Automated Inference of Cognitive Performance by Fusing Multimodal Information Acquired by Smartphone

Takashi Hamatani
NTT DOCOMO, INC.

Keiichi Ochiai
NTT DOCOMO, INC.

Akiya Inagaki
NTT DOCOMO, INC.

Naoki Yamamoto
NTT DOCOMO, INC.

Yusuke Fukazawa
NTT DOCOMO, INC.

Masatoshi Kimoto
NTT DOCOMO, INC.

Kazuki Kiriu
Graduate School of Engineering, The University of Tokyo

Kouhei Kaminishi
Graduate School of Engineering, The University of Tokyo

Jun Ota
Research into Artifacts, Center for Engineering, The University of Tokyo

Yuri Terasawa
Department of Psychology, Keio University

Tsukasa Okimura
Department of Neuropsychiatry, Keio University, School of Medicine

Takaki Maeda
Department of Neuropsychiatry, Keio University, School of Medicine

ABSTRACT

Recognizing human cognitive performance is important for preserving working efficiency and preventing human error. This paper presents a method for estimating cognitive performance by leveraging multiple information available in a smartphone. The method employs the Go-NoGo task to measure cognitive performance, and fuses contextual and behavioral features to identify the level of performance. It was confirmed that the proposed method could recognize whether cognitive performance was high or low with an average accuracy of 71%, even when only referring to inertial sensor logs. Combining sensing modalities improved the accuracy up to 74%.

CCS CONCEPTS

• **Applied computing** → **Consumer health**; *Psychology*.

KEYWORDS

cognitive performance; Go-NoGo task; smartphone log; machine learning

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3346275

ACM Reference Format:
Takashi Hamatani, Keiichi Ochiai, Akiya Inagaki, Naoki Yamamoto, Yusuke Fukazawa, Masatoshi Kimoto, Kazuki Kiriu, Kouhei Kaminishi, Jun Ota, Yuri Terasawa, Tsukasa Okimura, and Takaki Maeda. 2019. Automated Inference of Cognitive Performance by Fusing Multimodal Information Acquired by Smartphone. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341162.3346275

1 INTRODUCTION

The World Health Organization (WHO) reports mental disorders in a workplace cost the global economy US$ 1 trillion per year in lost productivity[1]. It also reports work is good for mental health, however negative working environment may lead to health problems. Although identification of positive and negative effects of work is difficult, however periodic screening of cognitive performance may help to judge whether a person is upon positive effect by working or not in workplace. Measuring cognitive performance also helps to preserve productivity and prevent human error.

There are several screening methods for cognitive performance such as Go-NoGo task [12] and Psychomotor Vigilance Task (PVT) [6]. These psychological tests can quantitatively measure human performance index as either success ratio of Go and NoGo responses or average response speed against visual stimulus. Nevertheless, they do not suit continuous measurement of cognitive performance since they

[1] https://www.who.int/mental_health/in_the_workplace/en/

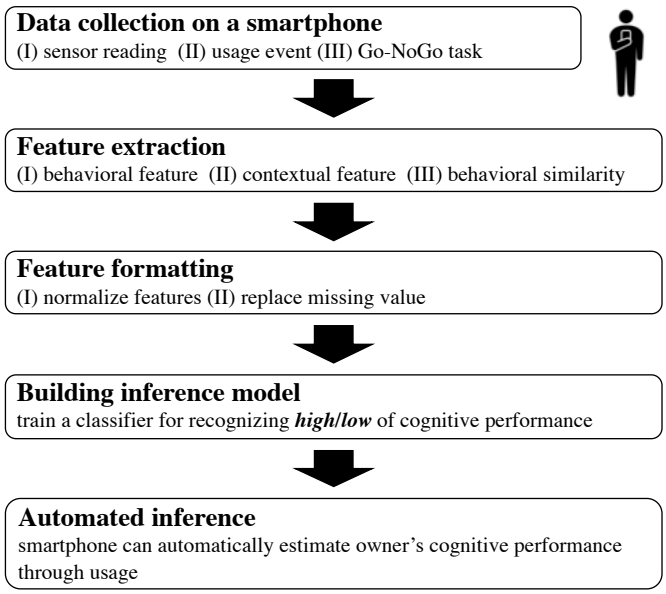


Figure 1: Overview of the proposed method.

