IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 3, 2021, accepted March 16, 2021, date of publication March 26, 2021, date of current version April 12, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3069049*

# Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone

**MASAKI GOGAMI[1], YUKI MATSUDA[1,2], (Member, IEEE), YUTAKA ARAKAWA[3], (Member, IEEE), AND KEIICHI YASUMOTO[1], (Member, IEEE)**

[1]Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan
[2]Japan Science and Technology Agency (JST), PRESTO, Tokyo 102-0076, Japan
[3]Department of Advanced Information Technology, Kyushu University, Fukuoka 819-0395, Japan

Corresponding author: Masaki Gogami (gogami.masaki.gg8@is.naist.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (KAKENHI) Grant Number 18H03233, and in part by the Japan Science and Technology Agency (JST), Precursory Research for Embryonic Science and Technology (PRESTO) Grant Number JPMJPR2039.

**ABSTRACT** Some respondents make careless responses due to the ''satisficing,'' which is an attempt to complete a questionnaire as quickly and easily as possible. To obtain results that reflect a fact, detecting satisficing and excluding the responses with satisficing from the analysis targets are required. One of the devised methods detects satisficing by adding questions that check violations of instructions and inconsistencies. However, this approach may cause respondents to lose their motivation and prompt them to satisficing. Additionally, a deep learning model that automatically answers these questions was reported. This threatens the reliability of the conventional method. To detect careless responses without inserting such screening questions, machine learning (ML) detection using data obtained from answer results was attempted in a previous study, with a detection rate of 55.6%, which is not sufficient from the viewpoint of practicality. Therefore, we hypothesized that a supervised ML model with a higher detection rate could be constructed by using on-screen answering behavior as features. However, (1) no existing questionnaire system can record on-screen answering behavior and (2) even if the answering behavior can be recorded, it is unclear which answering behavior features are associated with satisficing. We developed an answering behavior recording plug-in for LimeSurvey, an online questionnaire system used all over the world, and collected a large amount of data (from 5,692 people) in Japan. Then, a variety of features were examined and generated from answering behavior, and we constructed ML models to detect careless responses. We call this detection method the ML-ABS (ML-based answering behavior scale). Evaluation by cross-validation demonstrated that the detection rate for careless responses was 85.9%, which is much higher than the previous ML method. Among the various features we proposed, we found that reselecting the Likert scale and scrolling particularly contributed to the detection of careless responses.

**INDEX TERMS** Answering behavior, careless response, online questionnaire, satisficing, smartphone, supervised machine learning, touchscreen.

## I. INTRODUCTION

Questionnaire results have the problem of low reliability due to the practice of ''satisficing,'' which means to complete a task as quickly and easily as possible [1]. In response to this problem, some methods have been devised to detect satisficing by inserting screening questions, such as questions that determine instruction violations or detect inconsistency [2], [3]. However, the insertion of such questions may increase the psychological load on the respondents. Therefore, it is desirable to avoid such questions because they would undermine the intrinsic motivation to cooperate with the survey of respondents who are answering carefully, and the questions themselves may induce satisficing.

The associate editor coordinating the review of this manuscript and approving it for publication was Waleed M. Alsabhan.

In response to this problem, research in the field of social psychology has been conducted from the viewpoint of how to optimize a set of questions for detecting satisficing. For example, Miura and Kobayashi [4] attempted to narrow down the minimum number of questions from an existing satisficing indicator to detect satisficing efficiently and accurately. However, narrowing down the questions from the results of the experiment could not be realized. In addition, a deep learning model that automatically answers these questions was reported recently [5]. This threatens the reliability of conventional methods that insert additional questions, and requires a detection method based on answering behavior rather than screening questions.

Ozaki and Suzuki [6] attempted to detect careless responses by machine learning (ML). They reported a detection rate of 55.6% using features generated from response results, but the detection rate was not sufficiently practical.

To detect careless responses with a higher detection rate, we hypothesized that answering behavior on a touchscreen can be used. However, (1) no current questionnaire system records on-screen answering behavior and (2) even if answering behavior is recorded, the features of answering behavior associated with satisficing are unclear. Hence, we designed and developed a plug-in that can record answering behavior such as scrolling, choosing and changing options, and entering and deleting text (hereafter referred to as ''Operation Logger''). Operation Logger was developed as a plug-in that runs on LimeSurvey [7], a widely used online questionnaire system. Therefore, the questionnaire system can be used in the same way as without the plug-in, and the answering behavior can be recorded by simply adding Operation Logger to the survey. Thus, respondents do not need to install any additional applications or perform any other additional operations.

We collected data using Operation Logger on Yahoo! Crowdsourcing[1] in Japan. We asked 5,692 respondents (containing approximate 5% of careless response) to answer a questionnaire, which was developed based on a questionnaire[2] by Miura and Kobayashi [4], consisting of 128 questions with both Likert scale and open-ended formats. Existing satisficing indicators (e.g., questions checking instruction violations or inconsistency) were used to assign a label of ''careless'' or ''careful'' to each respondent as ground truth. The features were generated from the answering behavior data and question responses, and supervised ML models performed binary classification to detect careless responses. We call this detection method the ''ML-based answering behavior scale'' (ML-ABS). We evaluated these models by mainly focusing on the recall, representing the detection rate of careless responses.

Moreover, we constructed models using features generated from answering behavior data for a limited page range and verified the relationship between the number of questions and the detection rate.

The rest of the paper is organized as follows:

In Section II, related works are surveyed, and the challenges of this study are defined. We introduce proposed features and a system to record the data in Section III. Section IV describes the data collection experiments conducted using the system. Section V defines careless responses and describes the methods and results of careless response detection using ML. A discussion of the results is given in Section VI, and finally, Section VII concludes this paper.

## II. RELATED WORK AND CHALLENGES

The subject of the reliability of questionnaire surveys can be broadly divided into questionnaire content and response. For the former, Cronbach's alpha is often used [8]. However, this paper focuses on the latter subject, that is, the reliability of responses.

Simon [9] developed the concept of cognitive psychology, which holds that humans minimize their cognitive effort to achieve some goal. Applying this concept to questionnaire surveys, they initiated this research area by defining ''satisficing'' as the tendency of humans to minimize their efforts to complete the requests for answers [1].

Since then, it has been shown that satisficing degrades the quality of survey results [2], [3], [10]–[13]. In particular, in online questionnaires, this problem is a factor that significantly contributes to quality deterioration [14]. Oppenheimer *et al.* [2] proposed the instructional manipulation check (IMC), which inserts instructions such as ''Please move to the next page without answering any question'' in the instructional text, and conducted an online questionnaire using the IMC. If the instruction is not followed, this is regarded as satisficing. As a result of two experiments with university students, the percentage of satisficing was reported to be 46% and 35%. These values suggest that careless responses may distort the findings from the questionnaire results.

There are other satisficing indicators such as Directed Question Scale (DQS) and Attentive Responding Scale (ARS) [3]. These methods are used to detect satisficing based on violation of the instructions and inconsistency, and abnormality of responses by adding some questions for satisficing detection to Likert scale questions. These indicators are widely used in a number of studies addressing questionnaires [4], [6], [15]–[19]. However, adding the questions with these methods is like pushing the respondent through a suspicion filter. This imposes a psychological burden on the respondent. In particular, for respondents who are responding carefully, the intrinsic motivation to cooperate with the survey can be diminished, and then, this method may cause satisficing by itself. In addition, Pei *et al.* [5] built a deep learning model that automatically answered IMC and DQS questions and reported that the model answered correctly with about 78.5% accuracy. This result undermines the reliability of the satisficing detection method described above. For these

[1] http://crowdsourcing.yahoo.co.jp
[2] https://osf.io/6gu3q

reasons, there is a need for a method to detect satisficing that is stress-free and does not require the addition of filtering questions.

Ozaki and Suzuki [6] attempted to detect careless responses based on satisficing without any satisficing indicator. They tried to detect careless responses using various ML algorithms constructed with features that were generated from response results such as sex, age, total response duration, number of consecutive identical responses, Mahalanobis distance, and its p-value. The amount of data was 2,000 samples and the device for response were personal computers. As a result, the highest detection rate for careless responses was 55.6%. They reported that the detection rates of many algorithms (11 algorithms) were approximately in the range of 40% to 50%. We assumed that the factor of low detection rate may not be on the aspect of the algorithm but the quality of the features. In addition, using sex and age as features is not desirable in an actual questionnaire because some cases should limit the sex or age of the respondents.

Although Schroeders *et al.* [20] aimed to detect careless responses by machine learning. However, since they divided the respondents into two groups and asked them to answer either carefully or carelessly, they may have obtained data that deviates from the actual environment. In addition, since the definitions of careless and careful are random, midpoint, and fixed pattern responses, which differ greatly from our definitions. As Ozaki and Suzuki [6] also state in their paper, there are not many studies aimed at detecting careless responses by machine learning, but we judged that it was not appropriate to compare the results in this paper based on above reasons.

In recent years, smartphones have been replacing PCs as the terminals used to answer online questionnaires [21]. In response to this, Tourangeau *et al.* [22] investigated the differences in the quality of questionnaire results between PCs, tablets, and smartphones. The evaluation targets of the study were the total response duration, non-response rate, and number of consecutive identical responses. As a result, it was observed that the response duration for smartphones tended to be longer than that for PCs and tablets. However, they concluded that the reliability of the results was not particularly different between devices. Because smartphones have the same reliability as PCs and high portability, their use as a terminal for answering online questionnaires will continue to increase in the future. Therefore, in this study, we decided to focus on smartphones as the device used for responses.

In summary, the challenges of this study were as follows:

1) To record variations in answering behavior of respondents by simple implementation without modifying the general questionnaire system.
2) To detect careless responses with high accuracy based on the answering behavior.

