Towards Hybrid Profiling: combining digital phenotyping with validated survey questions to balance data entry effort with predictive power

Citation for published version (APA):

Hadian Haghighi, E., Nuijten, R. C. Y., & van Gorp, P. M. E. (2021). Towards Hybrid Profiling: combining digital phenotyping with validated survey questions to balance data entry effort with predictive power. In *Proceedings - 2021 International Symposium on Electrical, Electronics and Information Engineering, ISEEIE 2021* (pp. 585-593). (ACM International Conference Proceeding Series). Association for Computing Machinery, Inc. <https://doi.org/10.1145/3459104.3459199>

DOI:

[10.1145/3459104.3459199](https://doi.org/10.1145/3459104.3459199)

Document status and date:

Published: 19/02/2021

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Towards Hybrid Profiling: Combining Digital Phenotyping with Validated Survey Questions to Balance Data Entry Effort with Predictive Power

Ehsan Hadian Haghighi

Department of Industrial Engineering
Eindhoven University of Technology,
The Netherlands
e.hadian.haghighi@tue.nl

Raoul C.Y. Nuijten

Department of Industrial Engineering
Eindhoven University of Technology,
The Netherlands
r.c.y.nuijten@tue.nl

Pieter M.E. Van Gorp

Department of Industrial Engineering
Eindhoven University of Technology,
The Netherlands
p.m.e.v.gorp@tue.nl

ABSTRACT

Tailoring apps based on user traits has attracted tremendous interest in developing mHealth apps and understanding a users' personality is a key challenge in that context. This challenge is typically addressed via classic surveys, which pose a regrettably high burden on app users. This study aims to reduce the response burden of personality tests by introducing a model for predicting the user personality based on digital footprints of app usage. At the same time, skipping surveys completely turns out to undermine prediction accuracy. Therefore, this paper conceptualizes a hybrid framework that utilizes user event data in combination with surveys that have fewer questions than conventionally. The proposed method demonstrates a promising trade-off between the simplicity of using user event data and the accuracy of the validated survey methods: when applying the hybrid method to a retrospective case study, the accuracy is higher than when using the event data exclusively. Also, the number of survey questions needed is significantly lower. Since this is a novel method, we expect that results will strengthen as larger data sets available over time.

CCS CONCEPTS

• **Information systems** → Personalization; • **Computing methodologies** → Neural networks; • **Human-centered computing** → User centered design;; • **Software and its engineering** → Soft-ware usability; • **Social and professional topics** → User characteristics;

KEYWORDS

Digital phenotype, mHealth, Big-five, Personality Prediction, Health Interventions, Multi-layer Perceptron, Neural Network, Hybrid System

ACM Reference Format:

Ehsan Hadian Haghighi, Raoul C.Y. Nuijten, and Pieter M.E. Van Gorp. 2021. Towards Hybrid Profiling: Combining Digital Phenotyping with Validated Survey Questions to Balance Data Entry Effort with Predictive Power. In

2021 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021), February 19–21, 2021, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3459104.3459199>

1 INTRODUCTION

Recently, mHealth interventions have shown significant potential to empower its users to change their daily routines and adopt healthier lifestyles [10]. However, these digital support tools often fail to retain their impact in the long-term due to their users losing interest [12].

A promising strategy to retain users is by tailoring your intervention to the users' individual wants and needs [7]. Particularly, it was found that personalized mHealth interventions are more effective than interventions that adopt one-size-fits-all approaches [7].

User preferences are commonly derived from knowledge on a user's personality [11]. Therefore, to personalize an mHealth intervention, the tool first needs to understand the preferences a particular user has. Historically, personality traits are derived using questionnaires [22, 23].

However, users often perceive questionnaires as cumbersome [23], and the quality of their answers decreases with every next question [2]. Hence, lengthy personality tests increase the risk of users dropping out, and as such, the aim to reduce dropouts by applying trait-based personalization may fail a priori.

Although efforts have been invested in shortening popular personality tests, and user experience design techniques have been employed to reduce the perceived burden of questionnaires (e.g., using blocks and progress bars) [13], it was found that filling out popular personality tests still requires between five to fifteen minutes [15].

When the primary aim of such surveys is to personalize an mHealth intervention then they are typically part of a digital onboarding process. In digital onboarding processes, dropout is already a challenging issue so adding five to fifteen minutes there is probably unacceptable from the dropout point of view.

