Personality Trait Recognition Based on Smartphone Typing Characteristics in the Wild

Nikola Kovačević, Christian Holz, Tobias Günther, Markus Gross, and Rafael Wampfler

***Abstract*—As governed by personality trait theory, humans tackle problems differently depending on their long-term behavioral characteristics. Computational awareness of personality traits fuels affective computing research, which investigates how to reliably recognize and utilize personality traits. Applications are diverse, including therapy monitoring, learning assistance, and recommender systems. Data-driven approaches are a promising path forward towards personality-aware human-computer interactions. Thereby, central challenges are the non-disruptive data acquisition, the time frame over which data must be collected before predictions become accurate, and the feature-centered data reduction to train reliable and lightweight machine learning models. In this work, we address these challenges by presenting a fully-automatic feature extraction and machine learning pipeline that makes accurate personality trait predictions for the widely-used Five Factor Model from passively-collected, short-term smartphone typing data collected from 76 participants (68 university students) in the wild. Our model allows for personality trait assessments after one day of data collection, demonstrating that, despite being a long-term behavioral trend, personality traits can be inferred accurately from shorter time periods. We demonstrate that our system can accurately predict personality traits on two levels (low and high) with up to 74.5% accuracy and 0.72 AUC for a single day, and up to 84.5% accuracy and 0.79 AUC after subsequent refinement over 10 weeks.**

***Index Terms*—Affective computing, classification, machine learning, personality traits, smartphone typing data.**

I. Introduction

PERSONALITY trait theory is a field of psychology that emerged in the 1920s and tries to explain the individual differences in human behavior when experiencing or coping with different situations [1]. It claims an individual's personality to consist of a set of traits, each differing in intensity and revealing certain behavioral characteristics such as the tendencies to think, feel and behave in a certain way [2], [3], [4]. In contrast to emotions (i.e., characteristics that greatly vary in the short term), and except for a natural drift over the years [5], [6], personality traits remain relatively stable over decades and are therefore suited to represent a person's long term behavioral characteristics [7]. Knowledge of personality traits can be leveraged in several domains. For example, in context-aware recommender systems, suggestions become more accurate if the intensity of users' personality traits is taken into account [8], [9], because personality traits are connected to people's entertainment interests and preferences [10], [11]. Furthermore, in therapeutic settings, negatively connoted personality facets (unique aspects of the broader personality traits [12]) or facets that hinder people in their everyday lives can be changed by regularly attending coaching sessions or psychological therapy while tracking the progress over time and being in agreement with the therapist about changing the facets [13]. Despite concerns [14], [15], personality traits are also used in personnel recruitment for screening candidates based on how well their personality traits fit the personality of interest according to the company's needs [16], [17]. Finally, personality traits can be used in adaptive learning environments to improve the learning gain [18].

Personality traits are often described in terms of the Five Factor Model [19], [20], [21], [22], which quantifies personality as a combination of five traits: *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness* and *neuroticism* (OCEAN). Typically, pen-and-paper tests are used to assess personality traits [23], [24], [25], [26]. As with all tests that include subjective components, scores can be distorted through dishonest answers, which generally limits their suitability [27], [28]. Distortion can even be subconscious through self-deception phenomena where candidates lack objectivity when assessing their own personality characteristics [29], [30]. To overcome the shortcomings of subjective assessments, personality traits have been automatically recognized using objective and reliable tracking modalities. For example, several studies have investigated how smartphone usage patterns relate to personality traits [31], [32], [33], [34], leveraging today's ubiquity of smartphones as an integral part of our everyday lives, which produce vast amounts of data [35]. However, missing validation of their findings in the wild [31], [34] and substantially long inference times of multiple weeks [33], [36] make the proposed methods unsuitable for integration into personality-aware systems that operate on short-term data.

Manuscript received 29 December 2021; revised 2 February 2023; accepted 28 February 2023. Date of publication 6 March 2023; date of current version 29 November 2023. Recommended for acceptance by H. Hung. *(Corresponding author: Nikola Kovačević.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by ethics board of ETH Zurich under Application No. EK 2020-N-60.

Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler are with the Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland (e-mail: nikola.kovacevic@inf.ethz.ch; christian.holz@inf.ethz.ch; grossm@inf.ethz.ch; wrafael@inf.ethz.ch).

Tobias Günther is with the Department of Computer Science, Friedrich-Alexander University, 91054 Erlangen, Germany (e-mail: tobias.guenther@fau.de).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2023.3253202, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2023.3253202

In this work, we present a novel method for inferring the personality traits of the Five Factor Model from typing events on commodity smartphones. We chose this input modality, since a majority of the smartphone usage time is attributed to conversational apps that include typing [35], and typing data can be collected passively. Based on an in-the-wild data acquisition study [37] with smartphone typing data from 76 participants collected over 10 weeks that was annotated with the personality traits from the Five Factor Model, we introduce a fully automated technique for extracting and selecting relevant typing patterns and characteristics that does not require manual feature engineering. We incorporate our model into an aggregation scheme that refines and improves the model performance over time by leveraging previous predictions.

Using typing characteristics of one day, our system predicts personality traits on two levels (below or above the population median) with an accuracy of up to 74.5% and up to 0.72 AUC. After subsequent refinement of previous predictions over 10 weeks, our model reaches up to 84.5% accuracy and up to 0.79 AUC. Furthermore, our analysis reveals that stable personality trait predictions can be obtained after 10 days of data collection. We found that the most predictive features varied across traits, indicating that different patterns are relevant for each personality trait. In contrast to current inference models, our model leverages short-time characteristics and therefore allows for a much faster personality assessment including the tracking of personality trait drift over time, which constitutes an important step towards a deployment in personality-aware systems where only short-term data is available. In summary, we make the following three contributions:

- We infer the personality traits of the Five Factor Model from in-the-wild smartphone data by leveraging short-term typing characteristics.
- We present a novel, fully automated feature extraction and selection pipeline without the need for manual feature engineering.
- We improve the model performance by aggregating previous predictions over time.

II. Related Work

A. Pen-and-Paper Personality Assessment

The first use of large-scale personality tests in Europe was the Woodworth Personal Data Sheet (WPDS) [26], introduced before World War II, that assessed a military recruit's emotional stability without the need of a professional psychiatric evaluation, but rather with an easy, fast and cheap pen-and-paper test. In the following years, the systematic assessment and investigation of human behavioral characteristics was increasingly brought into focus, which spurred researchers to introduce various testing schemes assessing different sets of personality traits [38]. One of the personality tests that assesses the personality using the Five Factor Model is the Big Five Inventory (BFI) [2], which consists of 44 statements (8 to 10 statements per trait) rated on a 5-point Likert scale, indicating the strength of agreement with the statement. The ratings are aggregated into an intensity score per personality trait, resulting in a 5-tuple that represents the final personality model. The BFI was later improved, which yielded the Big Five Inventory 2 (BFI-2) [23]. The BFI-2 lifted some limitations of the BFI. In particular, the number of statements was increased to 60 (i.e., the same number of statements for each personality trait). Furthermore, the statements were revised and adapted by psychologists and linguists in order to make the statements more understandable and easier to interpret, which in turn allowed the BFI-2 to be easily translated into multiple languages. This enabled the investigation of cultural differences around the world and turned the BFI-2 into a tool that allows language-independent comparisons across multiple studies [39]. Variants of this test include the NEO-PI-R questionnaire [40] (240 statements), the NEO-FFI questionnaire [24] (a short version of the NEO-PI-R with 60 statements) and the TIPI questionnaire [25] (10 statements).