As an approach to Challenge 1, we developed a plug-in application written in JavaScript to record answering behavior on the touchscreen at the client side, and send behavior data to a database on the server. As an approach to Challenge 2, we generated features from answering behavior data recorded by the plug-in, and constructed ML models to detect careless responses. In addition, we examined features, such as inter- and intra-subject deviation, and aimed to improve the models.

## III. QUESTIONNAIRE SYSTEM

We examine answering behavior to detect careless responses in Section III-A, and examine the architecture and explain the implementation of the system recording the answering behavior in Section III-B.

**TABLE 1. Answering behavior related to satisficing.**

| Answering behavior | Unit | Scope | Original |
|---|---|---|---|
| Answering duration (Likert scale) | sec | Likert scale | – |
| Answering duration (open-ended) | sec | Open-ended | – |
| Changing Likert scale option | count | Likert scale | ✓ |
| Deleting text | count | Open-ended | ✓ |
| Deleting text rate | count/character | Open-ended | ✓ |
| Scrolling length | px | Entire | ✓ |
| Scrolling duration | sec | Entire | ✓ |
| Scrolling speed | px/s | Entire | ✓ |
| Reverse scrolling | count | Entire | ✓ |
| Long interval | count | Entire | ✓ |
| Straight-lining | question | Likert scale | – |
| Middle answer | question | Likert scale | – |
| Number of characters | character | Open-ended | – |

### A. ANSWERING BEHAVIOR

Table 1 lists the answering behaviors that are considered to be related to satisficing. The ''scope'' column indicates the question format type (Likert scale or open-ended) of each answering behavior. In this column, ''entire'' indicates that the behavior is related to the entire questionnaire regardless of the question format. In the ''Original'' column, ''✓'' indicates that the feature is newly proposed in this paper.

Answering duration has often been used to detect careless responses. Merrill *et al.* excluded respondents who completed the questionnaire earlier than an absolute threshold determined by analysis [23]. Furthermore, a Japanese survey company, NTT Com Online Marketing Research, Inc.,[3] considers responses in which the answering duration for the entire questionnaire is too short to be valid responses. There are some respondents who are not caught by such filtering for reasons such as giving careless responses only to certain parts of the questionnaire. However, there is a significant relationship between the answering duration and the quality of the responses [24]. In this study, the response duration was divided into ''answering duration (Likert scale)'' and ''answering duration (open-ended),'' and the average value for each form of question was used.

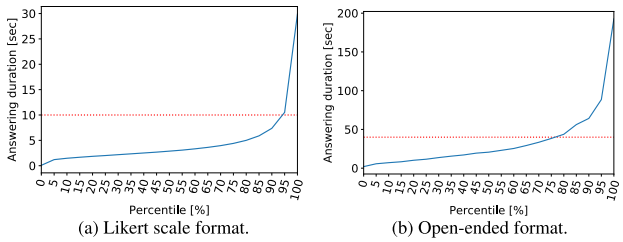
''Changing Likert scale option,'' ''deleting text,'' and ''deleting text rate'' are assumed to be answering behaviors that represent a state of thinking about answering accurately. Therefore, it is assumed that these behaviors will not increase if respondents give careless responses to complete a questionnaire quickly. In the research field of user authentication,

[3]http://research.nttcoms.com/service/qpolicy4.html

Alsultan *et al.* [25] succeeded in improving the FRR (False Rejection Rate) by extracting more advanced typing patterns using text deletion behavior, in contrast with previous work only focusing on keypress and release. Therefore, It is expected that more information could be obtained from text deletion behavior. The reason for adding the ''deleting text rate,'' which is calculated by dividing the amount of deleted text by the number of characters, is the probability of a text deletion occurring depending on the number of characters, and it might contribute more than the raw number of changes.

''Scrolling length'' is defined as the scrolling distance during one scrolling operation, ''scrolling duration'' is defined as the time spent for one scrolling operation, and ''scrolling speed'' is defined as the value obtained by dividing the scrolling length by the scrolling duration. Seo *et al.* [26] have achieved high accuracy by using scrolling speed and scrolling length for user authentication in mobile terminals. This work suggests that scrolling operation plays the role of original feature which is able to reveal individual differences. These can be regarded as parameters that represent the behavior of moving between questions. When a respondent is satisficing, it is assumed that the movement becomes coarse and fast due to the desire to finish quickly.

''Reverse scrolling'' is defined in the reverse direction of 100 pixels (px) or more. The reason for setting the threshold at 100 px is that this is the minimum scrolling distance needed to go back to the previous question when answering the LimeSurvey questionnaire with a standard smartphone. This behavior will represent changing the answer to that of the previous question while answering the questionnaire, or re-reading the question text at the beginning of the page. Because this behavior is considered to support the state of trying to answer carefully, it is assumed that the respondent is satisficing when this behavior is almost zero.




(a) Likert scale format. (b) Open-ended format.

**FIGURE 1.** Answering duration per question.

''Long interval'' is the number of inactive periods, during which the screen is not touched, that exceed the threshold. The threshold was determined based on responses in a preliminary experiment with the same experimental settings described in section IV. We asked 101 respondents to answer 128 questions, which consist of the Likert scale and the open-ended formats. The distribution of answering duration per question for each format is shown in Figure 1. From the shape of these graphs, as shown by the dashed line, we define the threshold of the long interval as 10 sec for the Likert scale format and 40 sec for the open-ended format. If the non-operation duration exceeds this threshold, it is considered to be a state of answering the questionnaire while doing some other task, because it is beyond the duration expected to be required for answering. If the value of this behavior is large, it is assumed that satisficing is occurring.

''Straight-lining'' is the number of times that a respondent provides the same response consecutively in a Likert scale question. Although non-satisficing respondents could give the same response consecutively, it is assumed that the value is greater for a respondent who is satisficing [27].

''Middle answer'' is the number of times a respondent selects an intermediate option, such as ''neither,'' in a Likert scale question. Although the middle answer can be selected often even by a non-satisficing respondent, people tend to choose the middle answer when they are satisficing because they do not want to pay the cognitive cost to confirm and express their opinion, and they practically abandon the answer [28]. Therefore, we believe that the number of intermediate responses is related to satisficing.

''Number of characters'' is the number of characters per open-ended question. When the number of characters and the degree of specificity of the content are not specified in the questionnaire, some respondents answer simply in one sentence, but others answer specifically in several sentences. This difference corresponds to the difference in the number of characters, and it is assumed that the tendency for satisficing is stronger when the number of characters is smaller.

### B. SYSTEM ARCHITECTURE

We examined a questionnaire system that records the answering behavior shown in Table 1. First, we investigated the use of a service called ClickTale,[4] which visualizes user behavior on a page, for example, using a heat map, for the improvement of web pages. However, while this service is highly versatile, we did not adopt it because it cannot record detailed answering behavior specific to questionnaires. Next, we considered the method of creating smartphone or web applications by ourselves. In this method, flexible implementation can be realized and desired answering behavior can be recorded. However, a smartphone application needs to be installed on the respondent's smartphone. For web applications, the questionnaire administrators need to learn how to operate the new system. In light of these shortcomings, and consideration of widespread adoption, we adopted the architecture of extending an existing questionnaire system rather than building a new system. With this approach, respondents do not need to add any software or change any setting. In addition, questionnaire administrators only need to learn how to operate the plug-in. Therefore, we focused on LimeSurvey, an open-source web questionnaire system. Unlike Google Forms[5] and Survey Monkey,[6] LimeSurvey allows users to create their own plug-ins. Using this mechanism, we developed the

[4]https://www.clicktale.com
[5]https://www.google.com/forms/about
[6]https://www.surveymonkey.com

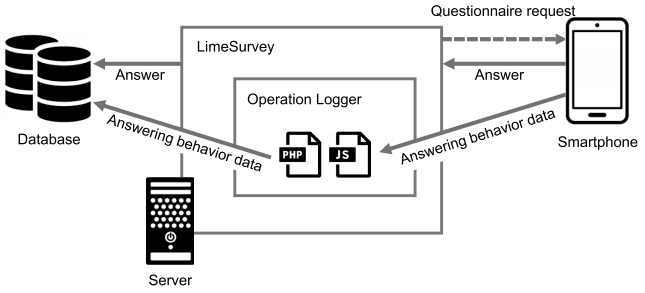



**FIGURE 2. Overview of questionnaire system that records on-screen answering behavior.**

Operation Logger plug-in to record the answering behavior shown in Table 1 using JavaScript, and built a questionnaire system that can record answering behavior. An overview of how the system with Operation Logger was implemented is shown in Figure 2. With only this plug-in, the answering behavior can be stored in the database along with the response results recorded by the standard function of LimeSurvey.

The detailed implementation of Operation Logger is described below. Three main types of events are used to acquire data: touch events, option taps, and text input. The touch events consist of touchstart, touchmove, and touchend, which acquire data at the moment when the screen detects a finger, while a finger is moving on the screen, and when a finger leaves the screen, respectively. For these touch events, the time, the coordinates in the screen, the scrolling distance from the top of the page, and the type of touch event are obtained and stored in a database. These data allow us to generate features concerning the scrolling length, scrolling duration, scrolling speed, reverse scrolling, and long interval. When the user taps a Likert scale option, the system records the answer time, the ID of the question, and the ID of the option. This makes it possible to generate the features concerning the answering duration (Likert scale), straight-lining, middle answers, and changing Likert scale option. When text is input, the response time, question number, input content, and type of recording trigger are recorded in the database. As a result of examining the recording units during text input, we decided that the trigger for generating a record is elapsing 1 sec of no input, switching from delete to input, switching from input to delete, and focusing out from the text box. This allows us to generate the features concerning the answering duration (open-ended), deleting text, deleting text rate, and number of characters.

