [43].

6.1.1 How Can an Algorithm Help Data Collection? The data input methods of mood, affect, or other mental health-related topics in the existing PI system largely rely on the participants' self-report. However, this approach is likely to cause recall bias and accuracy reduction [35, 55]. To complement this method, EMA was proposed to help users input data through repeated evaluation in the context of an event [69]. However, this approach relies on the user's subjective evaluation and judgment. Therefore, it may be a difficult process for users who lack self-awareness. Technical approaches such as mood detection [80] and algorithmic sensor feedback [80] using passive sensing technology can address the memory burden issue. However, there are concerns about the adequacy of the automatic tracking method in collecting the user's subjective and personal emotional state [62]. For example, in accepting algorithmic results for personal information, people often uncritically accept the results of algorithms by overriding their own interpretations or understanding [84].

MindScope offers a stress data collection method that combines data-based stress prediction information and user-driven stress level input. Field study results showed that this data collection method could effectively recall and track past stress levels in a user-driven manner. When comparing the modeling phase where participants report stress using an EMA process and the prediction phase where participants report their stress level with the prediction information, participants in the prediction phase utilized data-driven, objective algorithmic decisions to supplement their subjective, abstract, and intuition-based judgments, resulting in a less vague recall of the past. Further, participants used the explanation displaying the data related to stress prediction to reconstruct their own past stress-related events and to help with more detailed recall. Through these findings, we identified the possibility that the predictive algorithm could be used as a device to help individuals reminisce and analyze past events.

6.1.2 How Can an Algorithm Help Data Reflection? Many existing data-driven PI systems serve as data visualization that is mainly used for data analytics. However, the limitation of this approach is that it requires data literacy of the users [25]. AI-infused products or services–which often automatically generate insights and provide recommendations based on the data collected–have been found to be valuable for general users. A recent study reported that the ML model built on personal data (e.g., meal, blood sugar level) was effectively used to suggest actionable health behaviors [25]. In line with the studies, MindScope acted as an experimental platform for examining the role of algorithms that can generate actionable insights for stress management.

The deployment study also helped us determine an appropriate level of explainability to facilitate self-reflection. We confirmed that providing an explanation was more useful and effective than providing only prediction value. Our participants reported that the most detailed explanation enabled them to understand the context of the current stress level and generate concrete, actionable plans to offset the stressors identified. Explanations providing only category level information were also rated positively because they provided an opportunity for speculation and inspection at a glance in a user-driven way. However, we also determined that as the level of explainability increased, users became more sensitive to the accuracy of the system Based on our findings, we discuss what should be considered when applying the explainability for algorithm incorporated PI system for self-reflection in the following section.

6.2 Explainability for Supporting Self-Reflection in the Algorithm-Assisted PI System

In this study, we explored the role of explainability in algorithm-mediated self-reflection, which differs from the general goal of explanation that increases the acceptance and trust of algorithmic output [3, 20]. Our study results confirmed that explainability could be helpful in self-reflection by helping users reconstruct past stressful situations and plan stress intervention based on objective data-driven insight. We also found that giving a more detailed explanation is not always the best approach, consistent with the previous study [75]. In this section, we discuss considerations when designing an explanation for self-reflection in the PI system based on our research results.

6.2.1 Stress Prediction Visualization for Promoting User Initiative and Behavior Change. We have confirmed that the level of explanation can affect a user's initiative in the self-reflection process. Compared to providing detailed behavior-level *Type 3* explanations, the *Type 2* data-category-level explanations helped users infer and recall past stressors more proactively. Conversely, this finding also suggests detailed explanations might reduce users' initiative in the process of looking back on the past and gaining insight through it. Therefore, the explainability of the algorithm output for self-reflection should be designed in consideration of how much the user will take the initiative in the self-reflection process.

Consistent with the previous study on algorithmic self-reflection systems [25, 28], users planned specific and actionable behavior changes using the algorithm output and explanations. Considering one of primary goals of self-reflection is to influence future behaviors and attitudes based on retrospection, we believe the explanation for the algorithmic output in a PI system should focus on information that can help users change their future behavior. For example, global feature importance and decision tree approximation are known to be effective XAI methodologies for explaining how a model behaves [44], but they could be too complex and abstract to plan specific behavioral changes for users. On the other hand, although not explored in our study, an explanation from the "how to be that" perspective can be effective in answering how a specific behavior or instance needs to change to obtain different stress prediction results.

