Japanese Psychological Research THE JAPANESE PSYCHOLOGICAL ASSOCIATION

Japanese Psychological Research
2025, Volume 67, No. 2, 195–207
Special issue: Innovations in Clinical Psychological Science in the Era of Complexity

doi: 10.1111/jpr.12565

Exploring the Utility of a Machine Learning Approach with Mobile-Based Cognitive Function Tasks for Detecting Depression

MOMOKA TAKESHIGE, TAIKI OKA[†], MAI OHWAN and KEI HIRAI[*] *Osaka University*

Abstract: Self-report questionnaires, used for detecting major depressive disorder (MDD) in daily life, may incur biases stemming from social desirability and repetitive answers. Though detection based on mobile sensing was being developed recently, it cannot sufficiently promote self-help action due to the characteristics of passive feedback. Thus, an active self-monitoring and feedback system is crucial for individuals to recognize and address their malfunctions. In this study, we proposed to predict changes in MDD severity using cognitive tasks monitored on mobile devices. An online survey was conducted to evaluate the severity, incorporating cognitive tasks such as Navon task, Go/No-go task, and n-back task, along with the Quick Inventory of Depressive Symptomatology. Participants completed the survey three times on their mobile devices. The analysis included data from 75 participants, including 21 participants whose MDD score increased by at least one point during the second and third surveys; the first survey was excluded to avoid confounding effects. A random forest classifier was employed for classifying participants whose depression has and has not worsened. The learned model achieved modest accuracy (68.3%) with a significant mean area under the curve of 0.59 ($t(9) = 2.98$, $p = .016$, $d_z = 0.94$), suggesting the potential to predict depressive states based on cognitive domains. Moreover, working memory and attentional inhibition functions contributed to predicting the severity change mostly. Though improvements are required to reduce false negatives for practical applications, our result suggests that MDD aggravation could be assessed by mobile cognitive tasks.

Key words: major depressive disorder, detection, machine learning.

Depressive disorder remains a major problem worldwide that impacts individual well-being and socioeconomic situations. Considering the prognosis after treatment, depressive disorder needs to be detected more accurately and earlier before it becomes more severe (Cacheda et al., 2019). At present, several self-reported questionnaires have been used to detect depression (Lamoureux et al., 2010; Radloff, 1977). However, these methods have limitations, as subjective complaints are prone to overestimation or underestimation compared to objectively measured symptoms (Pavlova & Uher, 2020). In addition, the variability resulting from repeated self-measurement is significant (Long et al., 2020), so it is considered unsuitable for repeated daily assessment. Also, except for occasions such as annual health

*Correspondence concerning this article should be sent to: Kei Hirai, Graduate School of Human Sciences, and Graduate School of Medicine, Osaka University, Yamadaoka, Suita, Osaka 565-0871, Japan. (E-mail: khirai@grappo.jp)

[†]Additional affiliation: Advanced Telecommunications Research Institute International.

checkups, the questionnaires are often completed only after the awareness of the depressive symptoms, which is not suitable for early detection. On the other hand, behavioral changes (Seppälä et al., 2019), physiological changes (Moretta & Benvenuti, 2022), or cognitive function decline (LeMoult & Gotlib, 2019; X. Wang et al., 2020) caused by depression can be used to evaluate depression objectively. It is argued that the explicit nature of self-reports makes them susceptible to various biases (Hunt et al., 2003); thus, integrating data-driven tools and advanced data analysis techniques, such as behavioral and neurophysiological assessments, would be beneficial for enhancing the accuracy of psychiatric diagnostic processes (Richter, Fishbain, Fruchter, et al., 2021). By acknowledging the potential biases associated with subjective measures like self-reported assessments, it is possible to enhance overall predictive accuracy by complementing subjective evaluations with objective assessments. This study recognizes the important role that subjective scales play in understanding an individual's condition, while also proposing that integrating objective assessments can help mitigate the biases inherent in repeated subjective evaluations and can provide a complementary role.

There have been many attempts to detect such changes using devices such as smartphones. Seppälä et al. (2019) reported the effectiveness of mobile devices in detecting mental illness through the meta-analysis of studies investigating the relationship between depression and either physiological or behavioral data monitored from smartphone usage or wearable sensors. However, merely detecting depression through behavioral logs or sensors and providing feedback is insufficient to promote self-help behavior in individuals because these results are passive for people (Bauer et al., 2020; BinDhim et al., 2015). Active self-monitoring by individuals is crucial for seeing a doctor and stress-management behaviors (Bos et al., 2015; van Os et al., 2017). Therefore, it is necessary to create a system that provides feedback to individuals to detect their malfunctions for prevention, not only depending on passive data.

The question arises: Is there any objective indicator that can make us aware of such a decline? As mentioned, the decline in cognitive function serves as a suitable index. Depression symptoms are strongly associated with cognitive impairment led by dysfunction of the brain: attention, executive function, memory, processing speed, poor concentration, and impaired attention (Ahern & Semkovska, 2017; Henriques & Davidson, 1997; Semkovska et al., 2019). For example, Richter et al. (2020) and Richter, Fishbain, Fruchter, et al. (2021) attempted to predict depression using cognitive tasks and built a highly accurate model. However, this study is only laboratory-based and did not capture longitudinal changes in real-life situations. For daily depression screening, tools must be less burdensome and usable regardless of location or time. Although Cormack et al. (2019) proposed measuring cognitive tasks and depressed mood in daily life via smartwatch applications to support patient treatment and remediation, no paper has yet attempted depression detection using this method.