## IV. DATA COLLECTION

A large-scale experiment was conducted using LimeSurvey with Operation Logger introduced in Section III. Section IV-A explains satisficing indicators. Section IV-B describes the content of the questionnaire. Section IV-C describes the procedure of the experiment.

### A. SATISFICING INDICATORS

In this study, several conventional satisficing indices were employed as the ground truth for the ML models.

The first satisficing indicator is the DQS [3]. To measure this indicator, some questions are inserted to instruct the respondent to make a choice. Here, the instruction does not necessarily include universal options. If the respondent does not follow the instruction, he or she is considered to be satisficing. In this study, three DQS questions were included, and satisficing was defined as one or more of the DQS questions being answered in a manner contrary to the instructions.

The second indicator is the ARS [3], which has two types: inconsistency and infrequency. Inconsistency ARS focuses on the difference in answer scales to a pair of similar questions with slightly different wording. The responses are defined as satisficing if the sum of the differences is 11 or more for 11 question pairs. Infrequency ARS focuses on the difference between the assumed answer and the actual answer, using questions having an answer option that is assumed to be chosen by everyone. Satisficing is defined as the sum of the differences of 11 question pairs being 12 or more. These thresholds were calculated using the Receiver Operating Characteristic (ROC) curve to maximize the percentage of randomly selected data (764 samples) that were correctly identified. As a result, the thresholds were set to 10.5 (Inconsistency) and 11.5 (Infrequency), strictly. In this paper, we followed this definition.

As for the IMC, which is commonly used in studies focusing on satisficing [2], [12], [29], [30], we decided that the IMC could not be used as a satisficing indicator and did not use it because the questionnaire did not include questions with a special instruction adopting an IMC.

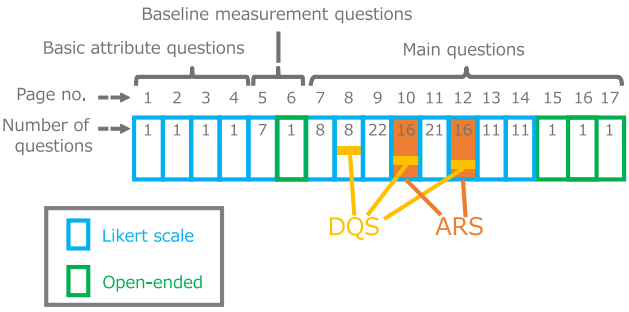
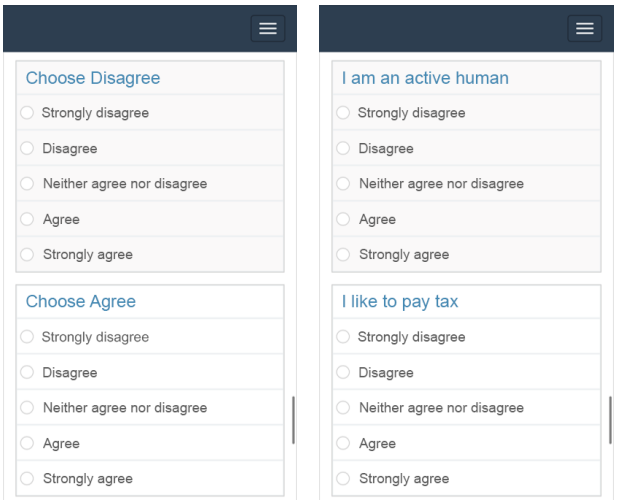



**FIGURE 3. Overview of the questionnaire.**

### B. QUESTIONNAIRE CONTENTS

Figure 3 shows a schematic of the questionnaire used in this experiment. The questionnaire, which asks respondents their personalities, mindsets, motivation for completing the questionnaire, and self-assessment of satisficing tendencies, such as "whether complete your answers as quickly as possible?" or "whether read the questions carefully?," was developed based on the questionnaire published by Miura and Kobayashi [11] with the following three changes. First, pages 1–6 and 15–17 were added as described below. Second, 11 dummy questions were added to make the ARS question pairs less obvious to respondents. Third, we reduced the number of DQS questions from five to three and changed

the placement of these questions to avoid the end and beginning of pages to further obscure their placement to respondents. The reason for this is that the DQS questions were placed at the end of five consecutive pages in the original questionnaire. As a result, the questionnaire consisted of 17 pages with 128 questions (120 questions in 5-point Likert scale and 6 questions in open-ended format and 3 questions of other types). The detailed questions are listed in the appendix. The estimated time needed to complete this questionnaire is about 15 minutes.




(a) Baseline measuring page. (b) Main question page.

**FIGURE 4. Examples of questionnaire screen on smartphone. Note that this questionnaire was conducted in Japanese.**

Pages 1–3 contained questions to obtain some basic attributes of respondents, and page 4 has a question to link the subject ID to the plug-in system. Page 5 questions measured the baseline of scrolling operation for each respondent. As an example, screenshots of the actual questionnaire screen of the baseline measurement page and the main question page are shown in Figure 4. Respondents scroll down to the next question and proceed to answer. The questions for baseline measurement ask the respondent to select a specified option, as shown in Figure 4(a). Because it avoids the process of expressing one's opinion, the cognitive cost of answering these questions is lower than that of general questions. Therefore, the scrolling behavior was measured when satisficing was less likely to occur. Page 6 is a question to measure the baseline of the delete rate for each respondent. The baseline for the delete rate was calculated by dividing the number of deletions by the number of characters when entering a sentence specified in the free text format. This feature was also used to calculate the relative delete rate for each respondent (delete_rate_Selfdev), which is a feature described below. Page 7 onwards contains the main question, and pages 7–14 contain questions on the big five personality scale, self-esteem scale, cognitive needs, and motivation to cooperate with the questionnaire as used by Miura and Kobayashi [11]. Then, pages 8, 10, and 12 contain DQS questions, and pages 9 and 11 contain ARS questions. On pages 15–17, a simple open-ended question was included. After page 7, which contains the main content, the "required answer" setting was turned off. No number of characters was specified for the open-ended question format.

### C. EXPERIMENTAL PROCEDURE

For this experiment, we used Yahoo! crowdsourcing as an environment because it is used to conduct actual online questionnaires. The sequence of our experiment is shown in Figure 5. The questionnaire was conducted in Japanese, and all participants were allowed to respond only once without filtering by platform indicators (e.g., a blacklist was employed). Respondents were rewarded with points worth 5 yen by completing this experiment.[7]

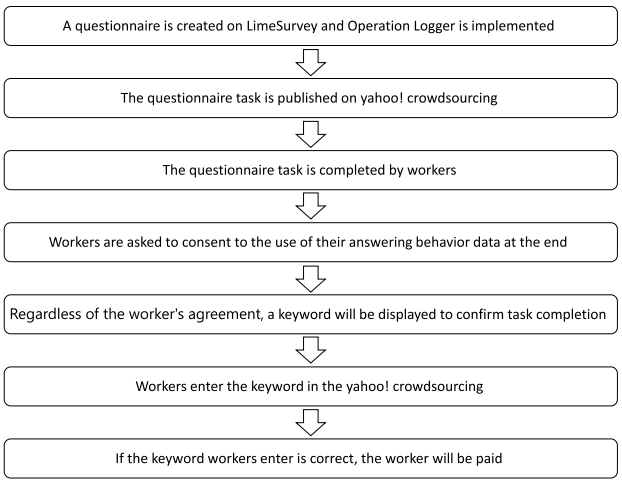



**FIGURE 5. Experiment flow.**

The method proposed in this paper can be applied not only to smartphones but also to tablets and laptops as long as they are equipped with a touch screen. However, when the screen size differs more than a certain level, the placement of questions and choices on the screen change drastically. This will greatly affect the distribution of answering behavior. In this experiment, we decided to limit the devices used for answering to smartphones to suppress the variation in feature distributions due to the response environment. Although a variety of smartphone models have different screen sizes, we did not set any particular model restrictions from the viewpoint of practicality. To limit the use of devices other than smartphones, we included instructions in the task description section on the crowdsourcing site and on the start screen of the questionnaire to encourage users to answer using a smartphone. The page displayed after the questionnaire was completed presented a question asking for consent for the use of answering behavior data. If the respondent did not give consent, the respondent was informed that the data would be deleted and not used, but no compensation would be paid.

[7]The reward of 5 yen (about 5 cents) seems extremely cheap, however, 5,692 people were actually attracted to this task within a month.

This study was approved by the Ethical Review Committee for Research on Human respondents, Nara Institute of Science and Technology (Approval No. 2020-I-2).

## V. CARELESS RESPONSE DETECTION

Section V-A describes the definition of ''careless'' responses as ground truth and describes the classification models and features. Section V-B provides an overview of the questionnaire responses and evaluates careless response detection.

### A. MACHINE LEARNING MODELS

In this paper, we define ''careless'' responses as responses that both indicators (ARS and DQS) register as satisficing. Although the purpose of this paper is to propose a method for detecting careless responses not using any satisficing indicator based on screening questions, we used these indices as ground truth to evaluate the proposed method. The reason why we used two indicators rather than one of them is that ARS and DQS are indicators that detect satisficing from different perspectives.