6.2.2 Open-Ended Algorithmic Stress Prediction Visualization for Promoting Self Reflection. In this study, we found the possibility that the prediction algorithm and its explainability can be used to help users better recall and reflect on stress-related data. Meanwhile, we also found that the level of explainability should be carefully considered as it can negatively affect users' system reliability and overall perception [44]. Therefore, we suggest presenting predictive information in an open-ended way that promotes users' self-understanding and reflection, rather than decisively diagnosing the user's stress level. For example, the interaction between a

user and an algorithm can be designed to assist the user in understanding himself or herself in a cooperative way. We can imagine a system's framing in this way: "`Today, I (system) guess you are experiencing some stress. What do you think?`" This would ask users for their own retrospection rather than mindlessly accepting the system's output. We further identified that observing how the system interpreted specific user behavior into stress levels led users to discover that the algorithm behaves differently from their perception of stress. Accordingly, excessive explanation could reveal this disparity and lower trust in the system [75]. In particular, the reasoning process of identifying stressors and patterns can be fundamentally different between people and machines because of the highly subjective nature of mental health. Therefore, when providing explanation in stress prediction algorithms, the system needs to provide cues implying that people and machines infer stress differently to prevent a reduction in the perceived reliability of the system. Furthermore, explainability should be designed to allow users to examine their past events and status in a non-decisive manner, rather than providing technical explanations of the prediction results.

6.3 Design Suggestions for Algorithm-Assisted Self-Reflection in PI systems

We proposed in the above section the general implications of how the prediction algorithm and explainability should support the user's self-reflection. In the following section, we summarize the concrete suggestions of design elements that should be considered to improve the user experience of the PI System using the algorithm.

6.3.1 Adjusting the Level of Explanation According to Stages of Interaction. While most users preferred the most detailed explanation, providing such detailed explainability in the early stage of interaction, especially when an algorithm provided an inaccurate or insufficient explanation, might decrease trust in the system. Therefore, careful consideration is required to determine the appropriate level of explanation provided, because the higher the level of explanation, the more likely the model's performance will be revealed in detail. MindScope was also not free from early stage accuracy problem (i.e., the cold-start problem) in which the system could not draw any inference for users about which it has not yet gathered sufficient information, which in turn negatively affected the user experience. [10, 36]. To alleviate this deficiency, the system can adjust the levels of explainability corresponding to the amount of data collected and the accuracy of the model. Our study revealed that the level of explainability influenced participants' perception of the algorithm's accuracy. Therefore, we suggest interface designs that progressively improve explainability as the interaction with the system increases and the model is trained sufficiently.

6.3.2 Building Co-performing Relationships. MindScope can be positioned as an intelligent PI system augmented by a prediction algorithm built upon users' self-generated data including EMA and smartphone usage data. Recent studies on user-algorithm interaction in intelligent computing systems emphasize establishing a cooperative and reciprocal relationship, known as building a co-performing agent [37]. In our study, some participants often perceived MindScope as a health partner that keeps learning about them. MindScope's 10-day data collection and modeling period without providing algorithm-generated stress reports allowed users to build a relationship and trust the system. Accordingly, our participants actively engaged in data input to train their models for better predictions. We also found that an interactive, anthropomorphized persona applied to the system (e.g., a bubble changes its color corresponding to a user's stress level) further helped users perceive the system as more of a co-performing agent. For instance, our participants noted that they were tolerant of the errors the system made and motivated by input data required to make the MindScope's agent intelligent. We also found that our participants appreciated that they received an opaque stress level every three hours. They said that waiting for a stress report and anticipating the algorithm's prediction was a particularly pleasant experience. MindScope's unique feature—delivering information about one's stress at regular intervals—was a lever for building a positive relationship with the algorithm. Our results emphasize that a co-performing relationship between a user and an algorithmic system is critical for an intelligent PI system where an individual's participation in the entire process of data collection and reflection is required [43]. Providing a pleasant user experience through interaction with a co-performing agent can be particularly beneficial to the domain of PI systems that deal with mental health issues such as stress, anxiety, and depression.