As explained thus far, depressive symptoms are difficult to recognize and often go untreated until they become severe. Our proposed method aims to detect depression at an early stage and promote self-help behaviors. While many neuropsychological tests directly measure cognitive function, they are typically conducted in medical or laboratory settings, are time-consuming, and are rarely used for daily measurements. Therefore, if it becomes possible to measure mild cognitive decline daily, like a thermometer, it could serve as a tool to quickly become aware of one's mental state. For this purpose, making predictions from data collected in an environment similar to a real-life situation would be valuable. The goal of this study is to create a predictor of changes in depressive symptoms using daily-measured cognitive function data. This approach involves measuring the decline in cognitive function that leads to depressive symptoms based on changes in brain condition. Instead of examining the state at a single point, a series of three surveys was conducted, and the values from the second survey onwards and their difference values were collected to make more dynamic predictions. Additionally, to facilitate the practical application of cognitive tasks in predicting depressive states in real-world settings, it is essential for the model to be robust against "noisy" data. Therefore, the goal should be developing a predictive model capable of effectively predicting depressive states without the need for a controlled environment.

Utilizing machine learning with a random forest classifier, we constructed a model composed of three cognitive functions to detect depression. The outcome of this study may help develop screening and preventive tools for depression in daily life. The present study attempted to create a predictive model for depression exacerbation using cognitive functions related to impaired function due to mental illness through three simple cognitive tasks via smartphones and tablets. The Navon, Go/No-go, and n-back tasks were employed for the cognitive tasks. These tasks were selected because the cognitive functions that can be assessed using them are closely related to MDD: the Navon task to assess the tendency to prioritize processing at the global level (de Fockert & Cooper, 2014), the Go/No-go task to measure attentional inhibition (Kaiser et al., 2003), and the n-back task to evaluate working memory (Nikolin et al., 2021; Rose & Ebmeier, 2006). In order to provide a practical tool for the early detection of depression in the real world, we tested whether short-term measures of cognitive function can effectively predict long-term depressive tendencies.

Method

Participants

One hundred and twenty-one people participated in our survey through Crowdworks (https://crowdworks.jp/). In the three surveys, 127 participants completed the first survey, 125 completed the second survey, and 121 completed the third survey. We decided on the sample size based on the previous research (Richter, Fishbain, Fruchter, et al., 2021; Richter, Fishbain, Richter-Levin, & Okon-Singer, 2021), following the guideline of sample size determination (Lakens, 2022). We primarily recruited workers because mental health issues are most prevalent in the working-age population (Ministry of Health, Labour, and Welfare of Japan, 2017). The collected data were excluded based on the following predetermined criteria detailed below. The number of valid samples was 75, including 36 males and 39 females, aged between 22 and 60 years ($M = 40.41$ years, $SD = 8.91$ years). Exclusions were as follows: two participants due to data mismatch caused by code miss-entry, four participants with the same IP address as another participant, one participant who responded with "did not understand" to any of the cognitive tasks, three participants who incorrectly answered a control question, and 36 participants with at least one outlier in each cognitive task. The data-reduction process is detailed in the Supplementary Methods in Data S1.

Procedure

The study consisted of three surveys, and participants were required to complete all three. Only in the first survey did participants select their age, sex, and employment status at the beginning. Each survey involved performing four cognitive tasks in random order, followed by the Quick Inventory of Depressive Symptomatology, Japanese version (QIDS-J). The tasks included multiple objects tracking (MOT), Navon task, Go/No-go task, and n-back task. The entire survey took approximately 50 min. Data acquisition occurred during three periods as follows: the first survey from February 14 to 16, 2022; the second survey from February 17 to 18; and the third survey from March 14 to 15. Details on equipment are provided in Supplementary Methods in Data S1.

Measurement

Questionnaire. The QIDS-J is a 16-item self-report rating scale assessing the severity of depression. It also corresponds to the diagnostic criteria for major depressive disorder

(American Psychiatric Association, 2000). The internal validity of the QIDS-J has been demonstrated by a Cronbach's alpha coefficient of .86 (Rush et al., 2003). In our sample, the Cronbach's alphas were .89 in the second survey and .87 in the third survey. We calculated the change in QIDS-J scores from the second to the third survey and used a cutoff point of an increase of one point. In this study, we adopted a smaller cutoff value than the clinical threshold (e.g., 3.5; McIntyre et al., 2021) to achieve the goal of predicting the risk of worsening depression in advance.

Cognitive tasks. Cognitive tasks were constructed using PsychoPy (Ver. 2020.2) and administered via Pavlovia (https://pavlovia.org/) from the participant's device. Too quick responses of less than 100 ms were treated as false responses. The detailed information on each task is shown in Supplementary Methods in Data S1.

1. In the Navon task, eight Navon stimuli were created using the target stimuli alphabet letters H and F and the other alphabet letters L and T (Figure S1). Participants were instructed to press the "H" button if the stimulus made up an H or if it consisted of small Hs, and the same rule was also applied to "F." Sixty-four trials (eight by eight stimuli) were conducted in random order (Figure 1A).