Another strategy to gain understanding of a user's personality is to apply unobtrusive learning methods from the domain of Artificial Intelligence (AI). For example, user event data is automatically recorded by users performing digital actions on mHealth tools. These actions have a certain meaning that may, or may not be, iconic for people with a certain personality. Hence, these digital traces can be tracked to construct an image of one's personality. For example, AI has been deployed to automatically determine a user's

This work is licensed under a Creative Commons Attribution International 4.0 License.

ISEEIE 2021, February 19–21, 2021, Seoul, Republic of Korea
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8983-9/21/02.
<https://doi.org/10.1145/3459104.3459199>

personality based on: the apps someone has installed [26], someone's behavior on social media platforms [4], and someone's call and messages logs [6]. These unobtrusive approaches are typically categorized as "Digital phenotyping, behavioral sensing, or personal sensing" [17]. Unfortunately, the current state-of-the-art in digital phenotyping in mHealth is insufficiently accurate to predict one's personality reliably [18].

In this study, we aim to combine classical personality tests with digital phenotyping techniques in order to explore the personalities of users of an mHealth tool. Particularly, we aim to increase the accuracy of digital phenotyping techniques, by enriching these unobtrusive learning methods with user reported survey responses of highly trimmed down personality tests.

Our findings demonstrate the relevance of investigating more subtly how hybrid approaches can be developed which balance well the user efforts with the predictive power. Since specific recommendations are largely dependent also on the quality of the data set that one has available to train a predictive model, our study aims to contribute also a mathematical model for rationalizing the design process of these new hybrid systems.

We showcase and validate our method using the GameBus platform, but the techniques are applicable more broadly: our proposed method can be implemented in any mobile application which (1) offers pages that have a strong link to psychological traits, and (2) which offers functionalities for prompting the user with short surveys. In this case study, we use user event data and a multilayer perceptron (MLP) neural network to have an estimation of two probability trait profiles of each user. Then, based on these predictions a limited number of questions will be asked to predict the user's personality. Results show that the proposed hybrid approach is promising and deserves more research attention.

2 BACKGROUND

2.1 GameBus

Recently, mHealth tools have been changing the face of healthcare. The rapidly increasing number of mHealth apps available in the market, clearly reveals that mHealth plays an important role in increasing health literacy, health management and patient engagement. It does remain a scientifically challenging question how one can design mHealth tools in a manner that users remain engaged over long periods of time.

GameBus (www.gamebus.eu) is an mHealth platform that uses gamification techniques as classified by Hamari [16]: it provides elements such as 'challenges', 'points', 'goals', 'progress', 'leaderboards', and 'rewards' in non-game contexts for promoting health activities in a playful manner. The GameBus platform is designed to persuade its end-users to a healthier lifestyle. Depending on the challenges chosen to take, different types of healthy activities - from social, to mental, to physical - are rewarded with virtual points. Such points can be linked to individual or team-based competitions, but also non-competitive goal tracking is supported. Moreover, Social support elements such as newsfeeds, chats and likes and comments can be employed for behavior change. The academic platform has already been used to support various controlled experiments that aim to show statistically the effect of certain platform configurations on user engagement [19, 20]. It turns out that results

often depend on personality traits of the study participants [21]. Personalization based on users' traits in GameBus can be applied to persuasive messages, often called nudges [9]. It is also possible to personalize the content, offered challenges, and design of the App in accordance with their personality. Therefore, we are preparing a dynamic variant that tailors the platform configuration based on user traits. Such plasticity in user experiences based on personal traits has been described in the context of marketing [22] but it seems not explored for health promotion well and the current state-of-art approaches are not accurate enough in mHealth [18]. Trait modeling can be done via a variety of models. In a gamification context, for examples, one could use the HEXAD model [25]. In the following section, we clarify a more established model from general psychology.

2.2 Big-Five Framework for Modeling User Preferences

The most common personality framework is the Big-Five [14], and there are several mappings framework between users' personality trait and actual (game) content [1, 25].

Personality testing in accordance with the Big-Five personality traits consists of five traits including extraversion, agreeableness, conscientiousness, neuroticism, and intellect representing personality at the broadest level of abstraction. Mini-IPIP [8] is one of the most commonly used short inventories measuring an individual on the Big Five dimensions of personality. Although Mini-IPIP with 20 questions provides a short-form IPIP survey delivering an acceptable proxy for a person's traits, it may still be felt like a burden since it typically takes 5 minutes to fill out. Although we contextualize the research using the GameBus app primarily because we have full data access for scientific research in compliance with the GDPR, this proposed method can be used by other mHealth platforms that collect similar data. Moreover, this hybrid approach is applicable to any psychological model.