**TABLE 2.** Hyperparameters for each models.

| Model | Parameter | Value | |
|---|---|---|---|
| | | Basic | Improved |
| LightGBM | lambda_l1 | 0.9977 | 0.003291 |
| | lambda_l2 | 0.1922 | 8.273 |
| | num_leaves | 148 | 75 |
| | feature_fraction | 0.7823 | 0.4974 |
| | bagging_fraction | 0.6284 | 0.8205 |
| | bagging_freq | 7 | 3 |
| | min_child_samples | 12 | 67 |
| Random Forest | bootstrap | True | True |
| | max_depth | 41 | 100 |
| | max_leaf_nodes | 16 | 15 |
| | n_estimators | 73 | 62 |
| | min_samples_split | 4 | 4 |
| | min_samples_leaf | 3 | 7 |
| SVM | kernel | rbf | rbf |
| | C | 5.505 | 2.383 |
| | gamma | 0.001231 | 1.066 |

We trained LightGBM [31], Random Forest [32], and Support vector machine (SVM) [33] as supervised classification models to detect careless response. For tuning the hyperparameters, we used Optuna [34], an automatic optimization tool based on the Bayesian optimization algorithm. The values of tuned hyperparameters for each model are shown in Table 2. We constructed offline training models under the computational environment shown in Table 3. The required computation time (from training to testing) under this condition is shown in Table 4. In our method, the set of features input to the detection models is nearly constant regardless of the content or volume of the questionnaire. Therefore, the computational time mostly depends on the computing environment, the number of respondents and the question format included in the questionnaire. In this paper, we used data collected from the questionnaire including both questions Likert scale and open-ended formats. However, if there are some questionnaires consisting of only a single format of those, the number of features would decrease and the computation time could be decreased accordingly. Otherwise, there would be a possibility of detection accuracy decrease. The following discussion explains how the ML models were constructed in two steps: feature generation and feature selection.

**TABLE 3.** Computational environment.

| Name | Description |
|---|---|
| Device | Surface Pro 4 |
| OS | Microsoft Windows 10 Pro (10.0.18363) |
| CPU | Intel Core i5 (2.4GHz) |
| Memory | 8 GB of RAM |

**TABLE 4.** Calculation time*[a].

| LightGBM | | Random Forest | | SVM | |
|---|---|---|---|---|---|
| Basic | Improved | Basic | Improved | Basic | Improved |
| 34 min | 32 min | 122 min | 110 min | 12 min | 10 min |

*[a] ''Basic'' and ''improved'' mean the type of models described in section V-B.

In the first step, in addition to the features shown in Table 1, the features defined in Table 6 were generated and the models were constructed. The coefficient of variation and the inter- and intra-subject deviation of some features listed below were added as features. The coefficient of variation was obtained by dividing the standard deviation by the mean value for each respondent. This was used instead of the standard deviation to absorb individual differences in the mean values for each respondent. The inter-subject deviation is the difference between the value of a respondent and the mean of all respondents. This represents the difference from the average answering behavior of the entire respondent population, and we thought it would contribute to the detection of careless responses. The intra-subject deviation is the difference between the baseline measured on pages 5 and 6 and the mean of all pages. The reason for adding this deviation is that careless responses increase as the respondents proceed through the questionnaire [14]. Based on this fact, we assumed that the feature difference between the beginning and end of the questionnaire might contribute to the detection of careless responses.

- Coefficient of variation: scrolling length, scrolling duration, scrolling speed, and answering duration (Likert scale)
- Inter-subject deviation: changing Likert scale option, deleting text, number of characters, scrolling length, scrolling speed, and reverse scrolling
- Intra-subject deviation: deleting text, deleting text rate, and scrolling speed.

In the second step, feature selection was applied based on the correlation coefficient between the features and the contribution of features on the LightGBM model (this model is called the ''improved model''). Figure 6 shows a heat map of correlation coefficients between features. Red and blue indicate positive and negative correlation, respectively, and the intensity of the color indicates the strength of the correlation.

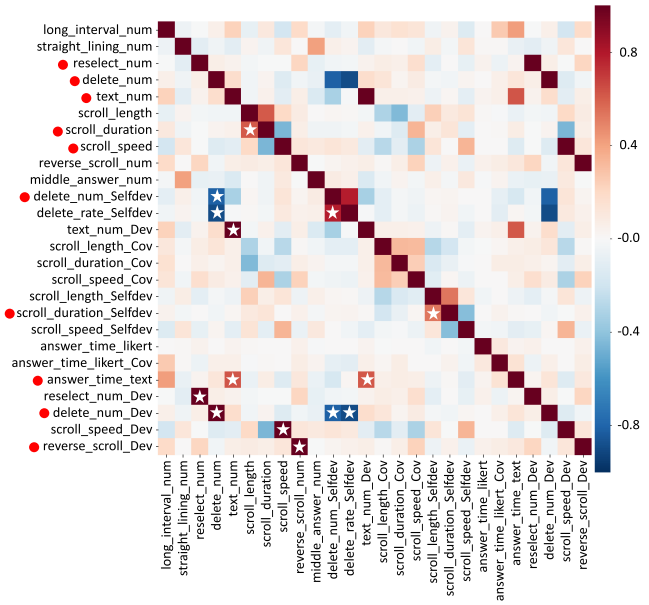



**FIGURE 6.** Correlation between features. White stars (⋆) in the figure indicate feature pairs with a correlation coefficient of 0.5 or higher picked up as candidates for deletion. Red circles (●) indicate the features that were excluded from the models during the feature selection step.

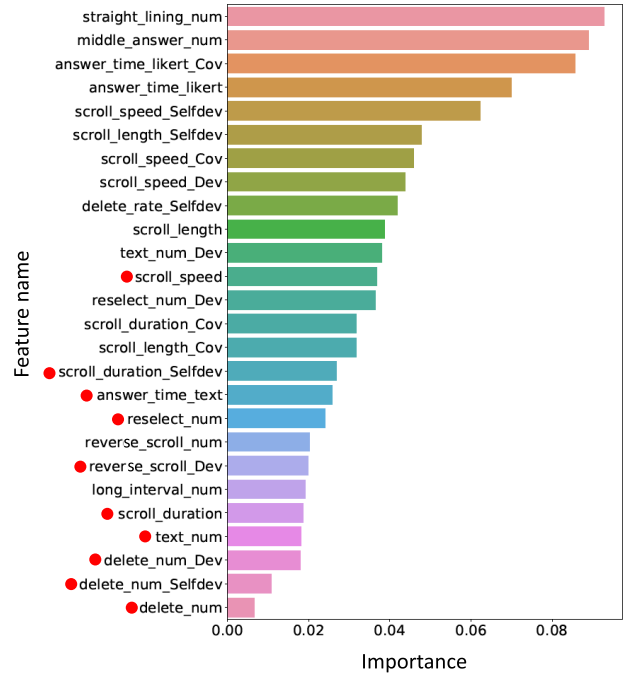



**FIGURE 7.** Feature importance in the LightGBM model constructed by using all features proposed in this paper. Red circles (●) in the figure indicate the features that were excluded from the model during the feature selection step.

Figure 7 shows the importance of the features calculated in the Lightgbm model. This is calculated using the Gini impurity according to how much of the target has not been classified for each node in a decision tree-based ML model. In other words, this indicates how much a node, which divides the tree according to feature, contributes to the classification of the target in a decision tree-based ML model. Because each node splits the tree based on a feature, the importance represents how much the feature contributes to the classification.

**TABLE 5.** Distribution of satisficing responses.

| Indicator | Label | Number of persons | % |
|---|---|---|---|
| DQS | Not satisficing | 4,520 | 91 |
| | Satisficing | 420 | 9 |
| ARS | Not satisficing | 4,066 | 82 |
| | Satisficing | 874 | 18 |
| Total | Careful | 4,693 | 95 |
| | Careless | 247 | 5 |

Feature pairs with a correlation coefficient of 0.5 or higher were picked up as candidates for deletion as indicated by white stars (⋆) in Figure 6. The paired features were compared in feature importance as shown in Figure 7, and the features with lower importance were excluded. Finally, the dimension of feature sets was reduced to 16. The excluded features are marked with a red circle (●) in Figures 6 and 7. Here, it can be confirmed that the excluded features are absolute rather than relative. Comparing the features of the same type of answering behavior, the pure absolute features tended to be of lower importance. These indicate that the relative features, such as inter- and intra-subject, and coefficient of variation contribute more to the classification than the absolute features.

Figure 7 suggests, the features additionally employed in this paper (scrolling speed, scrolling length, and deleting text rate) mark high importance, in addition to traditionally devised features such as straight-lining, middle answer, and answering duration (Likert scale). On the other hand, the response operations such as reverse scrolling and long interval were of low importance.

### B. RESULTS

We used data from 4,940 out of 5,692 respondents, who consented to the use of their answering behavior for this study. We assigned satisficing labels with both DQS and ARS criteria as described in Section V-A. The number and percentage of samples corresponding to each label are shown in Table 5. The results showed 247 "careless responses" and 4,693 "careful responses," as shown in the "total" column in Table 5. The average and standard deviation of each feature are shown in Table 6.