2. In the Go/No-go task, there were target alphabet letters F and J, and non-target P and T, and each letter was presented every 1,500 ms in random order. Participants were only asked to press the reaction button when they saw the target. The button was active for 1,000 ms after the stimulus was displayed. There were 200 trials, and the four alphabet letters appeared at equal rates (Figure 1B).

3. The n-back task consisted of three sets: $n = 0, 1, 2$. In one set, 25 numbers from 0 to 9 were displayed one by one with 2,000-ms intervals. Participants needed to press the reaction button in the following cases: in $n = 0$, when "3" was presented, and in $n = 1, 2$ if the number was the same as that presented one or two before it. In each condition, five out of 25 numbers were targets (Figure 1C).

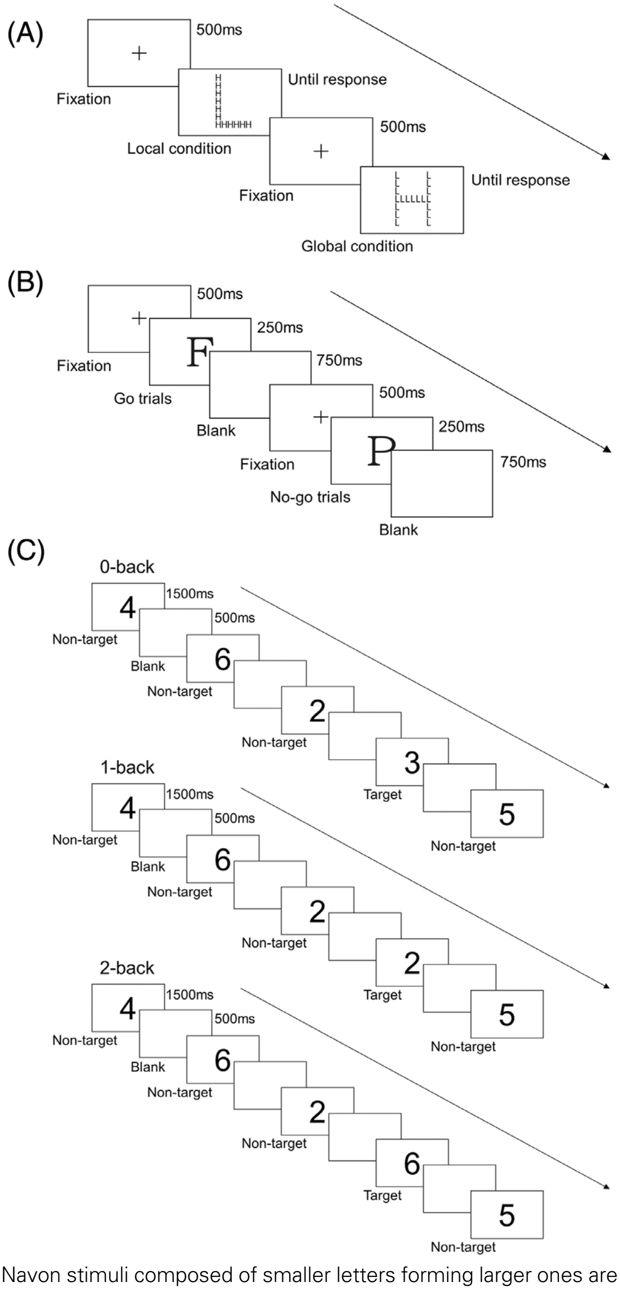
Several studies have shown that data from online cognitive tasks can yield performance and effect sizes comparable to those from laboratory experiments (Crump et al., 2013; Sauter et al., 2022). At the same time, it has been pointed out that it is difficult to control the experimental situation in online experiments, and there is a risk that the variance of the data may increase and outliers may be easily observed (Gagné & Franzen, 2023). Since the goal of this study was to assess depression through cognitive tasks in daily life, we did not restrict the surrounding environment in which participants took part in the study. We set explicit limits on the devices used by the participants (i.e., smartphones or tablets, see Data S1 for details).

In addition, we emphatically instructed the participants to keep the screen in portrait orientation during the task and not to interrupt other tasks in the middle of the task. We checked through practice whether the participants understood the instructions. We conducted rigorous data screening before the analysis (see Data S1 for details).

Analysis

Calculation of cognitive task scores. In the Navon task, the positive reaction time in the global condition was subtracted from the positive reaction time in the local condition to obtain an index of attentional focus. The smaller this index of attentional focus is, the narrower the attentional range becomes. In the Go/No-go task, commission error on No-go trials, omission error, and reaction time on the Go trials were calculated for all trials. In the 0-, 1-, and 2-back tasks, the number of hits (the number of trials in which the participant responded correctly to the target stimulus ≤5) and the average reaction time in hit trials were computed, respectively. The performance and standard deviation in the third survey as well as the difference between the second and third

Figure 1
Paradigms of cognitive tasks.