6.4 Limitation

Our study presents limitations related to experimental design and system design. Participants were able to interact with each type of explanation for only each five days chunk which could be relatively a short period to examine the impacts and differences between each explanation type. A longitudinal study should be done to compensate for the limitation. We did not have a control group to clearly distinguish the effects of the prediction and explanation on stress reduction and management. In addition, the order in which the three different explanations were provided is not strictly counterbalanced. To complement our result, we conducted in-depth interviews with most of the participants (n=34) to reveal their lived experiences on stress management and perception of prediction algorithm. This study was conducted not to present clinical, quantitative evidence of MindScope on stress management but to explore new possibilities and gain insights into the design of the future PI system utilizing a prediction algorithm for self-reflection. Therefore, we circumvented complexities and kept the field study to a reasonable size to focus on providing users with a seamless system experience for eliciting user experiences close to the real world. Because we conducted this study with college students, our research results may be difficult to generalize to a broader population. For example, the level of background knowledge in AI and data that university students have may differ from that of users with other demographic backgrounds, which can affect their perception of algorithmic output [77]. Future work could be done to compensate for this limitation by enrolling diverse participants. Our system design has room for improvement. We acknowledge that some microtask methods could be problematic (e.g., inducing excessive YouTube watching). However, rather than limiting the type of microtasks at the researcher's discretion, we focused on providing an opportunity

for users to freely choose their interventions from a list or create their own interventions. We decided on this design because the most appropriate and effective intervention for each individual may differ. Also, planning their own interventions can provide a sense of agency in managing stress. Regarding the model's performance, we expect there will be room for improvement through additional data and extension of the experimental period. We acknowledge that model performance could affect the user experience and efficacy of the system. However, although highly subjective, stress matters, and it is influenced by various factors and is difficult to predict. In this work, rather than focusing on the technical contribution to improving the model's performance, our main objective was to understand how users and prediction algorithms interact and the ways the visualization and explainability of prediction information affect them. Lastly, because our study focuses on how different levels of explainability affect a user's self-reflection, providing users with specific navigating interactions techniques (e.g., zoom, filter, details on demand) in the predictive information goes beyond the scope of our study. Nevertheless, these techniques allow users to effectively find the content they want in complex data as needed [70] and also can bring a positive user experience in the provision of an explanation of the algorithmic output [76]. Therefore, we suggest future work that allows users to explore the level and content of an explanation based on their needs and goals.

7 CONCLUSION

This paper addressed the design opportunities for utilizing prediction algorithms and explainability in the PI system for supporting users' retrospection. A MindScope is an algorithm-assisted PI system designed to explore how people perceive and utilize the prediction algorithm for their reflection on stressors. Through a 25-day real-world deployment study, we provided empirical findings on the impact of algorithmic stress prediction for self-reflection. In particular, we report how the explainability of stress prediction affected users' self-reflection and algorithmic perception. Our findings indicate that prediction could be used as a device to help individuals reminisce on past stress levels and related information by supplementing the user's subjective reasoning. The detailed explanations were used to reconstruct and understand stress-related events, while categorical explanations support users in understanding stressors in a user-led way. Drawing upon reflections on our field study, we propose exploiting a prediction algorithm for supporting users' retrospection with open-ended algorithmic prediction that promotes the user's self-understanding and reflection, rather than decisively diagnosing the user's stress level.

ACKNOWLEDGMENTS

We thank our participants for providing valuable data and the reviewers for constructive feedback. This work was supported by National Research Foundation of Korea (NRF) grant by Korea government (the Ministry of Science and ICT): (No. 2017M3C4A7083533, 2020R1F1066408).

REFERENCES

[1] Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Voida, Geri Gay, Tanzeem Choudhury, and Stephen Voida. 2014. Towards Personal Stress Informatics: Comparing Minimally Invasive Techniques for Measuring Daily Stress in the Wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare* (Oldenburg, Germany) *(PervasiveHealth '14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 72–79. https://doi.org/10.4108/icst.pervasivehealth.2014.254959

[2] William Albert and Thomas Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[4] Gerhard Andersson and Pim Cuijpers. 2009. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cognitive behaviour therapy* 38, 4 (2009), 196–205.