Accuracy, precision, recall, and F1 score were calculated to evaluate the models. Here, because the purpose of this research is to detect careless responses, we focus on recall, which represents the detection rate. The generalization performance of the models was verified by leave-one-out cross-validation, in which only one target sample is used for testing. Because the ratio of positive to negative examples was unbalanced in the data examined in this study, the negative examples were randomly downsampled as shown in Figure 8, and the evaluation was performed with positive and negative samples having the same ratio. Therefore, a dataset generated by single downsampling had 494 samples, and 494 times of leave-one-out cross-validation were performed on the dataset, and the metrics of each result were calculated by the average

**TABLE 6.** Description of features.

| Feature | Description | Unit | Mean | Standard deviation | Magnitude relation[*a] | Formula |
|---|---|---|---|---|---|---|
| answer_time_likert | Mean answering duration (Likert scale) | sec | $2.68 \times 10^{6}$ | $1.88 \times 10^{8}$ | Low | $\bar{t}^{q\_lik} = \sum_{qid=1}^{n^{q\_lik}} t_{qid}/n^{q\_lik}$ |
| answer_time_likert_Cov | Coefficient of variation of answering duration (Likert scale) | – | 2.61 | 1.17 | High | $\sqrt{\frac{1}{n^{q\_lik}} \sum_{qid=1}^{n^{q\_lik}} \left(t_{qid}^{q\_lik} - \bar{t}^{q\_lik}\right)^2} / \bar{t}^{q\_lik}$ |
| answer_time_text | Mean answering duration (Open-ended) | sec | 31,738.10 | 39,305.45 | Low | $\sum_{qid=1}^{n^{q\_text}} t_{qid}^{q\_text}/n^{q\_text}$ |
| reselect_num | Mean counts of changing Likert scale option | count | 10.56 | 7.77 | Low | $\bar{rn} = \sum n^{reselect}$ |
| reselect_num_Dev | Inter-subject deviation of reselect_num | count | $-2.11 \times 10^{-16}$ | 7.77 | Low | ${}_A\bar{rn} - \left(\sum_{uid=1}^{n^{user}} {}_{uid}\bar{rn}/n^{user}\right)$ |
| delete_num | Mean counts of deleting text | count | 7.71 | 12.32 | Low | $\bar{dn} = \sum n^{delete}$ |
| delete_num_Selfdev | Intra-subject deviation of delete_num | count | $-41.27$ | 97.10 | High | ${}_A\bar{dn} - {}_A^p\bar{dn}|_{5<p<6}$[*c] |
| delete_rate_Selfdev | Intra-subject deviation of deleting text rate | count/character | $-1.24$ | 2.63 | High | ${}_A\bar{dn}/{}_An^{char} - {}_A^p\bar{dn}/{}_A^pn^{char}|_{5<p<6}$ |
| scroll_length | Mean scrolling length | px | 118.84 | 62.60 | High | $\bar{sl} = \sum_{i=1}^{n^{scroll}} sl_i/n^{scroll}$ |
| scroll_length_Cov | Coefficient of variation of scrolling length | – | 0.431 | 0.129 | Low | $\sqrt{\frac{1}{n^{scroll}} \sum_{i=1}^{n^l} \left(sl_i - \bar{sl}\right)^2} / \bar{sl}$ |
| scroll_length_Selfdev | Intra-subject deviation of scroll_length | px | $-12.67$ | 42.28 | High | ${}_A^p\bar{sl} - {}_A\bar{sl}|_{5<p<6}$ |
| scroll_duration | Mean scrolling duration | sec | 332.18 | 209.98 | Low | $\bar{sd} = \sum_{i=1}^{n^{scroll}} sd_i/n^{scroll}$ |
| scroll_duration_Cov | Coefficient of variation of scrolling duration | – | 0.891 | 0.490 | High | $\sqrt{\frac{1}{n^{scroll}} \sum_{i=1}^{n^l} \left(sd_i - \bar{sd}\right)^2} / \bar{sd}$ |
| scroll_duration_Selfdev | Intra-subject deviation of scroll_duration | sec | $-41.70$ | 169.27 | Low | ${}_A\bar{sd} - {}_A^p\bar{sd}|_{5<p<6}$ |
| scroll_speed | Mean scrolling speed | px/s | 513.10 | 168.51 | High | $\bar{ss} = \sum_{i=1}^{n^{scroll}} sl_i/sd_i$ |
| scroll_speed_Cov | Coefficient of variation of scrolling speed | – | 0.471 | 0.155 | Low | $\sqrt{\frac{1}{n^{scroll}} \sum_{i=1}^{n^l} \left(ss_i - \bar{ss}\right)^2} / \bar{ss}$ |
| scroll_speed_Dev | Inter-subject deviation of scroll_speed | px/s | $2.98 \times 10^{-14}$ | 168.33 | High | ${}_A\bar{ss} - \left(\sum_{uid=1}^{n^{user}} {}_{uid}\bar{ss}/n^{user}\right)$ |
| scroll_speed_Selfdev | Intra-subject deviation of scroll_speed | px/s | 27.15 | 117.30 | High | ${}_A\bar{ss} - {}_A^p\bar{ss}|_{5<p<6}$ |
| reverse_scroll_num | Mean counts of reverse scrolling | count | 2.01 | 2.98 | High | $\bar{rs} = \sum n^{rev}$ |
| reverse_scroll_num_Dev | Inter-subject deviation of reverse_scroll_num | count | $2.26 \times 10^{-15}$ | 2.98 | High | ${}_A\bar{rs} - \left(\sum_{uid=1}^{n^{user}} {}_{uid}\bar{rs}/n^{user}\right)$ |
| long_interval_num | Counts of long interval | count | 4.86 | 4.47 | Low | $\sum_{i=1}^{n^{interval}} i$ |
| straight_lining_num | Maximum counts of straight-lining | count | 4.77 | 5.44 | High | $\max(\mathrm{n}^{lining})$[*b] |
| middle_answer_num | Counts of middle answer | questions | 30.53 | 16.34 | High | $\sum_{i=1}^{n^{middle}} i$ |
| text_num | Mean number of characters | characters | 12.29 | 10.15 | Low | $\bar{tn} = \sum_{qid=1}^{n^{q\_text}} n_{qid}^{char}/n^{q\_text}$ |
| text_num_Dev | Inter-subject deviation of text_num | characters | $-1.15 \times 10^{-14}$ | 10.10 | Low | ${}_A\bar{tn} - \left(\sum_{uid=1}^{n^{user}} {}_{uid}\bar{tn}/n^{user}\right)$ |

[*a] This column indicates whether the mean of the satisficing group is higher (High) or lower (Low) than that of the not satisficing group.
[*b] The function "max()" is assumed that it can extract maximum number.
[*c] The condition of $5 < p < 6$ means that the variables are baseline, which measured in the questionnaire pages for baseline measurement.

*Note.* $t$: The response duration; $n^{qlikert}$: The number of Likert scale questions; $qid$: The id of questions; $n^{qtext}$: The number of open-ended questions; $n^{reselect}$: The number of reselection about Likert scale; $A$: An arbitrary respondent; $n^{user}$: The number of respondents; $uid$: The id of respondents; $n^{delete}$: The number of text deletion; $p$: The page number of the questionnaire; $n^{char}$: The number of characters; $sl$: The length of scrolling; $n^{scroll}$: The number of scrolling; $sd$: The duration of scrolling; $ss$: The speed of scrolling; $n^{rev}$: The number of reverse scrolling; $n^{interval}$: The number of long interval; $n^{lining}$: The number of straight-lining; $n^{middle}$: The number of middle answer;

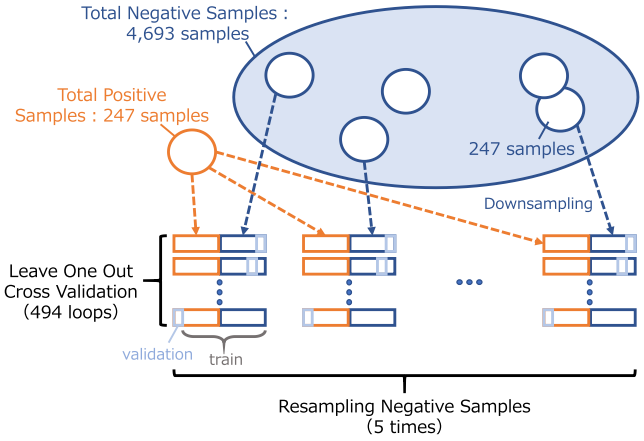



**FIGURE 8.** How data were separated for random downsampling and leave-one-out cross-validation.

of the 494 times. Furthermore, to evaluate the generalization performance, the down-sampling was done five times and the metrics of each result were calculated as the average value.

The evaluation results of the detection models proposed in Section V-B are shown in Table 7. For comparison of the detection rates, the results of the model reported by Ozaki and Suzuki [6] and the models using 16 features except for inter- and intra-subject deviation, which have "Dev" or "Selfdev" in the suffix of the basic feature name, are also shown. Although the definitions of the target labels in Ozaki *et al.* are not completely identical to those for the labels in this paper, they also used DQS questions and questions checking contradiction to define target labels. Because the properties of these satisficing detection questions are the same as those used in this paper, we consider that the effect of the difference of evaluation methods on the results is small.

The improved LightGBM model achieved the highest Recall among our proposed models. In particular, the noteworthy Recall was improved by 1% from the basic model, finally, it shows 85.9%. Compared to Ozaki *et al.*, we confirmed that this study succeeded in greatly improving the detection rate for careless responses (from 55.6% to 85.9%) by using answering behavior. On the other hand, Random Forest showed better performance than LightGBM for all indices except for Recall. Low precision in this classification problem means that there is a high probability of incorrectly classifying careful responses in careless cases, which leads to a waste of effort or money spent to collect data. In addition, excluding false-positive samples undermines the external validity of the data. Precision by improved Random Forest was improved by 1.5% rather than the basic model, reaching 87.3%. As for SVM, the results showed a significant decrease in accuracy compared to LightGBM and Random Forest. Furthermore, the detection rate of the improved model was higher than that of the basic model throughout all models and indices except for Recall of SVM. This indicates that the features of inter- and intra-subject deviation, which are the points that we devised, contributed to the improvement of the detection rate.

In addition, to verify the relationship between the number of questions and the quality of the data, the number of pages for feature generation (excluding the number of pages used for the baseline calculation) was set to 3 pages (average 17 questions) and 9 pages (90 questions). The resulting detection rates were 79.7% and 80.9%, respectively. Considering the detection rate of 85.9% for all pages (17 pages, 128 questions), this suggests that as the number of questions increases, the obtained quality of answering behavior also improves, along with the detection rate. However, the detection rate was about 80% even with a questionnaire of about 17 questions (3 pages).