B. Automatic Personality Trait Prediction

The field of personality computing has gained a lot of interest in recent years. Thereby, researchers tie methodologies from computer science and psychology together to drive forward the automatic assessment, perception, and recognition of human personality [41], [42]. Inferring personality traits is achieved through various modalities. For example, Suen et al. [43] predicted personality traits from facial video input obtained from job interviews. Carbonneau et al. [44] inferred speaker personality by analyzing spectrograms from speech. Zhao et al. [45] used motion patterns from computer mouse operations to detect personality. Subramnian et al. [46], Li et al. [47], and Zhao et al. [48] used biosensor signals for recognizing personality traits. Another line of research recognized personality traits from social media activity [49], [50], [51], [52]. Furthermore, several studies [31], [32], [33], [34], [36] showed a connection between smartphone-captured data and personality traits by leveraging the increasing importance of smartphones in people's everyday lives and the extensive amount of data they produce on a daily basis [35]. Both binary classification (using the population median as a separation margin) [31], [32], [33], [36] and regression [34] have been investigated in the past for inferring the personality traits of the Five Factor Model. However, the results suggest that binary classification is already a hard problem since all mentioned methods achieved rather low performance for two-class classification. In the following, we introduce models relying on different tracking modalities in more detail.

1) Touch-Based Input: Küster et al. [31] presented a touch-based method based on recorded touch interactions while users played a spelling game on a tablet during multiple sessions of around 50 minutes. From the touch data, several features such as swipe speed and touch frequency were computed. To obtain labels, the participants filled in the NEO-FFI questionnaire [24]. Different classifiers such as support vector machines (SVMs), random forests and logistic regression were trained on the data and achieved a classification accuracy of up to 67% (extraversion), suggesting that touch-based features are predictive for an individual's personality. However, this approach is unpractical for repetitive long-term personality tracking because of the

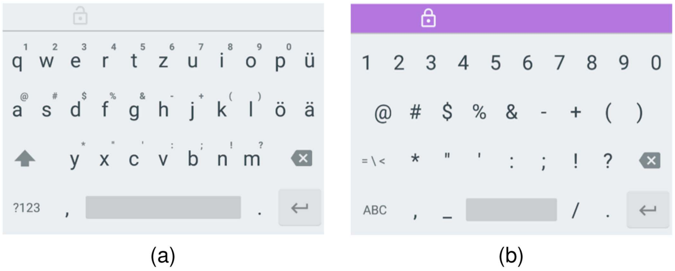


Fig. 1. The custom keyboard used during the data collection. The keyboard can be used in normal mode (a), and private mode visualized by a purple bar (b), where data collection is disabled.

required additional user effort, and it is unclear if the findings extend to externally valid assessments of larger cohorts and over a longer period of time in the wild.

2) Context-Based Input: Chittaranjan et al. [33] used context-related data extracted from application usage, Bluetooth and phone logs, and SMS in order to infer the personality traits. The dataset was collected in the wild during eight months and was aggregated on a monthly level. From the aggregated data, several features were computed (e.g., call durations, frequency of Bluetooth and application usage, message length and their statistical derivatives). The TIPI questionnaire [25] was used for obtaining the labels. Using an SVM classifier, an accuracy of up to 75.9% (extraversion) was achieved. In a follow-up data collection [36], the dataset was increased to 17 months of data and multiple features were added. The revised method achieved an F_1-score of up to 0.77 (extraversion). However, the long data collection time impedes a direct integration into personality-aware systems that have only short-term data available, and it is unclear how shorter data collection windows would influence the performance.

III. Dataset

A. Data Acquisition

We use a subset of the data collected in a previous work [37] where an experiment with 82 participants (43 female, 39 male) was conducted to collect smartphone typing and context data in the wild. The data collection was approved by the ethics board of ETH Zurich (EK 2020-N-60). On average, participants engaged in the experiment for 72 days (standard deviation SD = 2 days). The participants were between 18 and 43 years old (mean = 23.0 years, SD = 3.64 years). Seventy-seven of the participants were enrolled in different universities (61 bachelor, 13 master, 3 Ph.D. students). All participants were required to install an Android application on their smartphones, which collected activity records such as application usage logs, screen time and time zone changes for validational purposes. The typing events were recorded using a custom keyboard shipped with the Android application (see Fig. 1). As part of the ethics approval, only touch coordinates and the associated timestamps were recorded. Although the keystrokes were not explicitly stored, they could easily be deciphered. Therefore, a private mode, depicted as a lock on the keyboard top bar, allowed the users to temporarily deactivate the data collection manually at any time to increase user privacy. Furthermore, the private mode was automatically enabled based on the field of entry (e.g., password, e-mail, and numbers). Over the entire study, the private mode was activated 56 times (SD = 47) on average per user. In total, 7.6% were manual activations (SD = 11.6%). The app bundled the collected data and sent it periodically to a server whenever the phone was connected to WiFi or when the participants manually synchronized with the server via mobile data. The dataset was annotated with the personality traits of the Five Factor Model by requiring all users to fill in a German version of the Big Five Inventory 2 (BFI-2) questionnaire [23], [53]. Both the English [23] and the German [53] version of the BFI-2 show a high internal reliability (Cronbach's itemized alpha coefficient of up to 0.87 and 0.88, respectively). To enable data validation, the questionnaire was filled in before and after the data collection study.