[5] Jakob E. Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. 2012. The MONARCA Self-Assessment System: A Persuasive Personal Monitoring System for Bipolar Patients. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (Miami, Florida, USA) *(IHI '12)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2110363.2110370

[6] Jakob E Bardram and Aleksandar Matic. 2020. A decade of ubiquitous computing research in mental health. *IEEE Pervasive Computing* 19, 1 (2020), 62–72.

[7] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *ACM Trans. Comput.-Hum. Interact.* 20, 5, Article 30 (nov 2013), 27 pages. https://doi.org/10.1145/2503823

[8] Sofian Berrouiguet, David Ramírez, María Luisa Barrigón, Pablo Moreno-Muñoz, Rodrigo Carmona Camacho, Enrique Baca-García, and Antonio Artés-Rodríguez. 2018. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. *JMIR mHealth and uHealth* 6, 12 (2018), e197.

[9] A Bhattacharya, P Vasant, N Barsoum, C Andreeski, T Kolemisevska, Abdurrahman Talha Dinibütün, and Georgi M Dimirovski. 2006. Decision making in TOC-product-mix selection via fuzzy cost function optimization. *IFAC Proceedings Volumes* 39, 23 (2006), 51–56.

[10] JesúS Bobadilla, Fernando Ortega, Antonio Hernando, and Jesús Bernal. 2012. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems* 26 (2012), 225–238.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice* 92, 2 (2019), 277–297.

[13] Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. 2017. Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, 495.

[14] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[15] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A. Kientz. 2015. SleepTight: Low-Burden, Self-Monitoring Technology for Capturing and Reflecting on Sleep Behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 121–132. https://doi.org/10.1145/2750858.2804266

[16] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[17] Sheldon Cohen, Tom Kamarck, Robin Mermelstein, et al. 1994. Perceived stress scale. *Measuring stress: A guide for health and social scientists* 10, 2 (1994), 1–2.

[18] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.

[19] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. *Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle*. Association for Computing Machinery, New York, NY, USA, 1591–1602. https://doi.org/10.1145/3461778.3462131

[20] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. *Expanding Explainability: Towards Social Transparency in AI Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445188

[21] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961

[22] Elizabeth V. Eikey and Madhu C. Reddy. 2017. "It's Definitely Been a Journey": A Qualitative Study on How Women with Eating Disorders Use Weight Loss Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 642–654. https://doi.org/10.1145/3025453.3025591

[23] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 126 (Dec. 2020), 38 pages. https://doi.org/10.1145/3432231

[24] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A Lived Informatics Model of Personal Informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. https://doi.org/10.1145/2750858.2804250

[25] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, David J. Albers, Arlene M. Smaldone, and Lena Mamykina. 2021. From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 206, 17 pages. https://doi.org/10.1145/3411764.3445555

[26] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2015. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics* 20, 4 (2015), 1053–1060.

[27] Ben Green and Salomé Viljoen. 2020. Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 19–31. https://doi.org/10.1145/3351095.3372840

[28] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What does all this data mean for my future mood? Actionable analytics and targeted reflection for emotional well-being. *Human–Computer Interaction* 32, 5-6 (2017), 208–267.

[29] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 114 (Sept. 2018), 31 pages. https://doi.org/10.1145/3264924

[30] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[31] Bumsoo Kang, Chulhong Min, Wonjung Kim, Inseok Hwang, Chunjong Park, Seungchul Lee, Sung-Ju Lee, and Junehwa Song. 2017. Zaturi: We Put Together the 25th Hour for You. Create a Book for Your Baby. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1850–1863. https://doi.org/10.1145/2998181.2998186

[32] Anna Kantosalo and Sirpa Riihiaho. 2019. Quantifying co-creative writing experiences. *Digital Creativity* 30, 1 (2019), 23–38.