involves a few minutes. While there have been bountiful studies aiming at automated estimation of human internal context related to health (e.g. stress [13], depression [3], anxiety [7], etc.) by using smartphone, the estimation of human cognitive performance is still under explored, and to date, only a few studies have been reported to our knowledge [2, 11]. Therefore, it is not well studied what features are effective to estimate human cognitive performance regardless of plenty of sensing modalities installed in a smartphone.

In this study, it was hypothesized that the combination of multiple sensing modalities in a smartphone can sense changes of user behavior and smartphone usage, allowing for differentiating the level of human cognitive performance. To validate that, a feasible experiment was firstly designed to collect reliable groundtruth of cognitive performance in the real environment. Then, large-scale dataset composed of multimodal information and ground truth of performance has been collected through 34 participants.

This paper proposes a method of estimating human cognitive performance using a smartphone as illustrated by Figure 1. It illustrates the procedure of the proposed method, where the collected information on smartphone is translated into behavioral and contextual features related to one's cognitive performance, and then formatted in order to build inference model. Consequently, over 750 days of multimodal sensor logs and cognitive tests were collected, and then the inference model was built for automated inference of cognitive performance. The key idea of the method is to combine both of behavioral feature (e.g, physical movement of body and smartphone, spatial movement of user, usage of smartphone) and contextual feature (e.g. ambient environment, state of smartphone) to continuously track the situation facing the user and his/her behavior.

Through validation across 34 subjects, this study revealed real-behavioral features (i.e. features characterized by inertial sensors) could represent the owner's cognitive performance better than contextual features, and demonstrated that the cognitive performance levels of the subjects could be estimated with over 70% accuracy, using only inertial sensors in smartphones. Finally, fusing multimodal features improved robustness across different users and boosted the accuracy up to 74%. The primary contributions of this paper are as follows:

- Designing a feasible experiment to collect cognitive performance, and collecting large-scale dataset in the wild environment.
- Proposing behavioral and contextual feature engineering across plenty of sensor and usage logs in a smartphone
- Demonstrating that the proposed method showed 74% accuracy for identifying high and low cognitive performance, and physical movement of a smartphone and its owner was the dominant feature in recognition.

2 RELATED WORKS

Several studies have been conducted to measure cognitive performance using smartphone logs. Abdullah et al. investigated the relationship between cognitive performance, smartphone usage (screen on/off events), sleep, and chronotype [2]. They built a model to estimate cognitive performance by leveraging psychomotor vigilance task (PVT) result, and reported that screen on/off event was an effective feature of estimating cognitive performance. Similarly, Murnane et al. presented a correlation between cognitive performance and the usage of applications related to productivity (e.g. Evernote, OfficeSuite) [11]. There also have been many studies on the estimation of human cognitive performance by leveraging different sensing modalities. Hou et al. used electroencephalogram (EEG) to profile a subject's stress level using the combination of emotion and workload [9]. Abdelrahman et al. leveraged thermography to sense human forehead and nose temperatures since the balance of them represented subjective cognitive load [1]. However, the use of dedicated sensing modalities may hinder the scalability of their solution.

In contrast, the method proposed in this paper fuses many types of smartphone logs and explores the relationship between sensing modalities and human cognitive performance. It also employs Go-NoGo task result as the groundtruth of

Figure 2: Go-NoGo task application; users respond by tapping the screen for "go" stimulus and suppressing taps for "nogo" stimulus.

cognitive performance rather than PVT result, leading to building inference model to suit the situation that not only task execution but also task inhibition is important such as office work.