3 METHODOLOGY

Construction of user's trait profile is considered as one of the main challenges in personalization [3]. User preference as one of the most important aspects of user satisfaction should map into personality characteristics. Therefore, the user's personality as the users' most representative of personality trait makes this as a classification problem. The main goal of this study is to evaluate users' preferences from their user event data in order to reduce their effort by answering fewer survey questions. The user event database is a collection of behavioral and analytics data that represents how users act in the app. Big-five survey results are considered as the ground truth for one's actual preferences. We collected such data for pilots that enable us to train and test machine learning models such that future users do not have to fill out such long surveys. Specifically, we start from a 20-item Mini-IPIP survey (Appendix B) and aim to further reduce the number of survey items without significantly compromising the predictive power of the questions that remain. Figure 1 is the graphical representation of the study design and general user journey when using the digital phenotype platform within mHealth applications.

Figure 1: Study design and visual representation of general user journey when using the proposed method.

3.1 Data Collection

Data for this study was reused from a previous study [19] which evaluated the impact of competitiveness in a mHealth setting for 12 weeks among first-year high-school students. In total, 313 unique participants – including 290 students, and 23 teachers – have joined the study. Consent was collected to use the pseudonymized data for scientific studies on mHealth behavior change. Participants also were invited to complete a Mini-IPIP survey personality test in accordance with the Big-Five personality traits leaving a total of 94 participants in the dataset. Regarding event data, a total of 35 input parameters were derived. These parameters are fully listed in Appendix A and some of them are mentioned below

- x_1 : number of days a user has been online in the given period (excluding the automatically tracked events)
- x_2 : number of days a user has been tracked using automatic trackers
- x_3 : number of times a user has viewed their personal activity log
- $\vdots$
- x_7 : number of times a user has viewed their profile
- $\vdots$
- x_{32} : number of points a user has collected
- x_{33} : number of unique challenge rules a user has performed
- x_{34} : number of unique activity types a user has performed
- x_{35} : number of unique activity types a user has performed that were rewarded.

3.2 Data Preparation

As clarified in the introduction, our proposed method aims to classify users in psychological profiles in order to personalize the user experience and optimize user engagement. When aiming to reduce dropout rates, that classification should happen as fast as possible, with adequate accuracy. Therefore, we analyzed empirically different scenario's (ranging from classifying already after one day to classifying only after multiple weeks). The study took place in 12 weeks and datasets as it is expressed in the form Eq. 1) they

were prepared based on daily intervals T from users' data. Datasets include both the values of input parameters $X_{i,j}$ (user event data) and the actual degree of belonging to the Big-Five classes $y_{i,l}$ for all participants that completed the survey.

$$Dataset_T = [X_{i,j}, y_{i,l}]_T \quad (1)$$

$$X_{i,j} = \begin{bmatrix} x_{1,1} & \dots & x_{1,j} \\ \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} \end{bmatrix} \quad (2)$$

$$Y_{i,l} = \begin{bmatrix} y_{1,1} & \dots & y_{1,l} \\ \vdots & \ddots & \vdots \\ y_{i,1} & \dots & y_{i,l} \end{bmatrix} \quad (3)$$

where i represents users ($i = \{1, 2, \dots, 94\}$), j is the feature item number ($j = \{1, 2, \dots, 35\}$), l denotes the Big-Five traits group ($l = \{1, 2, \dots, 5\}$). In all datasets, two types of normalization were applied to each of the input variables $x_{i,j}$. As it can be seen in Eq. 4) ; first, to improve the scalability of our analysis, input parameters (user event data) for each user within these datasets were divided by the number of days that person has been online ($x_{i,1}$). Second, all user event data were transformed via z-score for better normalization and comparisons. In this equation n stands for the total number of users ($n = 94$).

$$x'_{i,j} = \frac{\frac{x_{i,j}}{x_{i,1}} - \frac{1}{n} \sum \frac{x_{i,j}}{x_{i,1}}}{\sqrt{\frac{\sum \left(\frac{x_{i,j}}{x_{i,1}} - \frac{1}{n} \sum \frac{x_{i,j}}{x_{i,1}} \right)^2}{n}}} \quad (4)$$

To normalize the Big-Five survey data against data that represents Dutch society and categorize each user into one of the five trait groups, the score of each individual participant is compared with the average μ_{nl} and standard deviation σ_{nl} score of people in the Netherlands (see figure 1 and table 1) (i.e., average scores and standard deviations of personality traits for the Dutch populations, we made use of data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg

Table 1: Average scores of Big-Five personalities test in the Netherlands.