[33] Ronald C Kessler, G Paul Amminger, Sergio Aguilar-Gaxiola, Jordi Alonso, Sing Lee, and T Bedirhan Ustun. 2007. Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry* 20, 4 (2007), 359.

[34] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 593–602.

[35] John F Kihlstrom, Eric Eich, Deborah Sandbrand, and Betsy A Tobias. 1999. Emotion and memory: Implications for self-report. In *The science of self-report*. Psychology Press, 93–112.

[36] Da-jung Kim, Yeoreum Lee, Saeyoung Rho, and Youn-kyung Lim. 2016. *Design Opportunities in Three Stages of Relationship Development between Users and Self-Tracking Devices.* Association for Computing Machinery, New York, NY, USA, 699–703. https://doi.org/10.1145/2858036.2858148

[37] Da-jung Kim and Youn-kyung Lim. 2019. *Co-Performing Agent: Design for Building User-Agent Partnership in Learning and Adaptive Services.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300714

[38] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. *TimeAware: Leveraging Framing Effects to Enhance Personal Productivity.* Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/2858036.2858428

[39] Susanne Kirchner, Jessica Schroeder, James Fogarty, and Sean A. Munson. 2021. *"They Don't Always Think about That": Translational Needs in the Design of Personal Health Informatics Applications.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445587

[40] Rafal Kocielnik, Fabrizio Maria Maggi, and Natalia Sidorova. 2013. Enabling self-reflection with LifelogExplorer: Generating simple views from complex data. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 184–191. https://doi.org/10.4108/icst.pervasivehealth.2013.251934

[41] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[42] Kwangyoung Lee, Hyewon Cho, Kobiljon Toshnazarov, Nematjon Narziev, So Young Rhim, Kyungsik Han, YoungTae Noh, and Hwajung Hong. 2020. *Toward Future-Centric Personal Informatics: Expecting Stressful Events and Preparing Personalized Interventions in Stress Management.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376475

[43] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. *A Stage-Based Model of Personal Informatics Systems.* Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[44] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. *Questioning the AI: Informing Design Practices for Explainable AI User Experiences.* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[45] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[46] Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-Aware Applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 415–424. https://doi.org/10.1145/2030112.2030168

[47] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[48] Therese H Macan, Comila Shahani, Robert L Dipboye, and Amanda P Phillips. 1990. College students' time management: Correlations with academic performance and stress. *Journal of educational psychology* 82, 4 (1990), 760.

[49] Sumit Majumder and M Jamal Deen. 2019. Smartphone sensors for health monitoring and diagnosis. *Sensors* 19, 9 (2019), 2164.

[50] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. *AffectAura: An Intelligent System for Emotional Memory.* Association for Computing Machinery, New York, NY, USA, 849–858. https://doi.org/10.1145/2207676.2208525

[51] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards multimodal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct*. 1132–1138.

[52] Martijn Millecamp, Robin Haveneers, and Katrien Verbert. 2020. *Cogito Ergo Quid? The Effect of Cognitive Style in a Transparent Mobile Music Recommender System.* Association for Computing Machinery, New York, NY, USA, 323–327. https://doi.org/10.1145/3340631.3394871

[53] Ranjita Misra and Michelle McKean. 2000. College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American journal of Health studies* 16, 1 (2000), 41.

[54] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 75 (Sept. 2019), 21 pages. https://doi.org/10.1145/3351233

[55] Debbie S Moskowitz and Simon N Young. 2006. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience* 31, 1 (2006), 13.

[56] Amir Muaremi, Bert Arnrich, and Gerhard Tröster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

[57] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.

[58] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. *I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174223

[59] Fredrik Ohlin and Carl Magnus Olsson. 2015. Intelligent Computing in Personal Informatics: Key Design Considerations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 263–274.

https://doi.org/10.1145/2678025.2701378

[60] Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. 2012. Habits make smartphone use more pervasive. *Personal and Ubiquitous computing* 16, 1 (2012), 105–114.

[61] Google PAIR. 2019. *People + AI Guidebook*. Retrieved Sep 8, 2021 from https://pair.withgoogle.com/guidebook

[62] Amon Rapp and Federica Cena. 2014. Self-monitoring and technology: challenges and open issues in personal informatics. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 613–622.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[64] Jens Riegelsberger, M Angela Sasse, and John D McCarthy. 2005. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies* 62, 3 (2005), 381–422.