The results for the challenges of this study described in Section II are as follows. In response to Challenge 1, we realized the recording of answering behavior with a plug-in that can be easily added to an existing questionnaire system. We actually recorded the desired data, so we can say that Challenge 1 was accomplished. For Challenge 2, we examined and generated features to express satisficing from the data collected in the experiments using the plug-in, and constructed ML models to detect careless responses. Because we succeeded in significantly improving the previously reported detection rate, we also achieved Challenge 2.

## VI. DISCUSSION

In Section VI-A, we evaluate the robustness of the improved LightGBM model, which had the best Recall, against outliers. Section VI-B describes the findings and usefulness of the features proposed in this paper based on their importance. Finally, Section VI-C describes the limitations of this paper.

### A. ROBUSTNESS AGAINST OUTLIER

Because the LightGBM is based on a decision tree, feature values that trend in the direction of the wrong class will cause false detection. From this perspective, we examine the robustness of the classification model against outliers.

First, we compared the average values of each feature for the positive and negative cases and clarified the magnitude relationship shown in the "magnitude relation" column in Table 6. Next, we divided the positive samples into true positives and false negatives, and calculated the quartile range for each feature in each group. The number of samples that exceeded this range and trended toward false detection, that is, in a direction opposite to the magnitude relationship shown in Table 6, was calculated and is shown in Figure 9. The horizontal axis shows the number of features that exceeded the quartile range in the direction leading to false detection. The bar plot shows the number of samples and the line plot shows the detection rate for careless responses. The number of samples is small in the range where the horizontal axis is 10 or more, so we mainly focus on the range below 9. False detection begins to occur when the value is higher than 2, but when the value is higher than 8, the detection rate is about 50%. In contrast, the model maintains a detection rate of 80% in the range of 6 or less, which accounts for about 85.7% of

**TABLE 7.** Evaluation results of ML models.

| Metrics | Existing | Proposed | | | | | |
|---|---|---|---|---|---|---|---|
| | Ozaki *et al.* [6]*a | LightGBM | | Random Forest | | SVM | |
| | | Basic | Improved | Basic | Improved | Basic | Improved |
| Accuracy | – | 0.844 | 0.862 | 0.852 | 0.865 | 0.690 | 0.736 |
| Precision | – | 0.841 | 0.864 | 0.858 | 0.873 | 0.691 | 0.823 |
| **Recall** | **0.556** | **0.849** | **0.859** | **0.844** | **0.854** | **0.690** | **0.603** |
| F1 Score | – | 0.845 | 0.862 | 0.851 | 0.863 | 0.690 | 0.696 |

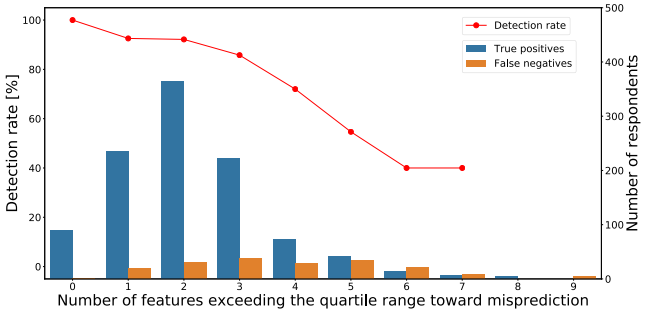



**FIGURE 9.** Careless response detection rate for the improved LightGBM model evaluated in terms of the number of features that exceeded the quartile range toward misprediction.

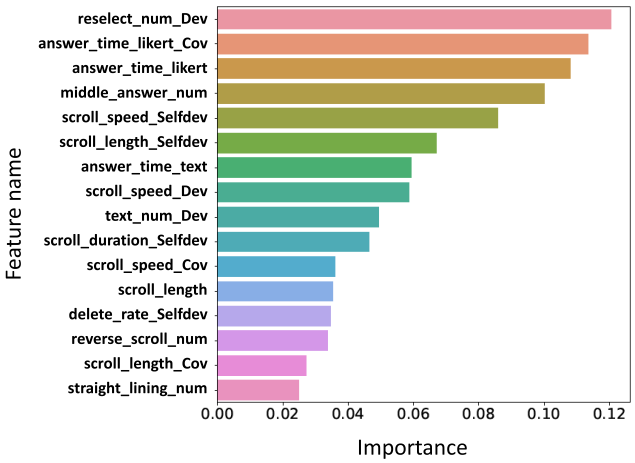



**FIGURE 10.** Feature importance in the improved LightGBM model.

the positive cases, so it can be said that the model is somewhat robust concerning the features used in this paper.

### B. FEATURE IMPORTANCE

First, we discuss Figure 7, which shows the contribution of features in the LightGBM model constructed using all features proposed in this paper. The previously existing features, such as the maximum number of straight-lining responses, the number of intermediate responses, and the answering duration (Likert scale) contributed the best. Of the 13 answering behaviors shown in Table 1, 6 absolute features were excluded in the improved LightGBM model, indicating that the contribution of inter- and intra-subject deviation, and coefficient of variation features with the suffix of _Dev, _Selfdev, and _Cov were high. This indicates that deviation and variability are good representations of careless responses.

As for the answering duration, the contribution rate for the Likert scale format was more than that for the open-ended format. However, the coefficient of variation may not be evaluated equally because only 4 questions employed the open-ended format versus 124 questions that employed the Likert scale. Also, the contribution rate for the variation in response duration was higher than that of response duration.

Among the features proposed in this paper, the contribution rate for the features related to scrolling, especially scrolling speed were high. However, the contribution rate for features related to the deletion of text were low. The reason for this is also attributed to the dearth of open-ended questions. The answering behavior such as reverse scrolling and long interval were also of low importance.

Next, the contribution of the improved LightGBM model is shown in Figure 10. Overall, the ranking relationship of contribution rates remained the same as that in the model constructed by using all features shown in Figure 7. On the other hand, the contribution rate for the feature related to reselection of the Likert scale is the highest. Moreover, the features relevant to scrolling (speed and length) also contribute to the detection of careless responses. These suggest that careless responses are well represented by reselecting and scrolling.

### C. LIMITATIONS

In this study, we evaluated the data obtained from a questionnaire that combines the Likert scale and open-ended questions. Thus, there is a possibility that appropriate detection may not be possible in different questionnaire environments (e.g., when drop-down formats are included or when only open-ended formats are used).

The inclusion of screening questions such as ARS and DQS may have increased the frequency of satisficing compared to the absence of these questions. Investigating this effect is one of the future tasks.

Moreover, because Operation Logger sends the answering behavior data to the server in addition to the answer results, it increases the server load when a large number of people answer at the same time. Hence, when conducting large-scale questionnaires (e.g., through crowdsourcing), it is necessary to pay attention to the number of simultaneous respondents, especially immediately after recruiting respondents.

Furthermore, in recent years, the handling of personal data has been discussed mainly with regard to European

Union regulations.[8] Therefore, when implementing the method proposed in this paper, it is necessary to account for the rules regarding personal data in each country.

## VII. CONCLUSION

In this paper, we aimed to detect careless responses, which are an attempt to complete questionnaires as easily and quickly as possible, with high accuracy in an environment that does not place a psychological burden on the respondent due to screening questions. We developed the Operation Logger plug-in to record the answering behavior and actually collected the desired data in a large-scale experiment using this plug-in, so we can say that Challenge 1, the recording of answer behavior, was accomplished. Using the data collected in the experiment, we constructed the ML-ABS method to detect careless responses. The results showed that the detection rate was much higher than that reported by Ozaki and Suzuki [6], who worked on a similar task. This indicates that Challenge 2, the detection of careless responses with high accuracy, was also accomplished. The method proposed in this paper was able to detect careless responses with high accuracy without adding the conventional screening questions. In addition, since careless responses can be detected automatically by the detection models, it is expected to allow survey administrators to avoid manual screening for careless responses.

The effect of the different screen sizes of smartphones and tablets on the response results has also been discussed previously [21]. Although our method is limited to smartphones, it is worth noting that we were able to achieve a high detection rate without normalizing for differences in screen size. Furthermore, as a result of examining the relationship between the number of questions and the detection rate, it became clear that the detection rate improved with the number of questions (number of pages). However, the detection rate did not decrease significantly even when the number of questions was small. Based on these results, we believe that the system is robust against differences in the screen size and number of questions.

Regarding the newly proposed features, the contribution rates of features concerning reselection of the Likert scale and scrolling (speed and length) were high. For the other response behaviors, intra-subject deviations tended to make a larger contribution than pure absolute features (i.e., raw counts). This result implies the possibility that careless respondents make a measurable change within their answering behavior. On the other hand, for the existing features, it was confirmed that the previously proposed answering duration, straight-lining, and middle answer features had high contribution rates. Concerning the answering duration, we found that its variation may better represent inappropriate responses than the raw answering duration feature.

To apply our careless responses detection method to practical problems, survey administrators need to follow the next 3 steps. First, the Operation Logger needs to be set on the LimeSurvey questionnaire. Second, the baseline questions have to be inserted at the beginning of the questionnaire. Third, the extracted features about response results and answering behavior have to be input to the careless response detection model proposed in this paper.

Therefore, it is necessary to examine whether the model can be applied to questionnaires of different fields and contents as one of the future works. In this paper, we dealt with questionnaires that investigate psychological states, but it will be interesting to see whether the model can be applied to questionnaires in the field of marketing, such as reviews of some products and services.