Note. (A) Navon task. Navon stimuli composed of smaller letters forming larger ones are presented on the screen. Participants identified letters ("H" or "F") within global letters or smaller components. (B) Go/No-go task. The alphabet stimuli are sequentially presented at 1-s intervals in a random order. Participants respond to "F" or "J" quickly by touching a button on the screen; they avoid responding to other letters. (C) N-back task. The single-digit stimuli are sequentially presented at 2-s intervals. Participants detect the number "3" (0-back), which matches the one before (1-back) or two before (2-back) in a sequence of presented numbers.

performance were used as features for each indicator.

Statistical analysis. Instead of using data from a single survey, changes between two surveys were analyzed. Data from the second and third surveys were analyzed to account for the training effect of the cognitive task and to utilize the performance and their variability as features. The analysis was performed using scikit-learn (Ver. 1.1.2) in Python (Ver. 3.8.5). Random Forest (Breiman, 2001) was used to predict the change in the depressive score. First, the data from 75 participants were randomly split into training sets (60% of the data) and test sets (40% of the data) with no overlapping. Parameter tuning and training were performed on the training data (45 participants). To identify the optimal hyperparameters, we conducted a grid search with cross-validation to tune the maximum depth of trees, the number of trees in the forest, and the number of features to contemplate when seeking the best division. Then, a random forest classification model was trained, given the selected parameters. The trained model was fitted to the test data (30 participants). The diagnostic results from the questionnaires were considered the ground truth, and the accuracy of the predicted diagnoses from the model was calculated. This process was repeated 10 times with different data splits. To evaluate model accuracy, we calculated the area under the receiver–operator curve (AUC) for a binary classifier. The AUC measures the model's ability to distinguish between classes, with values ranging from 0.5 (random) to 1 (perfect). This metric is particularly useful for imbalanced datasets.

Ethics

We obtained the approval of the Research Ethics Review Committee of the Division of Education, the Graduate School of Human Sciences, Osaka University (23017R). Informed consent was obtained from all participants through the recruitment page on Crowdworks.

Table 1
Mean and standard deviations of demographic information for each group

Group	Depressive ($N = 21$)	Non-depressive ($N = 54$)
Sex	10 female (47.6%)	29 female (53.7%)
Age (years)	41.3 [±10.1]	40.1 [±8.5]
1st QIDS-J score	6.1 [±3.8]	6.1 [±5.1]
2nd QIDS-J score	4.9 [±4.0]	5.8 [±5.3]
3rd QIDS-J score	7.1 [±4.0]	4.5 [±4.9]
Difference value of QIDS-J score	2.2 [±1.8]	−1.3 [±1.7]

Note. QIDS-J = Quick Inventory of Depressive Symptomatology.

Results

Demographic Results

Tables 1 and 2 provide sociodemographic and clinical scores by group, and Figure 2 shows the trajectory of QIDS-J scores across three surveys. In the first survey's QIDS-J scores, there was no significant difference based on gender ($t(72.8) = 0.14$, $p = .890$, $d_z = 0.03$). Similarly, no significant differences were observed based on age groups ($F(4, 71) = 1.43$, $p = .233$, $\eta_p^2 = .08$) or employment status ($F(3, 71) = 1.48$, $p = .228$, $\eta_p^2 = .06$).

The Results of Cognitive Measures

Table 3 illustrates the average data values collected during the initial survey of representative cognitive task indices. No significant differences were observed between individuals with and without worsening severity of depressive symptoms in any of the examined variables ($ps > .10$).

The Performance of the Random Forest Model

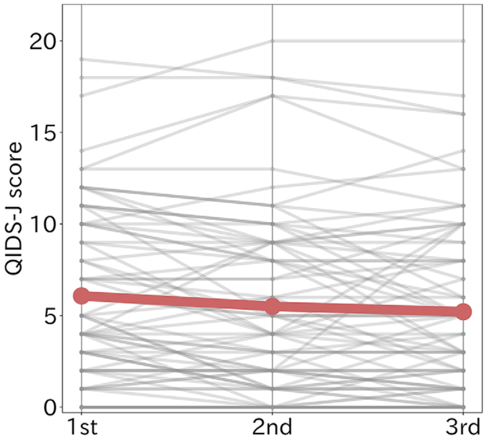
Among the 75 analyzed participants, 21 had an exacerbation in QIDS-J scores of one or more points over the approximately 1-month survey interval. The random forest classifier algorithm demonstrated an accuracy of 68.3%. The AUC across iterations was significantly higher than

Table 2
Mean and standard deviations (SDs) of demographic and clinical scores by group