Trait (l)	extra.	agree.	consc.	neuro.	intel.
Avg. (μ_{nl})	3.12	3.84	3.72	3.48	3.48
Std. (σ_{nl})	0.66	0.52	0.52	0.71	0.5

University, The Netherlands) [24]) to the Big-Five personality test by computing the Z-score for each trait group l (Eq. 5). Consequently, user i belongs to trait group Y_i which has higher value among other trait groups (Eq. 6).

$$y'_{i,l} = \frac{y_{i,l} - \mu_{nl,l}}{\sigma_{nl,l}} \quad (5)$$

$$Y_i = \left[\max_l (y'_{i,l}) \right] \quad (6)$$

3.3 Statistical Analysis

3.3.1 Optimum Time Frame and Key Features. One reason why user might abandon an app is we were not able to deliver personalized experiences that fit into their needs. The previous study shows most of the users abandon the app after only two weeks [19]. Also, analysis show users' activities reached their peak on the third day, but then started to decrease significantly. Hence, the decision came down to a trade-off between the accuracy and observation time. Three days was considered as the optimum minimum time span required to have a good balance of the estimation accuracy and time. This time span can be considered as the golden time to engage people and make a personalized intervention.

Feature selection is one of the most important concepts in Machine Learning techniques to eliminate redundant variables and keeps only the best subset of predictors in the model. In the first phase, after categorizing users in their trait groups, average and standard deviation values of the input variables $x_{i,1}$ to $x_{i,35}$ (features) for users in the same group were calculated. Based on the standard deviation and dispersion of features, the algorithm selects three features with a higher value of standard deviation between the average of that variable within groups. In another word, this filter method follows the reasoning scheme to choose variables which their value being close within their group but have a significant distance from the other group.

3.3.2 Inactive Users. The second set of statistical analyses was focused on the evaluation of the users' activities. A respondent labeled as an inactive user if their key user event data have not changed during the study period of 12 weeks. It means these features have zero standard deviation during this period.

3.4 Model Building

3.4.1 Artificial Neural Networks. Artificial neural networks (ANNs) are a soft computing technique inspired by simulating the behavior of the human brain. This study utilizes the power of Keras library in Python for this purpose and a multi-layer perceptron (MLP) with 1 input layer, 2 hidden layers, and 1 output layer are used to build the model. Each hidden layer consists of 10 nodes as depicted in figure 2. The hyperbolic tangent activation function in each hidden layer

Figure 2: MLP classifier structure.

Table 2: Confirmatory analysis for 6 Mini-IPIP questions.

Trait	extra.	agree.	consc.	neuro.	intel.
Item number	11	2	3	4	5
	6	7	8	9	15
	16	17	18	19	20

and softmax function for the output layer are supplied along with the number of nodes. Since this problem is a type of single-label multi-class classification, categorical cross-entropy was applied in its settings rather than binary cross-entropy as the loss function of the model. Therefore, the proposed model gives a dispersion of probabilities between all five traits groups. Two categories that receive the highest probability will be the output for the model for the next step.

3.4.2 Selective Questions. By predicting two possible personality groups for each user from the previous step, users' mini-IPIP test scores on those two personality groups are considered. In practice, it is equivalent to just 8 questions out of 20 questions of the Mini-IPIP test that are asked from a user according to the predicted probable personality traits. In the next experiment with the aim to decrease the burdensome of the test, even more, we reduced the number of questions to 6 questions which includes 1 question with a positive keyed direction and 2 questions with negative keyed directions for each group. These questions were selected based on the confirmatory factor analysis (CFA) of the Mini-IPIP by choosing the highest values [5]. Confirmatory factor analysis is a popular powerful statistical tool in social research for evaluating latent variables. This analysis measures whether a prespecified factor model (hypothetical constructs) provides a good fit to the data or not. Consequently, questions number 6, 11 and 16 for extraversion, 2, 7 and 17 for agreeableness, 3, 8 and 18 for conscientiousness, 4, 9 and 19 for neuroticism, 5, 15 and 20 for intelligence were selected (table 2). As can be seen from table 3, in the next step, the number of questions is reduced to 4 questions by using confirmatory factors and choosing the highest positive and negative keyed directions in each group.