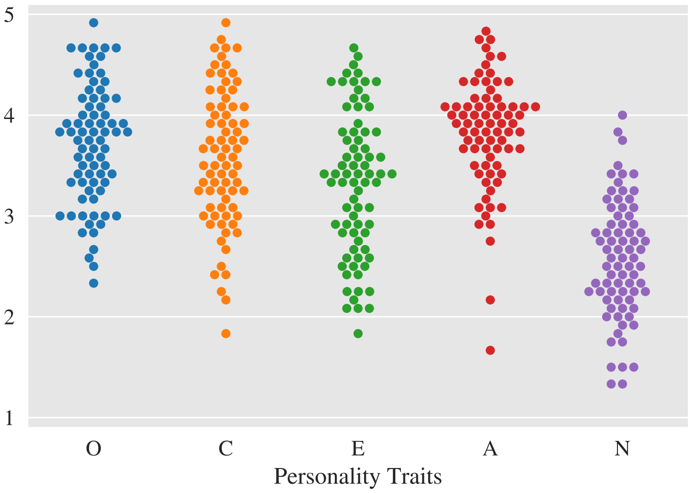


Fig. 2. The average personality trait scores from the BFI-2 over 76 participants on the range [1,5] grouped by personality trait: *openness to experience* (O), *conscientiousness* (C), *extraversion* (E), *agreeableness* (A), and *neuroticism* (N).

B. Data Evaluation and Validation

We aggregated the personality trait scores of each participant using the mean score of the questionnaires and grouped them by personality trait (see Fig. 2). The individual scores are widely spread over the range [1,5] and the average scores are similar to related findings by Schmitt et al. [39] where geographical differences were investigated (see Table I). The bias for neuroticism (0.41 compared to the western Europe mean) can be explained with our population being skewed towards student participants and students' academic performance being negatively correlated with neuroticism [54]. An analysis of the frequency of special characters (i.e., space bar, backspace, and punctuation characters) revealed that the touch frequencies are uniformly distributed over the course of the study (relative entropy of 0.042, SD = 0.003 with respect to the uniform distribution).

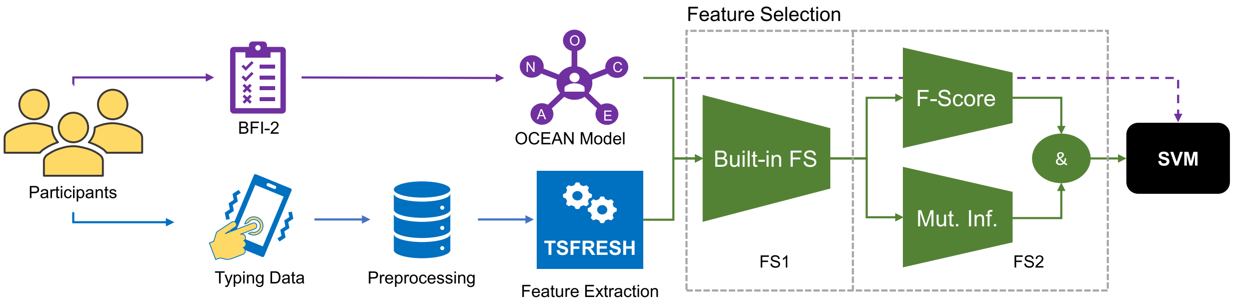


Fig. 3. Visualization of the main steps in our pipeline consisting of the preprocessing and feature extraction of the typing data (blue), the evaluation of the OCEAN model using the BFI-2 questionnaire (purple), and the feature selection pipeline (green), which feeds one selected feature set per personality trait into the classification model (black).