3 DATA COLLECTION

This section describes the method of dataset collection to verify the following hypothesis: the high/low performance of human cognitive performance affects human-smartphone interaction, finally showing different patterns across a variety of sensor types. An Android application was developed for collecting essential sensor logs and ground truth of human cognitive performance.

Design of Experiment

This study aims at assessing performance in workplace for screening mental health of workers. In pursuit of this goal, Go-NoGo task, a psychological test to measure execution and inhibition function, is employed to quantify working performance rather than PVT since execution speed and inhibition functions are equally important for dealing task precisely and switching task appropriately. Although Go-NoGo task requires a few minutes to be finished, it is desired to complete this task as much as possible for accurately track the fluctuation of performance in a day. To balance this trade-off, the participants were asked to complete Go-NoGo task three times a day in (1) morning (9:30 to 10:30), (2) noon (12:00 to 13:00), and (3) evening (16:00 to 17:00). These durations were selected to capture cognitive performance upon (1) the start of work , (2) during break, and (3) the end of work.

Table 1: Collected usage event logs.

log type	frequency
power connection, screen on/off, earbud connection, application launch & quit	event-driven

Table 2: Collected sensor readings and contexts.

log type	frequency
acceleration, gyroscope, slope of smartphone, acceleration without gravity	1Hz
pressure, illuminance, battery level, connected & visible Wi-Fi access points, google activity recognition result, GPS location	every 5 minutes
storage level, day of week, weekday and holiday	23:00 everyday

Data Collection Application

Figure 2 depicts the application for data collection. It consists of (1) an on-screen Go-NoGo task application and (2) a data-logging function always running in the background. The Go-NoGo task is widely used to gauge human cognitive performance in terms of execution and motor inhibition [12]. The Go-NoGo task application takes about one minute and randomly shows one of the eight designated characters 72 times. Users were instructed to respond (i.e. tap the screen) to six of the characters (Go-stimuli) and ignore two of them (NoGo-stimuli). They were also required to respond to the Go-stimuli as fast as possible. The characters for Go and NoGo stimulus are randomly chosen for individuals in order to remove the effect of characters themselves. The application collected the success ratio of Go and NoGo responses, as well as the average response time for Go-stimulus. It is noted that the average response time deviated significantly among the subjects, and thus the shown time of each character was calibrated to get an approximately 90% success ratio for each subject.

Additionally, the application always collects usage event logs and multiple sensor readings with different frequencies, as shown in Tables 1 and 2. The information was aggregated by windows with different time scales and used for cognitive performance estimation. The participants agreed to the collection of this information[2].

[2]This study was approved by the ethics committee of the Graduate School of Medicine, part of the Faculty of Medicine at the University of Tokyo.

Data Collection Scenario

A total of 39 healthy employees from the R&D division of a company were recruited. There were 34 males and 5 females, aged from their 20s to 50s[3]. They installed the data collection application to their own smartphones and collected data from November 13th, 2017 to January 31st, 2018. Finally, the 1906 (Mean = 52.9, S.D. = 29.5) surveys of the Go-NoGo task were collected. The summary of the dataset is as follows. The dataset included 779 days, which completed both of the sensor logs and at least one Go-NoGo task result. The 779 days consisted of 649 weekdays and 130 holidays. Unfortunately, the sensor readings were found to not be properly stored in some devices for 5 of the 39 participants. Hence, the Go-NoGo task results were analyzed through 39 participants, while the evaluation of cognitive performance estimation was conducted through 34 participants.

It should be noted that the participants installed the application to their smartphones for work rather than personal. That is, the dataset could easily track user activities during work days but had the potential for failure on days off. The dataset was also incomplete for some participants due to limited permission for the GPS sensor and application logs. Specifically, 11% of the GPS and 17% of the application usage logs could not be retrieved. Missing value regarding these sensing modalities is compensated during a feature formatting procedure as described later.