3.4.3 Cost function. Since this study uses the standard mini-IPIP questions to serve as ground-truth representing users' scores on

Table 3: Confirmatory analysis for 4 Mini-IPIP questions.

Trait	extra.	agree.	consc.	neuro.	intel.
Item number	11	2	3	4	4
	6	17	8	9	20

Table 4: Statistical analysis per trait group.

	vactivites		vprofile		points	
	Avg.	StD.	Avg.	StD.	Avg.	StD.
extra.	7.81	6.89	9.28	12.63	11.09	15.73
agree.	10.90	9.76	14.2	14.88	20.4	20.53
consc.	19.70	32.15	23.6	36.3	45.0	78.26
neuro.	7.25	6.08	7.25	5.14	11.5	12.7
intel.	10.85	9.46	13.54	22.88	13.19	24.85

Table 5: Statistical analysis of key features for all trait groups.

	vactivites	vprofile	points
Avg. between trait groups	11.3	13.57	20.24
StD. between trait groups	4.46	5.64	12.83

each of the Big Five dimensions, cognitive load is not contributing to our decision and the only factor that may have the greatest impact on the respondent burden is the length of the questionnaire. Therefore, to better visualization, we defined users' effort and model accuracy as the cost function, to make a balance between users' effort and the accuracy of the model. In Eq. 7 n is the number of questions and a is the accuracy of the model.

$$Cost = 0.55 - 0.55 \frac{a}{100} + 0.45 \frac{n}{20} \quad (7)$$

4 RESULTS

As previously mentioned, the main objective of this paper is to predict users' digital phenotype as the first step of personalization to create an optimized mobile app user experience based on users' trait. To ensure that the accuracy of the model persists, a 5-fold cross-validation method is applied for all the experiments.

4.1 Features Selection

Three features are selected based on their scores in the first statistical analysis. These features, as presents in table 4, are $x_{i,3}$, $x_{i,7}$, $x_{i,32}$ which respectively are the number of times a user has viewed their personal activity log, the number of times a user has viewed their profile, and the number of points a user has collected. Table 5 presents the average and the standard deviation of these features in all trait groups.

4.2 Active Users

Analyzing users based on the standard deviation of their key event data during these 12 weeks shows about 40% of participants who filled the survey (30 out of 94) were not active in that study at all.

Table 6: Prediction scores

Scores	
accuracy	42.7% ($\pm 7.7\%$)
recall	42.7% ($\pm 7.7\%$)
precision	34.5% ($\pm 16.7\%$)
f-score	32.6% ($\pm 11.8\%$)
cost	0.49 (± 0.04)

Figure 3: Active and inactive users.

By comparing users in each trait group, it appears that in three out of five categories we have fewer (active) users (see figure 3 and figure 4).

4.3 MLP Classifier

Initially, an MLP neural network was used to assess the feasibility of determining users' personality traits just using user event data. Considering the highest probability in groups as the predicted personality trait, the balance average accuracy of 42.7% ($\pm 7.7\%$) and the F-score 32.6% ($\pm 11.8\%$) for all 5-folds was obtained. Other scores of using MLP in predicting each trait group is summarized in table 6.

As discussed earlier, this paper proposes a model to combine user event data and ask some selective questions with the aim of reducing the response burden of the Mini-IPIP test and having a more accurate model for predicting user personalities. Therefore, by considering the two groups with the highest probability and rating those based on their scores to the questions for the predicted groups, the accuracy increases by 30% compared with using just MLP. Although the decline in the number of questions to 6 reduces the burden of the users by 10% more, the accuracy of the model slightly affected. Consequently, by using the defined cost function (Eq. 7) it is concluded asking 8 questions would be better. Table 7 shows the accuracy of proposed model for each trait group with 8 and 6 questions.

4.4 Evaluation of Model Performance

The values of accuracy are quite stable in all five folds of cross-validation. Hence, promising results are achieved for the prediction

Figure 4: Dropout rate.

Table 7: Prediction scores of using Hybrid method (user event data and asking selective related questions).