	Sample size	Ratio	1st QIDS-J average [±*SD*]	2nd QIDS-J average [±*SD*]	3rd QIDS-J average [±*SD*]	Difference value of QIDS-J score [±*SD*]
Total						
All	75	100.00%	6.1 [± 4.7]	6.1 [± 5.0]	5.2 [± 4.8]	−0.3 [± 2.3]
Age (years)						
20–29	10	13.30%	8.0 [± 3.8]	7.4 [± 4.0]	7.3 [± 4.5]	−0.1 [± 3.3]
30–39	22	29.30%	6.4 [± 4.5]	5.2 [± 4.4]	4.6 [± 4.2]	−0.6 [± 2.2]
40–49	29	38.70%	6.4 [± 5.3]	6.1 [± 5.8]	5.6 [± 5.5]	−0.5 [± 1.8]
50–60	14	18.70%	3.6 [± 3.8]	3.4 [± 4.3]	3.9 [± 4.2]	0.5 [± 2.7]
Sex						
Male	36	48.00%	6.0 [± 4.7]	5.1 [± 4.9]	5.0 [± 4.7]	−0.1 [± 2.4]
Female	39	52.00%	6.2 [± 4.8]	5.9 [± 5.1]	5.4 [± 4.9]	−0.5 [± 2.2]
Employment status						
Employee	19	25.30%	5.4 [± 5.4]	5.4 [± 5.9]	5.1 [± 5.4]	−0.3 [± 2.8]
Part-time job	15	20.00%	6.6 [± 4.0]	5.1 [± 4.1]	5.0 [± 3.4]	−0.1 [± 2.0]
Self-employment/ Freelance	33	44.00%	5.5 [± 3.9]	4.5 [± 3.5]	4.2 [± 3.7]	−0.4 [± 2.4]
Other	8	10.70%	9.1 [± 7.0]	10.6 [± 6.9]	10.1 [± 7.0]	−0.5 [± 1.7]

Note. QIDS-J = Quick Inventory of Depressive Symptomatology.

Figure 2
Transition of Quick Inventory of Depressive Symptomatology (QIDS-J) scores.



Note. The red line represents the average score. The gray lines represent individual participants' data.

the chance level (mean AUC = 0.59, $t(9) = 2.98$, $p = .016$, $d_z = 0.94$, Figure 3). We also constructed another model including the baseline depression severity and sociodemographic variables, but the performance did not improve (see Data S1).

Additionally, our analysis revealed the marginal contribution of each behavioral measure derived from the input (Figure 4). The most contributing feature was the standard deviation (*SD*) of reaction times in the 2-back task, which measures the working memory. This was followed by mean reaction time in the 2-back task, the local condition of the Navon task, and the 1-back task. The features of importance differed depending on the model, and the results needed to be more consistent.

Discussion

In the present study, we developed a predictive model to detect an increase in the QIDS-J score, an indicator of MDD status, based on participants' aggregated performances and changes in several cognitive tasks. This study revealed a substantial number of false negatives, so it did not achieve high accuracy. According to Šimundić (2009), an AUC of 0.59 falls just

Table 3
The first survey's mean and standard deviations (SDs) of cognitive scores

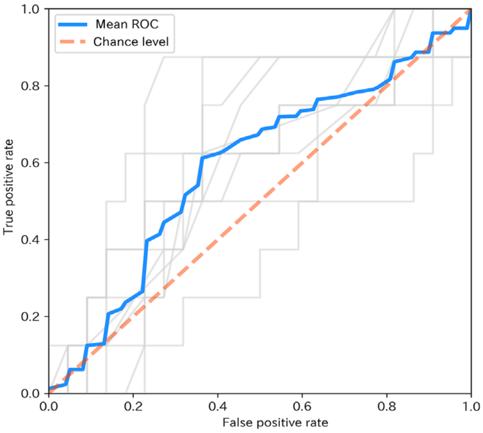
				Depressive		Non-depressive		*t*-value	*p*-value
N				21		54			
QIDS-J				6.1	[±3.7]	6.1	[±5.1]	0.02	1
Navon	In Global condition	RT		1040.4	[±185.4]	1043.3	[±297.9]		
Navon	In Local condition	RT		1101.5	[±158.3]	1039.6	[±228.5]	0.05	1
Go/No-go		FA count	*Max = 100	3.9	[±2.6]	4.2	[±4.7]	0.39	1
Go/No-go		RT		488.4	[±36]	494.6	[±46.0]	0.61	1
n-back	(0-back)	FA count	*Max = 20	0.2	[±0.6]	0.5	[±2.5]	0.72	1
n-back	(0-back)	RT		569.3	[±156.5]	560.3	[±168.9]	0.56	1
n-back	(1-back)	FA count	*Max = 20	0.1	[±0.4]	0	[±0.1]	0.79	1
n-back	(1-back)	RT		630.9	[±101.8]	611	[±127.1]	0.69	1
n-back	(2-back)	FA count	*Max = 20	0.4	[±0.6]	0.4	[±0.6]	0.42	1
n-back	(2-back)	RT		743	[±212.9]	796.3	[±163.6]	0.42	1

Note. The *p*-values were corrected using the Holm method. QIDS-J = Quick Inventory of Depressive Symptomatology; RT = reaction time; FA = false alarm.

below the "sufficient" range, which suggests that the diagnostic performance is relatively low. Furthermore, our findings did not achieve the accuracy reported by Richter et al. (2020) and Richter, Fishbain, Fruchter, et al. (2021) in laboratory experiments. However, it is also true that our results (AUC = 0.59) surpassed chance levels. Possibly, with future revisions, there remains the potential to predict depression to some extent even from "noisy" data collected in real-world mobile-based settings.

Compared with previous studies (Richter et al., 2020; Richter, Fishbain, Fruchter, et al., 2021), which primarily consist of laboratory experiments with a predominantly young participant demographic, although not attaining an equivalent level of performance, this study provided suggestive evidence for the potential feasibility of prediction using data derived from real-world situations. In addition, our study has confronted the issue of data imbalance. A potential future direction could involve targeted sampling of individuals diagnosed with depression. Furthermore, as the foremost issue, the current algorithms must be revised to

Figure 3
Receiver–operator curve (ROC).