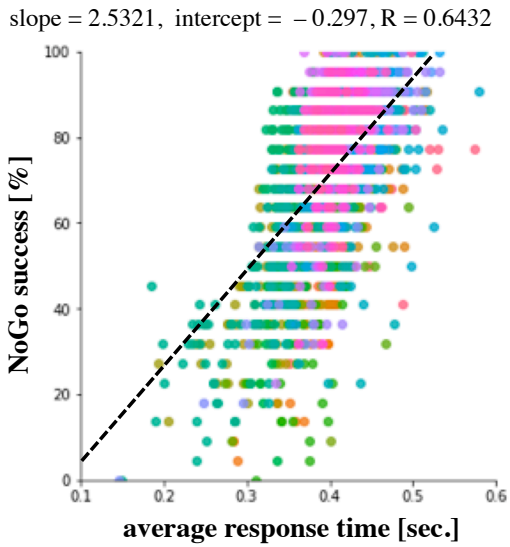
4 DEFINITION OF COGNITIVE PERFORMANCE

A total of 1906 instances of Go-NoGo task results were collected and analyzed from the following perspectives. In Go-NoGo task, execution function can be measured through Go response, meanwhile inhibition function can be measured through NoGo response. Hirose et al. proposes *efficiency* index as below.

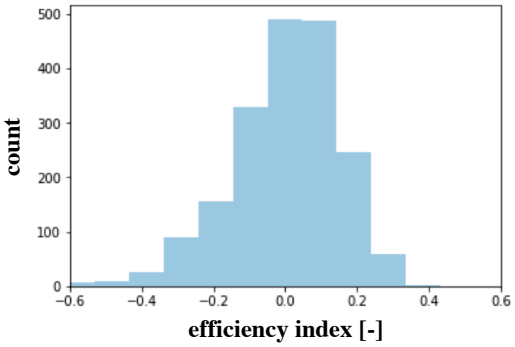
$$efficiency = y - f(x) \quad (1)$$

Here, efficiency is derived by subtracting average NoGo success ratio ($f(x)$) obtained by giving actual reaction speed x to regression line $f()$ presented in Figure 3(a) from actually obtained NoGo success ratio y. Namely, *efficiency* represent the difference in the inhibition performance (accuracy of NoGo response) against an ordinary level under the same level of execution performance (i.e. response speed) [8]. This metric is employed to represent our assumption: both of response speed and accuracy are important for dealing task precisely and switching task appropriately in workplace. This is why we employ *efficiency* as practical cognitive performance in workplace. We expect that long-term measurement of *efficiency* would reflect the change of cognitive load and/or mental health of workers.

[3]participants detail: 27 in 20s, 9 in 30s, 1 in 40s, and 2 in 50s.



(a) Response time and success rate of NoGo response with a regression line of them (color indicates individual user.)



(b) Distribution of efficiency index.

Figure 3: Summary of Go-NoGo task with user highlights.

Figure 3(b) depicts the distribution of calculated *efficiency* values across 1906 instances of Go-NoGo task. As described in the previous section, this study aims at day-by-day assessment of cognitive performance level in order to assess long-term effect of work upon one's mental health. Therefore, the collected groundtruth (i.e. *efficiency*) in the morning, noon, and evening were averaged as representative ground truth of daily performance. However, performance of execution function (average response time) and inhibition function (NoGo success rate) was widely distributed among subjects as shown in Figure 3(a). Accordingly, *efficiency* values was normalized using z-score within a subject, and then the class "high performance" was defined by z-score > 0, where

Table 3: Features regarding interaction extracted from usage event.

log type	extracted feature
charge event, earbud connection	*count, ratio* of connection
screen on/off	*avg**, *S.D.**, *max, min* of on-duration, *count* of on & off
application usage	*count, ratio* of app launch, *average, total* of use duration, *count, duration* of usage sessions, *average* number of used applications & application switching in sessions

**Note: avg, S.D.* indicate average and standard deviation, respectively.

Table 4: Features extracted from smartphone sensor readings (i.e. inertial sensors, context information, location).