	Scores for 6q	Scores for 8q
accuracy	63%(±7.2%)	72%(±4.7%)
recall	63%(±7.2%)	72%(±4.7%)
precision	53.3%(±10%)	60.6%(±11.7%)
f-score	56%(±8.1%)	63.7%(±7.7%)
cost	0.33(±0.04)	0.33(±0.02%)

of the digital phenotype by the proposed framework. To test the effect of asking smart selective questions, the model was tested by random questions. It was ensured that at least one question from each group was asked about that type of personality. The accuracy in this scenario drops to the 33.7%(±5.6%) .

4.5 Performance Analysis

To analyze the performance of the proposed model, it was tested by a set of 14 users' data representing in table 8.

Moreover, users' personality was defined by the Mini-IPIP test as the ground truth table. The results of this test appear in Tables 8 as well. In the first step, an MLP classifier is used to predict the test groups' traits. Figure 5 shows confusion matrix of using MLP classifier and as can be seen the accuracy of using user event data without asking any question is 43%. Also, precision, recall and F-score, in this case, are 25%, 43%, 31% respectively.

To follow our goals, instead of using binary values in classification with MLP, we use predicted probabilities that of to what extent a user belongs to each trait group. Therefore, Z-score for related questions to the two groups having the highest probabilities is calculated. In this method, the accuracy of using hybrid technique, as it is shown in table 9 and figure 6, is 71.43% which is about 30% higher than using the MLP neural network.

By reducing the number of questions to 6, the performance of the model dropped. Figure 7 shows the confusion matrix of that and table 10 clearly shows that when users are asked 6 questions, the result is worse than when they are asked 8 questions.

Table 8: Test group dataset.

	points	vprofile	vactivities	Phenotype
User 1	29	10	11	extra.
User 2	5	1	2	extra.
User 3	0	6	5	intel.
User 4	29	11	13	neuro.
User 5	143	51	49	consc.
User 6	5	9	10	extra.
User 7	24	12	12	extra.
User 8	34	20	21	agree.
User 9	0	2	1	extra.
User 10	13	7	9	intel.
User 11	150	133	57	extra.
User 12	4	10	4	intel.
User 13	49	10	22	intel.
User 14	0	7	2	extra.

Figure 5: Confusion matrix of using MLP classifier.

Figure 6: Confusion matrix for 8 questions.

Table 9: Prediction scores of using the proposed hybrid method by asking 8 questions.

Scores	
accuracy	71.43%
recall	71.43%
precision	56.55%
f-score	62.86%
cost	0.33

Figure 7: Confusion matrix 6 Mini-IPIP questions.

Table 10: Prediction scores for 6 Mini-IPIP questions

Scores	
accuracy	57.14%
recall	57.14%
precision	45.54%
f-score	50.48%
cost	0.37

5 DISCUSSION AND CONCLUSIONS

Quickly and accurately learning the personality traits of mobile app users is challenging. Our results demonstrate the potential of combining validated survey methods with unobtrusive trait modeling based on digital footprints. Specifically, we demonstrated a significant reduction in the number of survey questions that users need to answer during an onboarding process while maintaining acceptable accuracy. Unfortunately, our retrospective case study does suffer from the lack of samples in three trait groups leading to low accuracy for those groups. Others applying our proposed method on their own data set may experience similar yet slightly different data set limitations. Therefore, we also contributed a formal approach to rationalizing the engineering process for hybrid systems.

Results of using MLP for users' trait classification show that for the two groups for which the retrospective data set has most active users, the accuracy and corresponding F-score are highest. Therefore, hopefully increasing in number of samples may help the model to obtain better results. Having Fewer samples in the three groups cannot be a tangible proof or adequate reason that most users have an extroverted or Intellect personality. Rather, it is quite probable that the users who did not participate in the survey can belong to those three groups whom for some reason we could not encourage them to fill out the survey.

To improve reliability and prediction accuracy for determining users' digital phenotype, a limited number of questions based on estimated personalities from users event data is asked. The proposed model has established a good interaction between the simplicity of using user event data and the accuracy of the conventional survey methods. In light of this, the optimal solution occurs when users are asked by 8 questions subsequently. It means we reduced the burden of the survey by 60% and we still have acceptable accuracy in our prediction. Even when users do not actively participate in the survey, by using user event data exclusively, we would reach only about 50% accuracy. Although diminishing the number of questions to 6 reduces the burden of the survey by 10%, it reduces the performance of the model by the same level as well.