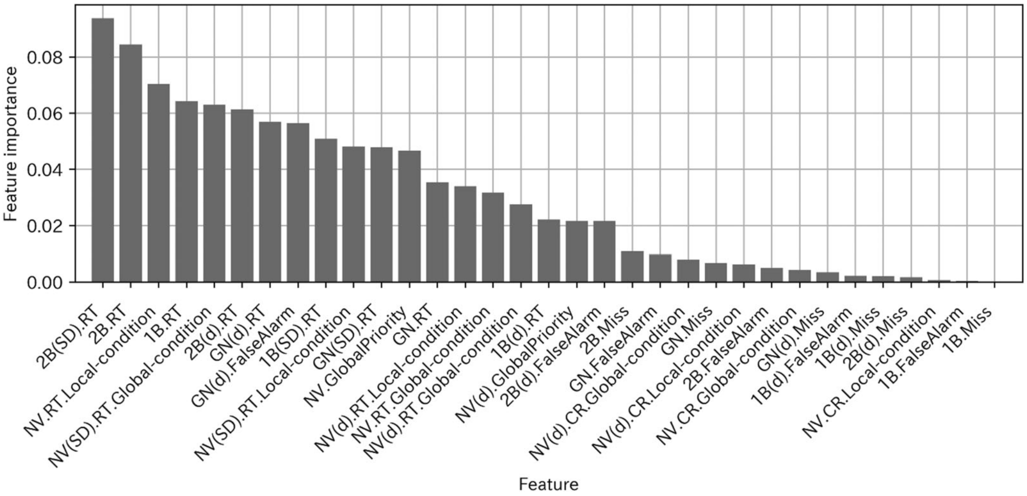
Note. The blue line represents the ROC plotted based on the average estimated class probabilities across 10 modeling iterations. The red dashed line serves as a reference point for the performance of a classifier that classifies the condition entirely at random.

identify true positives correctly. One contributing factor to the limited efficacy of true positive detection in this investigation is the deliberate decision not to customize predictions to individual idiosyncrasies. To achieve pragmatic predictive efficacy, it is advisable to formulate a predictive model tailored to each individual's characteristics (Aalbers et al., 2023). Given the inherent diversity in cognitive task performance across individuals, establishing a universal classification threshold presents challenges of considerable magnitude. Moreover, in other studies, classification modeling utilizing behavioral logs has also been conducted (R. Wang et al., 2018), and by combining these approaches, it may be possible to enhance the detection performance of true positives.

Measures from the 2-back task and the Navon task contributed to the classification prediction in this model. This finding suggests that working memory and attentional focus impacted mental state fluctuations. Unexpectedly, the differential values of the cognitive indicators contributed relatively little. Regarding this matter, the assessment of 2-point variations may not have fully captured the fluctuations in cognitive function. In the future, it will be imperative to accumulate cognitive function data longitudinally from the same cohort of participants to enhance the precision of cognitive decline extraction. Furthermore, the feature importance ranking needed to be more consistent, depending on the bias due to splitting the training and testing data. In light of these results, it can be speculated that a diverse range of cognitive measures are involved in the prediction, and it is desirable to develop a comprehensive tool that can measure a variety of cognitive functions to predict MDD severity more easily.

To enhance predictive performance in forthcoming endeavors, it is essential to address real-world data noise, which necessitates adaptable preprocessing and robust data infrastructure. Given device-related impacts on cognitive tasks, developing data-efficient applications for stable collection is anticipated. Despite the challenges associated with device variability, the ultimate goal of this study is to ensure predictability across devices.

Figure 4
Mean marginal contribution of each behavioral measure.



Note. The importance values were derived from the Gini impurity reduction criterion employed during construction of decision trees within the random forest ensemble. The Gini impurity is a measure of the degree of disorder in a set of data points, where lower values indicate a purer distribution of classes. The Gini impurity reduction resulting from each feature's use in decision splits is accumulated during the tree construction process. The greater the reduction, the more influential the feature is. NV = Navon task; GN = Go/No-go task; 1B = 1-back task; 2B = 2-back task; CR = correct rate; RT = reaction time; d = difference value between the second and third performances.

Considering participant fatigue, task refinement for shorter assessments is required, along with the development of composite tasks assessing diverse cognitive functions, possibly integrating gamification (Jamaludin et al., 2021).

Study Limitations and Future Directions

This study has several limitations. The first point is the absence of a clear gold standard for depression diagnosis (Richter, Fishbain, Richter-Levin, & Okon-Singer, 2021). This signifies that self-report bias is inherent in questionnaire-based assessments. In this study, we constructed a predictive model utilizing data derived from the depression questionnaire. For subsequent model validation, it is advisable to incorporate datasets consisting of individuals diagnosed through clinical interviews. The second point concerns the influence of participant fatigue resulting from survey engagement. To mitigate such effects, the strategic integration of technologies like gamification into the assessment methodologies assumes paramount importance, offering not only conciseness but also an engaging experience that counters the impact of fatigue. Third, we did not assess individual characteristics and lifestyle habits, which might be considered confounding factors for cognitive dysfunctions. However, the previous evidence suggests that cognitive dysfunction may have a specific impact on depression, even considering other factors (Richter, Fishbain, Richter-Levin, & Okon-Singer, 2021). Therefore, we think that we can ignore this problem here to some extent.