log type	extracted feature
acceleration, gyroscope, slope	*avg, S.D., max, min, diff** for each axis, *correlation coefficient* for each pair of 3 axes
pressure, illuminance, battery level	*avg, S.D., max, min, diff*
google activity recognition	*count, ratio* for each activities (7 types)
Wi-Fi	*ratio* of Wi-Fi on & connection, *count* to the mostly connected AP & found APs
GPS location	*max distance* from home, *distance, max radius* of daily trace, *max, min, diff* of latitude, longitude and altitude, *count* of places visited
storage level	*raw value*
day of week, weekday and holiday	*one-hot value*

**Note: diff* indicates difference between max and min.

the alternative was "low performance". This binary transformation is based on assumption that behavioral characteristics when high or low *efficiency* within the person are common to different person, even if the value of *efficiency* index can locate in different range depending on the person. This allows comparing the relationship between smartphone logs and *efficiency* index among different subjects. Consequently, the number of *high* and *low* labels of *efficiency* in the collected dataset were 444 (57.0%) and 335 (43.0%), respectively. Thereafter, this label was employed as ground truth in terms of cognitive performance for supervised learning.

5 PROPOSED METHOD

Overview

Figure 1 shows that the method is comprised of the following five consecutive tasks: (1) essential data collection through Android application (described in the previous section), (2) translation of raw sensor readings to behavioral and contextual features, and behavioral similarity, (3) formatting feature values to successfully train a machine learning model, (4) training and optimization of the model to infer user cognitive performance, and (5) automated estimation using a model built by supervised learning.

Design of Features

First, behavioral and contextual features were extracted from raw sensor values using time windows of 1-, 6- and 24- hour. Using different resolution of window ensures to observe macro- and micro- scale activities of user. The statistical values extracted for each sensor type are summarized in Tables 3, 4, and 5, and described below.

Feature of interaction events

Human-smartphone interaction events were tracked by earbud and power connection, screen-on, and application usage logs. Using these event logs representing active usage of smartphone, features regarding interaction shown in Table 3 were extracted as below. Power connection and earbud connection events were captured as the timings of connection and disconnection. In addition to counting up the number of connection events, *ratio* was calculated as the ratio of connection duration against total duration. User interaction events were tracked by screen-on and application logs. The number and duration of screen on event were aggregated as *count, avg. S.D., max* and *min*. An application event consisted of a timestamp and application package name. It was first translated into a pair consisting of a timestamp and application category (published on the Google Play Store) due to the huge number of unique package names. It should be noted that some applications are not publicly available (e.g. built-in applications by smartphone makers and telecommunications carriers). Accordingly, category labels were manually given according to function (e.g. OS setting, OS home screen). For each category, the number of launch, ratio across a series of hourly launch, and launch duration were calculated. The application usage was also integrated by session unit, which represents a series of application usage from screen on to off.

Table 5: Features of activity similarity.

log type	extracted feature
GPS location	*Jaccard, Dice Index* of visiting points compared to 1 day/week ago
screen on	*Bhattacharyya coefficient* of on-distribution, difference of total on-count & on-duration against his/her average usage & 1d/1w ago
application	*Bhattacharyya coefficient, difference of total count & duration* for each application against his/her average usage & 1d/1w ago

Feature of Sensor Logs of Smartphone

Table 4 represents features extracted from sensor readings equipped with a smartphone. Each feature is derived by the following procedures. For sequential values, such as inertial sensor (i.e. acceleration, gyration and slope) readings, statistical values were calculated as follows: average, standard deviation, max value, min value, and difference between max and min (as denoted by *avg, S.D., max, min, diff*). The correlation coefficient was also calculated for each pair of sensor axes. For the measured scalar values (e.g. pressure, ambient illuminance, and battery level), average, standard deviation, max, min, and difference were calculated.