The retrospective case study was based on a health gamification app. In that setting, digital traces included variables such as the number of times a user has viewed their personal activity log, the number of times a user has viewed their profile, and the number of points a user has collected. These variables aided to predict significant differences in users with different personality traits. On one hand, users in the Conscientiousness group are more loyal to the platform; on the other hand, the app has not engaged the interest of users with agreeableness and extraversion trait to use the app and they have a higher rate of dropout.

Running a trial on the impact of personalization of challenges and users' experience would be an opportunity to have a better understanding of the effect of it in the context of mHealth tools. More broadly, having a good proportion of samples in all traits' groups could improve the accuracy of the proposed model.

REFERENCES

- [1] F. R. Bartle. Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD research*, 1(1):19, 1996.
- [2] F. J. R. Busemeyer and J. T. Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432, 1993.
- [3] F. S. Ceri and M. Brambilla. *Search computing: trends and developments*, volume 6585. Springer Science & Business Media, 2011.
- [4] F. D. N. Chin and W. R. Wright. Social media sources for personality profiling. In *UMAP Workshops*, 2014.
- [5] F. A. J. Cooper, L. D. Smillie, and P. J. Corr. A confirmatory factor analysis of the mini-ipip five-factor model personality scale. *Personality and Individual Differences*, 48(5):688–691, 2010.
- [6] F. Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 48–55. Springer, 2013.
- [7] F. A. DeSmet, D. Van Ryckeghem, S. Compennolle, T. Baranowski, D. Thompson, G. Crombez, K. Poels, W. Van Lippevelde, S. Bastiaensens, K. Van Cleemput, et al. A meta-analysis of serious digital games for healthy lifestyle promotion. *Preventive medicine*, 69:95–107, 2014.
- [8] F. M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

- [9] F. J. E. d'Hondt, R. C. Nuijten, and P. M. Van Gorp. Evaluation of computer-tailored motivational messaging in a health promotion context. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 120–133. Springer, 2019.
- [10] F. L. V. Eckerstorfer, N. K. Tanzer, C. Vognrinic-Haselbacher, G. Kedia, H. Brohmer, I. Dinslaken, and K. Corcoran. Key elements of mhealth interventions to successfully increase physical activity: Meta-regression. *JMIR mHealth and uHealth*, 6(11):e10076, 2018.
- [11] F. B. Ferwerda, E. Yang, M. Schedl, and M. Tkalcic. Personality traits predict music taxonomy preferences. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2241–2246, 2015.
- [12] F. E. A. Finkelstein, B. A. Haaland, M. Bilger, A. Sahasranaman, R. A. Sloan, E. E. K. Nang, and K. R. Evenson. Effectiveness of activity trackers with and without incentives to increase physical activity (tripipa): a randomised controlled trial. *The lancet Diabetes & endocrinology*, 4(12):983–995, 2016.
- [13] F. M. Galesi' c. Dropouts on the web: Influence of changes in respondents' interest and perceived burden during the web survey. *Journal of Official Statistics*, 22:313–328, 2004.
- [14] F. L. R. Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.
- [15] F. S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.
- [16] F. J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work?—a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*, pages 3025–3034. Ieee, 2014.
- [17] F. D. C. Mohr, K. Shilton, and M. Hotopf. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *npj Digital Medicine*, 3(1):1–2, 2020.
- [18] F. R. Nuijten, U. Kaymak, P. Van Gorp, M. Simons, P. Van Den Berg, and P. Le Blanc. Fuzzy modeling to understand personal preferences of mhealth users: a case study. In *2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*. Atlantis Press, 2019.
- [19] F. R. Nuijten, P. Van Gorp, T. Borghouts, P. Le Blanc, P. Van den Berg, A. Kemperman, E. Hadian Haghighi, and M. Simons. Different implementations of social comparison as drivers for health behavior change: Evaluating engagement levels of preadolescent students with an mhealth intervention (submitted). *Journal of Medical Internet Research*, 2020.
- [20] F. R. C. Nuijten, P. M. Van Gorp, U. Kaymak, M. Simons, A. D. Kemperman, and P. E. Van den Berg. Evaluation of the impact of extrinsic rewards on user engagement in a health promotion context. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3600–3604. IEEE, 2019.
- [21] F. A. Nunes, T. Limpo, and S. L. Castro. Individual factors that influence the acceptance of mobile health apps: The role of age, gender, and personality traits. In *International Conference on Information and Communication Technologies for Ageing Well and e-Health*, pages 167–179. Springer, 2018.
- [22] F. R. Orji, R. L. Mandryk, and J. Vassileva. Improving the efficacy of games for change using personalization models. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–22, 2017.
- [23] F. M. Revilla and C. Ochoa. Ideal and maximum length for a web survey. *International Journal of Market Research*, 59(5):557–565, 2017.
- [24] F. A. C. Scherpenzeel and M. Das. True" longitudinal and probability-based internet panels: Evidence from the netherlands. *Social and behavioral research and the internet: Advances in applied methods and research strategies*, pages 77–104, 2010.
- [25] F. G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke. The gamification user types hexad scale. In *Proceedings of the 2016 annual symposium on computer-human interaction in play*, pages 229–243, 2016.
- [26] F. R. Xu, R. M. Frey, E. Fleisch, and A. Ilic. Understanding the impact of personality traits on mobile app adoption—insights from a large-scale field study. *Computers in Human Behavior*, 62:244–256, 2016.