TABLE I
MEAN PERSONALITY TRAITS FROM OUR DATASET COMPARED TO THE GLOBAL MEAN AND WESTERN EUROPE MEAN FROM SCHMITT ET AL. [39] IN THE RANGE [1,5] FOR *OPENNESS TO EXPERIENCE* (O), *CONSCIENTIOUSNESS* (C), *EXTRAVERSION* (E), *AGREEABLENESS* (A), AND *NEUROTICISM* (N). FOR OUR DATASET, THE STANDARD DEVIATIONS ARE SHOWN IN BRACKETS

	Our Dataset	World	Western Europe
O	3.79 (0.59)	3.64	3.71
C	3.58 (0.68)	3.52	3.39
E	3.42 (0.72)	3.34	3.38
A	3.92 (0.56)	3.69	3.56
N	2.58 (0.58)	3.01	2.99

There were no significant correlations between special characters and personality traits. Alphanumeric characters were not further analyzed due to restrictions from the ethics board (see Section III-A).

C. *Exclusion of Participants*

Initially, eighty-two participants completed the experiment. Six of them were excluded during the data evaluation and validation process. In particular, two participants were excluded because they did not fill in the second personality test, hence their scores could not be validated. Three participants were excluded because they showed a high discrepancy between the two test results for one or more traits. The corresponding threshold was set at 20% because the 5-point Likert scale used in the BFI-2 induces a discretization error of 12.5%, and intentional personality changes of up to 20% per facet (i.e., approximately 4% per trait) are achievable through mental coaching [13], which amounts to approximately 16.5% and was rounded up to 20% as a tolerance margin. The absolute discrepancy between the two test results (before and after the data collection) was on average 5.7% (SD = 4.5%) for openness, 5.6% (SD = 4.6%) for conscientiousness, 5.9% (SD = 4%) for extraversion, 5.6% (SD = 4.2%) for agreeableness, and 7.6% (SD = 6.2%) for neuroticism. One participant was excluded because of non-monotonically increasing timestamps, which is an indicator for corrupted timestamp values. Hence, the final dataset consisted of 76 users.

IV. METHODOLOGY

The automatic prediction of personality traits from short-term information is fruitful for many applications, yet it is coupled to challenging demands: the data collection should not be disruptive, and derived feature descriptors should be rich enough for accurate data-driven predictions from only short time spans of data. Our approach to this problem is illustrated in Fig. 3. As we follow a data-driven approach, a cohort of 76 test subjects first completed a BFI-2 questionnaire to determine a classification label for each personality trait of the OCEAN model (purple branch in Fig. 3). To meet the non-disruption constraints, we utilized daily smartphone usage data passively collected from the participants over a course of 10 weeks (blue branch in Fig. 3). Since a substantial part of the overall smartphone usage time can be attributed to conversational applications [35], we focused on typing data. From the typing data, we automatically extracted more than 5,000 feature descriptors. In order to derive a small yet rich set of features to be used at inference time, we subsequently reduce the feature descriptors in a two-stage process, including linear and non-linear correlation metrics (green branch in Fig. 3). The resulting feature sets are optimized per personality trait, contain about two dozen features, and are used to train a support vector machine (black branch in Fig. 3), aiming to predict the corresponding label. At inference time, only the reduced feature set is needed to infer personality traits. In the following subsections, we elaborate on the data preprocessing, feature extraction, feature selection, and model training.