The other sensor features were derived as following: the results of the Google activity recognition package were aggregated into two values (*count* and *ratio*): *count* naturally counts up the number of detected activities for each type, and *ratio* translates *count* values into the probability of occurrence among all activity classes. The Wi-Fi log consists of connection and observation logs. The *ratio* of Wi-Fi standby and connection was calculated, and the number of mostly connected Wi-Fi access point and found access points was aggregated. GPS trace was used for tracking spatial and semantic locations. First, the coordinates of their homes and lists of stay points were inferred. Then the total and maximum distances from the home were calculated from daily trace. The stay points were also counted to track the subjects' semantic locations [3]. Storage level was recorded to capture the change in the remaining amount due to user activity on smartphone. Date and work information were added to represent users' work styles.

Feature of Behavioral Similarity

The regularity of human behavior is expected to depend on the one's mental state. To capture the periodic signal and changes in the behavior patterns of subjects, behavioral similarity features were designed using the semantic location and interaction of smartphones as summarized in Table 5. The spatial location captured by GPS was first translated into a set of semantic places. Then, Jaccard and Dice indices were calculated to measure similarities of mobility patterns. For the patterns of screen activation and application usage, Bhattacharyya coefficient was calculated for their probability distribution of screen on event or application launch over every hour.

Feature Formatting

To ensure the effective training of classifier, it is essential to understand the difference among users and deal with missing feature values. First, each feature value was scaled to a standard distribution with zero mean and unit variance to fairly compare feature values across different types of value range. Thereafter, feature compensation was applied for incomplete data, by using the average data available from either the subject or the other subjects.

Detection of Cognitive Performance

A classifier was built to recognize the two states of user cognitive performance (*high* or *low*) by leveraging a variety of feature values representing user behavior on a given day. The XGBoost [5] model, which presents state-of-the-art performance on classification task, ranking task, etc., was employed. Due to incompletely balanced ground truth labels, a SMOTE [4] oversampling algorithm was applied to avoid overfitting caused by a larger weight of the major class. The balanced feature values and ground truth labels are fed to XGBoost model to be tuned.

Optimization of Features for Inference

High dimensional features often cause over-fitting and increase computational costs. To reduce the dimension of the feature space, feature importance values derived by the XGBoost model were referred to, and features with relatively lower importance were filtered out. Finally, approximately 5% of all feature values were left.

6 PERFORMANCE EVALUATION

This section elaborates upon the evaluation results of the proposed method based on the following goals: to clarify the contributions of each log type available on smartphone for detecting cognitive performance, and confirm the overall capability of the proposed method to estimate cognitive performance by combining multimodal features.

Evaluation Setting

To measure the capability of system generalization for a new subject, a leave-one-subject-out (LOSO) cross-validation method was used over 34 subjects. The participant's own dataset (i.e. test dataset) was used for formatting his/her features. However, the test dataset was never leaked to training and

Table 6: Result of performance recognition with different feature types.

used feature type	performance metrics				
	Acc.	Spe.	Sen.	Pre.	AUC
baseline	51.0%	57.0%	43.0%	43.0%	*N/A*
(1) inertial sensor	72.0%	76.2%	66.3%	68.3%	0.781
(2) context information	62.7%	63.4%	62.7%	56.0%	0.670
(3) location	59.4%	56.0%	63.9%	52.4%	0.641
(4) interaction information	61.7%	64.0%	58.8%	55.3%	0.664
(5) behavioral similarity	59.4%	60.2%	58.1%	52.6%	0.610
(6) calendar	50.5%	48.3%	53.9%	43.8%	0.508
using all features	**74.1%**	**77.1%**	**70.7%**	**70.1%**	**0.809**