APPENDIX

A EVENT DATA

x1: number of days a user has been online in the given period(excluding the automatically tracked events)

x2: number of days a user has been tracked using automatic trackers
x3: number of times a user has viewed their personal activity log
x4: number of times a user has viewed their friends' activity log
x5: number of times a user has viewed the details of an activity
x6: number of times a user has viewed their newsfeed
x7: number of times a user has viewed their profile
x8: number of times a user has viewed their friends' profile
x9: number of times a user has viewed the details of (one of their)circles (= teams)
x10: number of times a user has viewed their overview of challenges
x11: number of times a user has viewed the leaderboard of a particular challenge
x12: number of times a user has viewed the leaderboard of a particular circle (= team) within a particular challenge
x13: number of times a user has given support to others
x14: number of unique players a user has supported
x15: number of times a user has received support from other players (i.e., not including the number of times a user has supported herself)
x16: number of unique players that have supported the user (i.e., not including the user herself, in case she has supported herself)
x17: number of reactions on activities a user has made
x18: number of unique players a user has reacted on (i.e., a reaction on their activities)
x19: number of reactions a user has made on her own activities
x20: number of reactions a user has received on her activities, not including the reactions she has made on her own activities
x21: number of unique players a user has received reactions from on her reactions, not including herself (if she has made a reaction on her own activities)
x22: number of circle (= team) chat messages send by a user
x23: number of unique circles a user has sent a chat message to
x24: number of circle chat messages a user has received, to including the chat messages she has received from herself
x25: number of unique players that have sent the user a circle (= team) chat message, excluding the user herself (if she has sent any messages herself)
x26: number of unique circles (=teams) that have sent the user a circle chat message
x27: number of activities a user has performed (both manually and automatically registered)
x28: number of activities that were automatically registered for a user (for example via Google Fit, Fitbit, or any other automatic tracker)
x29: number of activities (both annually and automatically registered) that were rewarded by a challenge
x30: number of automatically registered activities that were rewarded by a challenge
x31: number of challenges a user was enrolled in
x32: number of points a user has collected
x33: number of unique challenge rules a user has performed
x34: number of unique activity types a user has performed
x35: number of unique activity types a user has performed that were rewarded

B BIG-FIVE MINI-IPIP SURVEY

Table 11: Mini-IPIP survey

Item	Factor	Text	Keyed direction
1	Extra.	Am the life of the party.	Positive
2	Agree.	Sympathize with others' feelings.	Positive
3	Consc.	Get chores done right away.	Positive
4	Neuro.	Have frequent mood swings.	Positive
5	Intel.	Have a vivid imagination.	Positive
6	Extra.	Don't talk a lot.	Negative
7	Agree.	Am not interested in other people's problems.	Negative
8	Consc.	Often forget to put things back in their proper place.	Negative
9	Neuro.	Am relaxed most of the time.	Negative
10	Intel.	Am not interested in abstract ideas.	Negative
11	Extra.	Talk to a lot of different people at parties.	Positive
12	Agree.	Feel others' emotions.	Positive
13	Consc.	Like order.	Positive
14	Neuro.	Get upset easily.	Positive
15	Intel.	Have difficulty understanding abstract ideas.	Negative
16	Extra.	Keep in the background.	Negative
17	Agree.	Am not really interested in others.	Negative
18	Consc.	Make a mess of things.	Negative
19	Neuro.	Seldom feel blue.	Negative
20	Intel.	Do not have a good imagination.	Negative