In addition, there is another issue of how to handle response data detected as a careless response. Excluding careless responses from the sample increases the internal validity of

[8]https://gdpr.eu/cookies/

**TABLE 8. Questionnaire contents (Page 1–9).**

| Page | Q No.[*a] | Question | Type [*b] |
|---|---|---|---|
| 1 | 1 | Please answer your sex. | Man or woman |
| 2 | 1 | Please answer your age. | Open-ended |
| 3 | 1 | Please answer your occupation. | Dropdown |
| 4 | H[*a] | Please copy and paste the text shown below into the answer field. | |
| | 1 | (Auto-generated unique respondent ID) | Open-ended |
| 5 | H | For the following items, please choose the one that best applies to you. | |
| | 1 | Please choose "Disagree." | 5-point LS[*b] |
| | 2 | Please choose "Agree." | 5-point LS |
| | 3 | Please choose "Neither agree nor disagree." | 5-point LS |
| | 4 | Please choose "Strongly disagree." | 5-point LS |
| | 5 | Please choose "Disagree." | 5-point LS |
| | 6 | Please choose "Neither agree nor disagree." | 5-point LS |
| | 7 | Please do not choose any option. | 5-point LS |
| 6 | H | This page is for operational verification. Please type the following text into the answer field by hand, but do not copy and paste it. | |
| | 1 | "I am a cat. I don't have a name yet. I have no idea where I was born. All I remember is that I was meowing in a dark, dank place. This was the first time I saw a human being." (This is a passage from a famous Japanese novel named "I am a cat" wrote by Soseki Natsume.) | Open-ended |
| 7 | H | For the following items, please choose the one that best applies to you. | |
| | 1 | I prefer complex problems to simple ones. | 5-point LS |
| | 2 | I tend to think long and hard about problems, even if they are not directly related to me. | 5-point LS |
| | 3 | I like to work hard to come up with new ways to solve problems. | 5-point LS |
| | 4 | I don't like to think about what I should be doing with my life. | 5-point LS |
| | 5 | I find it satisfying to think hard for hours on end. | 5-point LS |
| | 6 | I do not enjoy the process of slow thinking. | 5-point LS |
| | 7 | I would rather do something that requires less thinking than something that requires more thinking. | 5-point LS |
| | 8 | I prefer to think about small, everyday plans rather than long-term plans. | 5-point LS |
| 8 | H | What are your reasons for participating in this survey? For the following items, please choose the one that best applies to you. | |
| | 1 | To cooperate in academic research. | 5-point LS |
| | 2 | To cooperate in academic research. | 5-point LS |
| | 3 | Because it sounds interesting. | 5-point LS |
| | 4 | Please choose "Agree." | 5-point LS |
| | 5 | Because I am interested in the research topic. | 5-point LS |
| | 6 | To receive a reward. | 5-point LS |
| | 7 | To pass the time. | 5-point LS |
| | 8 | Because I happen to have time. | 5-point LS |
| 9 | H | Here are some questions about your mindset. For the following items, please choose the one that best applies to you. | |
| | 1 | I am a happy person. | 5-point LS |
| | 2 | I like children. | 5-point LS |
| | 3 | I am always worried. | 5-point LS |
| | 4 | I do not want to get a speeding ticket. | 5-point LS |
| | 5 | I cannot wait for the holidays. | 5-point LS |
| | 6 | I have a good exercise routine. | 5-point LS |
| | 7 | I spend my free time relaxing. | 5-point LS |
| | 8 | I get irritated when people wait for me. | 5-point LS |
| | 9 | I like the music of Marlene Sandersfield. | 5-point LS |
| | 10 | I am a very energetic person. | 5-point LS |
| | 11 | I have a memorable place. | 5-point LS |
| | 12 | I am not afraid to open up to my friends. | 5-point LS |
| | 13 | I eat breakfast every day. | 5-point LS |
| | 14 | My favorite subject is Noh. | 5-point LS |
| | 15 | I sometimes feel annoyed with others. | 5-point LS |
| | 16 | I don't like to be ridiculed or embarrassed. | 5-point LS |
| | 17 | I am a very compassionate person. | 5-point LS |
| | 18 | I take a long bath. | 5-point LS |
| | 19 | I am an active person. | 5-point LS |
| | 20 | I like to pay taxes. | 5-point LS |
| | 21 | I am an active person. | 5-point LS |
| | 22 | I enjoy the company of my friends. | 5-point LS |

[*a] In the "Q no." column, the instructions for the question in the page are shown in the line represented by "H" as Header.
[*b] In the "Type" column, "Likert scale" is abbriviated as "LS."

**TABLE 9. Questionnaire contents (Page 10–16).**

| Page | Q No. | Question | Type |
|---|---|---|---|
| 10 | H | Here are some questions about your personality. For the following items, please choose the one that best applies to you. | |
| | 1 | I am a talkative person. | 5-point LS |
| | 2 | I am a quiet person. | 5-point LS |
| | 3 | I think i am cheerful. | 5-point LS |
| | 4 | I am prone to worry. | 5-point LS |
| | 5 | I get anxious easily. | 5-point LS |
| | 6 | I think I am a worrier. | 5-point LS |
| | 7 | I am a creative person. | 5-point LS |
| | 8 | I am a versatile person. | 5-point LS |
| | 9 | I am a progressive thinker. | 5-point LS |
| | 10 | I am lazy most of the time. | 5-point LS |
| | 11 | Please do not choose any of the options. | 5-point LS |
| | 12 | I lead a loose life. | 5-point LS |
| | 13 | I am a lazy person. | 5-point LS |
| | 14 | I am usually short-tempered. | 5-point LS |
| | 15 | I am a mild-mannered person. | 5-point LS |
| | 16 | I think I have an angry personality. | 5-point LS |
| 11 | H | Here are some questions about your mindset. For the following items, please choose the one that best applies to you. | |
| | 1 | I lead an active life. | 5-point LS |
| | 2 | I like to spend time with my friends. | 5-point LS |
| | 3 | I am interested in politics. | 5-point LS |
| | 4 | It feels good to be appreciated. | 5-point LS |
| | 5 | I like to receive sales calls. | 5-point LS |
| | 6 | It is not difficult for me to confide something important to my friends. | 5-point LS |
| | 7 | I always try to be considerate of others. | 5-point LS |
| | 8 | I save money in a systematic way. | 5-point LS |
| | 9 | I would rather people hate me than like me. | 5-point LS |
| | 10 | I have a good memory. | 5-point LS |
| | 11 | I have a lot of things to worry about. | 5-point LS |
| | 12 | I get irritated with people sometimes. | 5-point LS |
| | 13 | I am looking for a new hobby. | 5-point LS |
| | 14 | I think I would be happy if I won the lottery. | 5-point LS |
| | 15 | I am happy most of the time. | 5-point LS |
| | 16 | I have a habit of reading books. | 5-point LS |
| | 17 | My hobbies are coin collecting and creative dancing. | 5-point LS |
| | 18 | I have a lot of energy. | 5-point LS |
| | 19 | I am quite active. | 5-point LS |
| | 20 | I am uncomfortable when people are late for appointments. | 5-point LS |
| | 21 | I like to relax in my free time. | 5-point LS |
| 12 | H | Here are some questions about your personality. For the following items, please choose the one that best applies to you. | |
| | 1 | I think I am a diplomatic person. | 5-point LS |
| | 2 | I am a dark person. | 5-point LS |
| | 3 | I am usually unsociable. | 5-point LS |
| | 4 | I am a difficult person to deal with. | 5-point LS |
| | 5 | I am weak most of the time. | 5-point LS |
| | 6 | I am vulnerable. | 5-point LS |
| | 7 | I am insightful. | 5-point LS |
| | 8 | I am imaginative. | 5-point LS |
| | 9 | I have a keen sense of beauty. | 5-point LS |
| | 10 | I live a spontaneous life. | 5-point LS |
| | 11 | I think I am a sluggish person. | 5-point LS |
| | 12 | I lead a well-planned life. | 5-point LS |
| | 13 | I think I am a generous person. | 5-point LS |
| | 14 | Please choose the option at the bottom. | 5-point LS |
| | 15 | I am kind most of the time. | 5-point LS |
| | 16 | I am a conscientious person. | 5-point LS |
| 13 | H | For the following items, please choose the one that best applies to you. Please tell us how you feel about yourself, not how others see you. | |
| | 1 | I am at least as worthy as others. | 5-point LS |
| | 2 | I have many good qualities. | 5-point LS |
| | 3 | I often feel like a loser. | 5-point LS |
| | 4 | I can do things as well as others. | 5-point LS |
| | 5 | I am not very proud of myself. | 5-point LS |
| | 6 | I am positive about myself. | 5-point LS |
| | 7 | I am satisfied with myself most of the time. | 5-point LS |
| | 8 | I have a vague sense of confidence. | 5-point LS |
| | 9 | I would like to be able to respect myself more. | 5-point LS |
| | 10 | I sometimes feel that I am completely useless. | 5-point LS |
| | 11 | I feel that I am useless in many ways. | 5-point LS |
| 14 | H | What do you usually do when you answer these surveys? For the following items, please choose the one that best applies to you. | |
| | 1 | I read the questionnaire carefully. | 5-point LS |
| | 2 | I pay attention to all the questions. | 5-point LS |
| | 3 | Take enough time to answer the questions honestly. | 5-point LS |
| | 4 | I answer in a pattern. | 5-point LS |
| | 5 | I give the same answer to the same set of questions. | 5-point LS |
| | 6 | I think one at a time, even with a group of questions. | 5-point LS |
| | 7 | I try to answer quickly and without much thought. | 5-point LS |
| | 8 | I answer with an idea without thinking it through. | 5-point LS |
| | 9 | I tries to finish the answer as quickly as possible. | 5-point LS |
| | 10 | I quickly skim the question text. | 5-point LS |
| | 11 | I try to understand the meaning of the question text. | 5-point LS |
| 15 | 1 | Please describe the most memorable event of yesterday. | Open-ended |
| 16 | 2 | Please describe one recent news item. | Open-ended |
| 17 | 3 | Please enter your goals for the current year. | Open-ended |
| 18 | 1 | In this questionnaire, the log data of the screen operation at the time of answering was recorded along with the answer result. We are investigating the reliability of the online questionnaire using the screen operation data. If you understand that the data will be used only for this research, and if you agree to the use of the screen operation data, please select "I agree". If you do not agree to the use of the screen operation data, please select "I do not agree". In either case, you will be paid. | Agree or disagree |

the survey because this exclusion reduces noise, but it also reduces the diversity of the sample, which may compromise the external validity. To avoid such trade-offs, some believe that it is preferable to have a policy of transforming non-satisficing respondent data to valid data by providing some intervention for satisficing respondents [12], [29]. Oppenheimer *et al.* [2] conducted an experiment in which people who violated the IMC were repeatedly redirected to the same IMC until they cleared it. They reported that this could potentially correct the subsequent survey responses of respondents who violated the IMC the first time. By this work, we will also apply the detection method to develop some kind of real-time intervention technique for correcting careless respondents as future work.