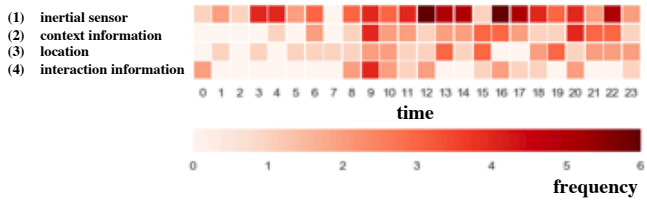


Figure 4: Distribution of the number of selected features for feature types and time periods.

feature selection procedure. The metrics of system performance and abbreviations were as follows. (1) *Accuracy* ***(Acc.)****:* the overall success rate of *high* and *low* classification. (2) *Specificity* ***(Spe.)****:* the coverage of *high* state recognition. (3) *Sensitivity* ***(Sen.)****:* the coverage of *low* state recognition. (4) *Precision* ***(Pre.)****:* the ratio of correctly retrieved *low* state out of all detections. (5) *ROC-AUC* ***(AUC)****:* the area under ROC-curve. A baseline estimation method was introduced for comparison; it randomly estimates according to the probability of *high* and *low* labels.

Relationship between Feature and Performance

Table 6 depicts the performance variation with different feature types. Each column represents performance when using particular feature for estimation. The labels (1) **inertial sensor**, (2) **context information**, (3) **location**, (4) **interaction information**, (5) **behavioral similarity**, and (6) **calendar** mean the performance when using features regarding (1) accelerometer, gyroscope, and orientation sensor, (2) pressure and illuminance sensors, battery and Wi-Fi states, and google activity recognition, (3) location given by GPS, (4) events of power connection, earbud connection, screen on/off, and application usage, (5) behavioral similarity features in Table 5, and (6) calendar information, respectively. It can be seen that using calendar information does not improve detection performance compared to baseline, highlighting that cognitive performance is not predictable by periodical rules. These results indicate that using inertial sensors contributes mainly to allowing the classifier to detect whether the cognitive performance is high or low. Note that behavioral similarity features are incomplete due to the lack of samples, since a comparison between logs and the logs in up to one week ago is required. Accordingly, feature values are not complete in the first week of experiment.

Comparing the results, it was found that the system could recognize higher/lower states of efficiency index with over 70% (approximately 21% increase against baseline) accuracy using only inertial sensors. In the other metrics, the proposed method could improve performance compared to the baseline performance. It was also confirmed that user-smartphone interaction events, such as application usage frequency and screen-on duration, ensured the recognition of user cognitive performance, similar to previous studies [2]. However, accuracy was limited to approximately 60% when using only features regarding interaction; this suggests user-smartphone interaction log is not fully capable to infer users' cognitive performance, since it depends on the frequency of user interaction. It was also confirmed in another evaluation that XGBoost algorithm had showed better performance on accuracy (up to 72.0%) than Random Forest algorithm (64.3%).

Figure 4 depicts the distribution of the number of selected important features. Vertical axis means feature type and horizontal axis means time period by an hour as well as the color represents freqency of selection. Note that features of calendar and behavioral similarity are omitted since they are calculated by day unit. The figure briefly represents temporal contribution of each feature type for cognitive performance estimation. It is found here that all the feature types contribute from 8:00 to 23:59 and inertial sensor features are the most important among them. It is also notable that inertial sensor features respond even in approximate bed time (i.e. 3:00 - 6:59). This indicates that owner's behavioral feature during bed time has potential to represent the duration and quality of sleep, and further may be leveraged to estimate his/her cognitive performance. This finding is partially related to a previous study on the relationship between sleep duration and quality, and cognitive performance [10].