Conclusions

In this study, we developed a predictive model for detecting an increase in depression scores using real-world mobile-based data and found the key predictors might include working memory and attentional focus. However, the model had a high rate of false negatives and an AUC of 0.59, which was not superior to a previous laboratory-centric study. This finding highlights the importance of developing more effective mobile tools and constructing more practical models for predicting MDD severity change based on cognitive dysfunctions in real life.

Conflict of Interest

The corresponding author received remuneration as a board member for conducting business using behavioral economics from CoBe-Tech, Ltd. This author also holds 22% of the shares as the founder of a business using behavioral economics.

Acknowledgments

The authors are grateful to our laboratory staff for their support. We would also like to thank all the participants who spared their time and effort for this study.

Funding

This study was funded by Daikin Industries, Ltd. The funding sources had no role in the design and conduct of the study.

Author Contributions

M.T. and T.O. contributed equally to this work. T.O. and K.H. made substantial contributions to the study conception and design. M.T., T.O., and M.O. contributed to data acquisition. M.T., T.O., and M.O. conducted statistical analyses. T.O., K.H., M.T., and M.O. contributed substantially to data interpretation. M.T. and T.O. drafted the first version of the manuscript. All authors contributed to critical revisions and approved the final version of the manuscript. K.H. assumes responsibility for the integrity of the work.

Declaration of AI Use

We have used AI-assisted technologies (DeepL, Grammarly) to improve the readability and language of the work.

Data Availability Statement

All statistical data that support the findings of this study are available in this paper. Unfortunately, owing to company cohort data-sharing restrictions, individual-level data cannot be publicly posted. Data are, however, available from the authors upon reasonable request and with permission of Daikin Corporation.

References

Aalbers, G., Hendrickson, A. T., Vanden Abeele, M. M., & Keijsers, L. (2023). Smartphone-tracked digital markers of momentary subjective stress in college students: Idiographic machine learning analysis. *JMIR mHealth and uHealth*, *11*, e37469. https://doi.org/10.2196/37469

Ahern, E., & Semkovska, M. (2017). Cognitive functioning in the first-episode of major depressive disorder: A systematic review and meta-analysis. *Neuropsychology*, *31*(1), 52–72. https://doi.org/10.1037/neu0000319

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). American Psychiatric Association.

Bauer, M., Glenn, T., Geddes, J., Gitlin, M., Grof, P., Kessing, L. V., … Whybrow, P. C. (2020). Smartphones in mental health: A critical review of background issues, current status and future concerns. *International Journal of Bipolar Disorders*, *8*(1), 2. https://doi.org/10.1186/s40345-019-0164-x

BinDhim, N. F., Shaman, A. M., Trevena, L., Basyouni, M. H., Pont, L. G., & Alhawassi, T. M. (2015). Depression screening via a smartphone app: Cross-country user characteristics and feasibility. *Journal of the American Medical Informatics Association*, *22*(1), 29–34. https://doi.org/10.1136/amiajnl-2014-002840

Bos, F. M., Schoevers, R. A., & aan het Rot, M. (2015). Experience sampling and ecological momentary assessment studies in psychopharmacology: A systematic review. *European Neuropsychopharmacology*, *25*(11), 1853–1864. https://doi.org/10.1016/j.euroneuro.2015.08.008

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

Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, *21*(6), e12554. https://doi.org/10.2196/12554

Cormack, F., McCue, M., Taptiklis, N., Skirrow, C., Glazer, E., Panagopoulos, E., … Barnett, J. H. (2019). Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: Longitudinal observational study. *JMIR Mental Health*, *6*(11), e12814. https://doi.org/10.2196/12814

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

de Fockert, J. W., & Cooper, A. (2014). Higher levels of depression are associated with reduced global bias in visual processing. *Cognition & Emotion*, *28*(3), 541–549. https://doi.org/10.1080/02699931.2013.839939

Gagné, N., & Franzen, L. (2023). How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience. *Swiss Psychology Open*, *3*(1), 1. https://doi.org/10.5334/spo.34

Henriques, J. B., & Davidson, R. J. (1997). Brain electrical asymmetries during cognitive task performance in depressed and nondepressed subjects. *Biological Psychiatry*, *42*(11), 1039–1050. https://doi.org/10.1016/s0006-3223(97)00156-x

Hunt, M., Auriemma, J., & Cashaw, A. C. A. (2003). Self-report bias and underreporting of depression on the BDI-II. *Journal of Personality Assessment*, *80*(1), 26–30. https://doi.org/10.1207/s15327752jpa8001_10

Jamaludin, N. F., Tengku Wook, T. S. M., Mat Noor, S. F., & Qamar, F. (2021). Gamification design elements to enhance adolescent motivation in diagnosing depression. *International Journal of Interactive Mobile Technologies*, *15*(10), 154. https://doi.org/10.3991/ijim.v15i10.21137