[65] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[66] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H. Epstein, Kenzie L. Preston, C. Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, Mustafa al'Absi, and Santosh Kumar. 2016. *Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data.* Association for Computing Machinery, New York, NY, USA, 4489–4501. https://doi.org/10.1145/2858036.2858218

[67] Jessica Schroeder, Chia-Fang Chung, Daniel A. Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A. Munson. 2018. Examining Self-Tracking by People with Migraine: Goals, Needs, and Opportunities in a Chronic Health Condition. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 135–148. https://doi.org/10.1145/3196709.3196738

[68] Jussi Seppälä, Ilaria De Vita, Timo Jämsä, Jouko Miettunen, Matti Isohanni, Katya Rubinstein, Yoram Feldman, Eva Grasa, Iluminada Corripio, Jesus Berdun, et al. 2019. Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: systematic review. *JMIR mental health* 6, 2 (2019), e9819.

[69] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

[70] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[71] Åsa Smedberg, Hélène Sandmark, and Andrea Manth. 2017. Online Stress Management for Self- and Group-Reflections on Stress Patterns. In *Biomedical Engineering Systems and Technologies*, Ana Fred and Hugo Gamboa (Eds.). Springer International Publishing, Cham, 387–404.

[72] Jaime Snyder, Mark Matthews, Jacqueline Chien, Pamara F. Chang, Emily Sun, Saeed Abdullah, and Geri Gay. 2015. MoodLight: Exploring Personal and Social Implications of Ambient Display of Biosensor Data. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 143–153. https://doi.org/10.1145/2675133.2675191

[73] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2018. Mood modeling: accuracy depends on active logging and reflection. *Personal and Ubiquitous Computing* 22, 4 (2018), 723–737.

[74] Aaron Springer and Steve Whittaker. 2018. Progressive disclosure: designing for effective transparency. *arXiv preprint arXiv:1811.02164* (2018).

[75] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322

[76] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 29 (Oct. 2020), 32 pages. https://doi.org/10.1145/3374218

[77] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[78] Anja Thieme, Jayne Wallace, Thomas D. Meyer, and Patrick Olivier. 2015. Designing for Mental Wellbeing: Towards a More Holistic Approach in the Treatment and Prevention of Mental Illness. In *Proceedings of the 2015 British HCI Conference* (Lincoln, Lincolnshire, United Kingdom) *(British HCI '15)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2783446.2783586

[79] Karmen Toros and Michael C LaSala. 2019. Child protection workers' understanding of the meaning and value of self-reflection in Estonia. *Reflective Practice* 20, 2 (2019), 266–278.

[80] Alina Trifan, Maryse Oliveira, and José Luís Oliveira. 2019. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR mHealth and uHealth* 7, 8 (2019), e12649.

[81] Greg Wadley, Frank Vetere, Liza Hopkins, Julie Green, and Lars Kulik. 2014. Exploring Ambient Technology for Connecting Hospitalised Children with School and Home. *Int. J. Hum.-Comput. Stud.* 72, 8 (aug 2014), 640–653. https://doi.org/10.1016/j.ijhcs.2014.04.003

[82] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[83] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

[84] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A. Smith. 2015. *Can an Algorithm Know the "Real You"? Understanding People's Reactions to Hyper-Personal Analytics Systems.* Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/2702123.2702274

[85] Paweł W. Woźniak, Przemysław Piotr Kucharski, Maartje M.A. de Graaf, and Jasmin Niess. 2020. *Exploring Understandable Algorithms to Suggest Fitness Tracker Goals That Foster Commitment.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3419249.3420131

[86] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[87] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. *Understanding the Effect of Accuracy on Trust in Machine Learning Models.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[88] Bin Zhu, Anders Hedman, and Haibo Li. 2016. Design Digital Mindfulness for Personal Wellbeing. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction* (Launceston, Tasmania, Australia) *(OzCHI '16)*. Association for Computing Machinery, New York, NY, USA, 626–627. https://doi.org/10.1145/3010915.3011841