Combined Performance

Table 6 also presents the inference performance when using combination of all the available features. It is found that fusing multiple sensing modalities can improve accuracy by 2.1% compared to using only inertial sensors. This indicates that the performance of human cognitive function can be

Table 7: Statistics of Acc. for 34 participants.

used feature type	statistics of **Acc.**			
	avg.*	S.D.*	max	min
inertial sensor	72.0%	13.2%	100%	25.0%
all features	75.1%	9.7%	100%	56.0%

largely found with the motion behaviors captured by inertial sensors, and other sensing modalities can improve robustness by considering ambient environment and interactive behaviors. Indeed, in the evaluation for each participant unit, *average, standard deviation, max, min* of accuracy were summarized as Table 7. This result demonstrated the worst case of performance of the proposed method showed poorer than baseline method in Table 6 when using only inertial sensor features for a particular subject. In contrast, it also indicates that the combination of various features available on smartphone can consistently estimate a user's cognitive performance, where the minimum performance is still higher than baseline method as well as the standard deviation is much smaller. This highlights the advantage of using multimodal information to improve robustness of inference model.

Consequently, the proposed method could distinguish *high* or *low* state of cognitive performance calculated as *efficiency* with up to 74% accuracy. However, the performance evolution appears limited since the estimation model is built by LOSO cross validation and not optimized for individuals. The performance is expected to be improved by training the inference model with further collected dataset in the consecutive study.

7 CONCLUSION

This paper presented a method to identify the level of human cognitive performance by leveraging multimodal information in a smartphone. Behavioral and contextual features were designed over 15 types of sensor logs, and the method was examined through 779 traces of 34 participants. The results demonstrated that the method could classify high and low states of cognitive function with over 70% accuracy using inertial sensor features, and consistently estimate cognitive performance across 34 subjects by combining sensing modalities in a smartphone. This study is currently limited in terms of dataset diversity. The forthcoming study aims at validating generalizability of the proposed method across a variety of subject. We also aim at updating our algorithm by employing deep learning and better feature selection method.

REFERENCES

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 33.

[2] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 178–189.

[3] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 1293–1304.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] 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*. ACM, 785–794.

[6] David F Dinges, Naomi L Rogers, and Jillian Dorrian. 2004. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In *Sleep deprivation*. CRC Press, 67–98.

[7] Yusuke Fukazawa, Taku Ito, Tsukasa Okimura, Yuichi Yamashita, Takaki Maeda, and Jun Ota. 2019. Predicting anxiety state using smartphone-based passive sensing. *Journal of Biomedical Informatics* 93 (2019), 103151.

[8] Satoshi Hirose, Junichi Chikazoe, Takamitsu Watanabe, Koji Jimura, Akira Kunimatsu, Osamu Abe, Kuni Ohtomo, Yasushi Miyashita, and Seiki Konishi. 2012. Efficiency of go/no-go task performance implemented in the left hemisphere. *Journal of Neuroscience* 32, 26 (2012), 9059–9065.

[9] Xiyuan Hou, Yisi Liu, Wei Lun Lim, Zirui Lan, Olga Sourina, Wolfgang Mueller-Wittig, and Lipo Wang. 2016. CogniMeter: EEG-based brain states monitoring. In *Transactions on Computational Science XXVIII*. Springer, 108–126.

[10] Seiko Miyata, Akiko Noda, Kunihiro Iwamoto, Naoko Kawano, Masato Okuda, and Norio Ozaki. 2013. Poor sleep quality impairs cognitive performance in older adults. *Journal of sleep research* 22, 5 (2013), 535–541.

[11] Elizabeth L Murnane, Saeed Abdullah, Mark Matthews, Matthew Kay, Julie A Kientz, Tanzeem Choudhury, Geri Gay, and Dan Cosley. 2016. Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 465–477.

[12] Daniel J Simmonds, James J Pekar, and Stewart H Mostofsky. 2008. Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. *Neuropsychologia* 46, 1 (2008), 224–232.

[13] Naoki Yamamoto, Keiichi Ochiai, Akiya Inagaki, Yusuke Fukazawa, Masatoshi Kimoto, Kazuki Kiriu, Kouhei Kaminishi, Jun Ota, Tsukasa Okimura, Yuri Terasawa, and Takaki Maeda. [n.d.]. Physiological Stress Level Estimation Based on Smartphone Logs. In *Mobile Computing and Ubiquitous Networking (ICMU), 2018 11th International Conference on*.