Kaiser, S., Unger, J., Kiefer, M., Markela, J., Mundt, C., & Weisbrod, M. (2003). Executive control deficit in depression: Event-related potentials in a go/Nogo task. *Psychiatry Research*, *122*(3), 169–184. https://doi.org/10.1016/s0925-4927(03)00004-0

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lamoureux, B. E., Linardatos, E., Fresco, D. M., Bartko, D., Logue, E., & Milo, L. (2010). Using the QIDS-SR16 to identify major depressive disorder in primary care medical patients. *Behavior Therapy*, *41*(3), 423–431. https://doi.org/10.1016/j.beth.2009.12.002

LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical Psychology Review*, *69*, 51–66. https://doi.org/10.1016/j.cpr.2018.06.008

Long, E. E., Haraden, D. A., Young, J. F., & Hankin, B. L. (2020). Longitudinal patterning of depression repeatedly assessed across time among youth: Different trajectories in self-report questionnaires and diagnostic interviews. *Psychological Assessment*, *32*(9), 872–882.

McIntyre, R. S., Lipsitz, O., Lui, L. M. W., Rodrigues, N. B., Gill, H., Nasri, F., … Rosenblat, J. D. (2021). The meaningful change threshold as measured by the 16-item Quick Inventory of Depressive Symptomatology in adults with treatment-resistant major depressive and bipolar disorder receiving intravenous ketamine. *Journal of Affective Disorders*, *294*, 592–596. https://doi.org/10.1016/j.jad.2021.07.035

Ministry of Health, Labour, and Welfare of Japan. (2017). *The patient survey*. Statistics and Information Department, Minister's Secretariat. https://www.mhlw.go.jp/toukei/saikin/hw/kanja/17/index.html

Moretta, T., & Benvenuti, S. M. (2022). Early indicators of vulnerability to depression: The role of rumination and heart rate variability. *Journal of Affective Disorders*, *312*, 217–224. https://doi.org/10.1016/j.jad.2022.06.049

Nikolin, S., Tan, Y. Y., Schwaab, A., Moffa, A., Loo, C. K., & Martin, D. (2021). An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis. *Journal of Affective Disorders*, *284*, 1–8. https://doi.org/10.1016/j.jad.2021.01.084

Pavlova, B., & Uher, R. (2020). Assessment of psychopathology: Is asking questions good enough? *JAMA Psychiatry*, *77*(6), 557–558. https://doi.org/10.1001/jamapsychiatry.2020.0108

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Richter, T., Fishbain, B., Fruchter, E., Richter-Levin, G., & Okon-Singer, H. (2021). Machine learning-based diagnosis support system for differentiating between clinical anxiety and depression disorders. *Journal of Psychiatric Research*, *141*, 199–205. https://doi.org/10.1016/j.jpsychires.2021.06.044

Richter, T., Fishbain, B., Markus, A., Richter-Levin, G., & Okon-Singer, H. (2020). Using machine learning-based analysis for behavioral differentiation between anxiety and depression. *Scientific Reports*, *10*(1), 16381. https://doi.org/10.1038/s41598-020-72289-9

Richter, T., Fishbain, B., Richter-Levin, G., & Okon-Singer, H. (2021). Machine learning-based behavioral diagnostic tools for depression: Advances, challenges, and future directions. *Journal of Personalized Medicine*, *11*(10), 957. https://doi.org/10.3390/jpm11100957

Rose, E. J., & Ebmeier, K. P. (2006). Pattern of impaired working memory during major depression. *Journal of Affective Disorders*, *90*(2–3), 149–161. https://doi.org/10.1016/j.jad.2005.11.003

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., … Keller, M. B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*(5), 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8

Sauter, M., Stefani, M., & Mack, W. (2022). Equal quality for online and lab data: A direct comparison from two dual-task paradigms. *Open Psychology*, *4*(1), 47–59. https://doi.org/10.1515/psych-2022-0003

Semkovska, M., Quinlivan, L., O'Grady, T., Johnson, R., Collins, A., O'Connor, J., … Gload, T. (2019). Cognitive function following a major depressive episode: A systematic review and meta-analysis. *Lancet Psychiatry*, *6*(10), 851–861. https://doi.org/10.1016/s2215-0366(19)30291-3

Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., … Bulgheroni, M. (2019). Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: Systematic review. *JMIR Mental Health*, *6*(2), e9819. https://doi.org/10.2196/mental.9819

Šimundić, A. M. (2009). Measures of diagnostic accuracy: Basic definitions. *Electronic Journal of the International Federation of Clinical Chemistry*, *19*, 203–211.

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., … Delespaul, P. (2017). The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depression and Anxiety*, *34*(6), 481–493. https://doi.org/10.1002/da.22647

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–26. https://doi.org/10.1145/3191775

Wang, X., Zhou, H., & Zhu, X. (2020). Attention deficits in adults with major depressive disorder: A systematic review and meta-analysis. *Asian Journal of Psychiatry*, *53*, 102359. https://doi.org/10.1016/j.ajp.2020.102359

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site: http://onlinelibrary.wiley.com/doi/10.1111/jpr.12565/suppinfo.

Data S1. Supporting Information.

(Received January 31, 2024; accepted October 2, 2024)