## APPENDIX
## QUESTIONNAIRE CONTENTS

See Tables 8 and 9.

## REFERENCES


[1] J. A. Krosnick, "Response strategies for coping with the cognitive demands of attitude measures in surveys," *Appl. Cognit. Psychol.*, vol. 5, no. 3, pp. 213–236, May 1991.

[2] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *J. Exp. Social Psychol.*, vol. 45, no. 4, pp. 867–872, Jul. 2009.

[3] M. R. Maniaci and R. D. Rogge, "Caring about carelessness: Participant inattention and its effects on research," *J. Res. Personality*, vol. 48, pp. 61–83, Feb. 2014.

[4] A. Miura and T. Kobayashi, "Exploring tips to detect 'satiscing' in an online survey: A study using university student samples," (in Japanese), *Jpn. J. Social Psychol.*, vol. 48, pp. 61–83, Sep. 2016.

[5] W. Pei, A. Mayer, K. Tu, and C. Yue, "Attention please: Your attention check questions in survey studies can be automatically answered," in *Proc. Web Conf.*, Apr. 2020, pp. 1182–1193.

[6] K. Ozaki and T. Suzuki, "Using machine learning to predict inappropriate respondents," (in Japanese), *Kodo Keiryogaku (Jpn. J. Behaviormetrics)*, vol. 46, no. 2, pp. 39–52, 2019.

[7] LimeSurvey GmbH. *LimeSurvey: The Online Survey Tool—Open Source Surveys*. Accessed: Mar. 1, 2021. [Online]. Available: https://www.limesurvey.org/

[8] H. Taherdoost, "Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research," *Int. J. Acad. Res.*, vol. 5, no. 3, pp. 28–36, 2016.

[9] H. A. Simon, "Rational choice and the structure of the environment," *Psychol. Rev.*, vol. 63, no. 2, pp. 129–138, 1956.

[10] A. Miura and T. Kobayashi, "Survey satisficing inflates stereotypical responses in online experiment: The case of immigration study," *Frontiers Psychol.*, vol. 7, p. 1563, Oct. 2016.

[11] A. Miura and T. Kobayashi, "Survey satisficing biases the estimation of moderation effects," *Jpn. Psychol. Res.*, vol. 61, no. 3, pp. 204–210, Jul. 2019.

[12] A. J. Berinsky, M. F. Margolis, and M. W. Sances, "Can we turn shirkers into workers?" *J. Exp. Social Psychol.*, vol. 66, pp. 20–28, Sep. 2016.

[13] D. Steger, U. Schroeders, and T. Gnambs, "A meta-analysis of test scores in proctored and unproctored ability assessments," *Eur. J. Psychol. Assess.*, vol. 36, pp. 174–184, Sep. 2020.

[14] N. A. Bowling, A. M. Gibson, J. W. Houpt, and C. K. Brower, "Will the questions ever end? Person-level increases in careless responding during questionnaire completion," *Organ. Res. Methods*, p. 1, Aug. 2020.

[15] T. R. Kwapil, G. M. Gross, P. J. Silvia, M. L. Raulin, and N. Barrantes-Vidal, "Development and psychometric properties of the multidimensional Schizotypy scale: A new measure for assessing positive, negative, and disorganized Schizotypy," *Schizophrenia Res.*, vol. 193, pp. 209–217, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920996417304024

[16] M. Malesza and M. C. Kaczmarek, "Predictors of anxiety during the COVID-19 pandemic in poland," *Personality Individual Differences*, vol. 170, Feb. 2021, Art. no. 110419. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886920306103

[17] N. Plohl and B. Musil, "Modeling compliance with COVID-19 prevention guidelines: The critical role of trust in science," *Psychol., Health Med.*, vol. 26, no. 1, pp. 1–12, Jan. 2021, doi: 10.1080/13548506.2020.1772988.
[18] W. Rote, M. Olmo, L. Feliscar, M. Jambon, C. Ball, and J. Smetana, "Helicopter parenting and perceived overcontrol by emerging adults: A family-level profile analysis," *J. Child Family Stud.*, vol. 29, no. 11, pp. 3153–3168, 2020.
[19] A. L. Nichols and J. E. Edlund, "Why don't we care more about carelessness? Understanding the causes and consequences of careless participants," *Int. J. Social Res. Methodol.*, vol. 23, no. 6, pp. 625–638, Nov. 2020, doi: 10.1080/13645579.2020.1719618.
[20] U. Schroeders, C. Schmidt, and T. Gnambs, "Detecting careless responding in survey data using stochastic gradient boosting," *PsyArXiv*, Jul. 2020, doi: 10.31234/osf.io/vs37k.
[21] P. Lugtig and V. Toepoel, "The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error," *Social Sci. Comput. Rev.*, vol. 34, no. 1, pp. 78–94, Feb. 2016.
[22] R. Tourangeau, H. Sun, T. Yan, A. Maitland, G. Rivero, and D. Williams, "Web surveys by smartphones and tablets: Effects on data quality," *Social Sci. Comput. Rev.*, vol. 36, no. 5, pp. 542–556, Oct. 2018.
[23] M. Warkentin, S. Goel, and P. Menard, "Shared benefits and information privacy: What determines smart meter technology adoption?" *J. Assoc. Inf. Syst.*, vol. 18, no. 11, p. 3, 2017.
[24] M. Revilla and C. Ochoa, "What are the links in a Web survey among response time, quality, and auto-evaluation of the efforts done?" *Social Sci. Comput. Rev.*, vol. 33, no. 1, pp. 97–114, Feb. 2015.
[25] A. Alsultan, K. Warwick, and H. Wei, "Non-conventional keystroke dynamics for user authentication," *Pattern Recognit. Lett.*, vol. 89, pp. 53–59, Apr. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865517300429
[26] H. Seo, E. Kim, and H. K. Kim, "A novel biometric identification based on a user's input pattern analysis for intelligent mobile devices," *Int. J. Adv. Robot. Syst.*, vol. 9, no. 2, p. 46, Aug. 2012, doi: 10.5772/51319.
[27] Y. Kim, J. Dykema, J. Stevenson, P. Black, and D. P. Moberg, "Straightlining: Overview of measurement, comparison of indicators, and effects in mail–Web mixed-mode surveys," *Social Sci. Comput. Rev.*, vol. 37, no. 2, pp. 214–233, Apr. 2019.
[28] H. Baumgartner and J.-B.-E. M. Steenkamp, "Response styles in marketing research: A cross-national investigation," *J. Marketing Res.*, vol. 38, no. 2, pp. 143–156, May 2001.
[29] E. Anduiza and C. Galais, "Answering without reading: IMCs and strong satisficing in online surveys," *Int. J. Public Opinion*, vol. 29, no. 3, pp. 497–519, May 2016.
[30] D. J. Hauser and N. Schwarz, "It'sa trap! Instructional manipulation checks prompt systematic thinking on 'tricky' tasks," *SAGE Open*, vol. 5, no. 2, pp. 1–6, 2015.
[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 3146–3154.
[32] L. B. Statistics and L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[33] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.
[34] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

**MASAKI GOGAMI** was born in 1997. He received the B.E. degree from the Advanced Course of Kobe City College of Technology, Japan, in 2019. He is currently pursuing the degree with the Nara Institute of Science and Technology, Japan. His research interest includes careless response detection in online surveys, with a focus on answering behavior. He is a member of IPSJ.

**YUKI MATSUDA** (Member, IEEE) was born in 1993. He received the B.E. degree from the Advanced Course of Mechanical and Electronic System Engineering, National Institute of Technology, Akashi College, Japan, in 2015, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016 and 2019, respectively. During his Ph.D. degree, he studied as a Visiting Researcher with Ulm University, Germany, from 2017 to 2018. His current research interests include urban sensing, civic computing, ubiquitous computing, and affective computing. He is a member of IPSJ. He received the IEEE PerCom Best Demonstration Award in 2019.

**YUTAKA ARAKAWA** (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Keio University, Japan, in 2001, 2003, and 2006, respectively. He is currently a Professor with the Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, and a Visiting Professor with the Nara Institute of Science and Technology and Osaka University. His current research interests include human activity recognition, behavior change support systems, and location-based information systems. He is a member of ACM, IPSJ, and IEICE.

**KEIICHI YASUMOTO** (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. He is currently a Professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology. His research interests include distributed systems, mobile computing, and ubiquitous computing. He is a member of ACM, IPSJ, SICE and IEICE.