A. *Data Preprocessing*

Due to the physiologically inherent circadian rhythm [55], the typing time series naturally align on a daily time scale. Its smallest unit is a day, which results in 5472 sample points (72 days on average times 76 users), which proved to be a sufficiently long time frame to collect information from which behavioral patterns can be extracted [56]. Therefore, we first aligned all collected Unix timestamps to the correct time zones using the time zone change logs collected during the study because users may have switched time zones (e.g., holidays, business trips, and altering the phone settings). Then, all timestamps were normalized to the range $[0, 1]$ where 0 indicates the start of a day and 1 indicates the end of a day. Phone usage heatmaps revealed that a cut-off at 4 a.m. is the most suitable threshold for the start

TABLE II
Overview of Input Features and Derived Higher-Order Features Used in the Feature Extraction Process and Their Definitions in Terms of Time Series With N Time Steps

Input Feature	Definition as Time Series
x-Coordinate	$x = [x_i]_{i=1,\ldots,N}$
y-Coordinate	$y = [y_i]_{i=1,\ldots,N}$
Day-Timestamp	$t = [t_i]_{i=1,\ldots,N}$

Higher-order Feature	Definition as Time Series
Sum of x and y	$\text{sum}(x, y) = [x_i + y_i]_{i=1,\ldots,N}$
Elapsed Time	$\text{elapsed}(t) = [t_{i+1} - t_i]_{i=1,\ldots,N-1}$
Euclidean Distance	$\text{distance}(x, y) = \left[\sqrt{x_i^2 + y_i^2}\right]_{i=1,\ldots,N}$
Typing Speed	$\text{speed}(x, y, t) = \frac{\text{distance}(x, y)}{\text{elapsed}(t)}$

and end of a day since many users were still active after midnight. In a next step, the Google PlayStore was crawled to obtain category information for all apps the participants used over the course of the study. Only touch events that occurred in the scope of a conversation app (e.g., WhatsApp, Telegram, and Viber) or other apps that allowed typing (e.g., browsers and notepads) were kept. Then, all touch events that exceeded the bounds of the custom keyboard (e.g., when scrolling through chat messages or when the keyboard was not shown) were removed. In a last step, the remaining touch events were normalized to the range $[0, 1] \times [0, 1]$ based on the screen orientation and resolution, and the keyboard size to ensure comparability across participants.

B. *Feature Extraction*

Since our typing data is timestamped, it can be considered as a time series of N typing events where the timestamp $t = [t_i]_{i=1,\ldots,N}$ represents the time axis, and $x = [x_i]_{i=1,\ldots,N}$ and $y = [y_i]_{i=1,\ldots,N}$ represent the x-coordinates and y-coordinates of the typing events, respectively. We used the Feature Extraction Based On Scalable Hypothesis Tests (FRESH) algorithm by Christ et al. [57], which extracts various features and their statistical derivatives from time series and offers a variety of statistical significance tests based on which the features can be filtered. The multiple tests problem that arises when a high number of such statistical tests is conducted, is tackled by controlling the false discovery rate using the Benjamini-Yekutieli procedure [58]. The algorithm is available as part of the Python package *tsfresh* [59], which offers 63 features and a set of recommended default parameter settings for each, resulting in 794 features. We extracted all features with all recommended default parameters for all seven higher-order features from Table II (5558 features in total). For an exhaustive list of all available features and their parameter sets, we refer to the official documentation [59].

Since the FRESH algorithm does not consider combinations of time series, but extracts the features over each time series separately, we additionally formed higher-order features using the following combinations: The pairwise sum of the touch coordinates, the elapsed time between two consecutive typing events, the Euclidean distance between two typing events and the resulting typing speed measuring the distance between two consecutive typing events per elapsed time (see Table II for the definitions). Thus, the features are extracted over a total of seven time series separately, which all have been previously normalized to the range $[0, 1]$ using min-max normalization.

C. *Feature Selection*

Selecting informative features reduces noise, and can thus improve model performance [60], while also positively influencing the run time since the number of features is decreased. In the following, we present our individual feature selection steps as depicted in the green branch of Fig. 3 in more detail.

1) Preprocessing: In a first step, all constant features (i.e., features that are identical for all users) are removed since they contain no discriminative information and increase training time. Then, features with missing values or not-a-number (NaN) values are dropped, which decreases the feature space by 7.8% from 5558 to 5128 features. In a final step, the feature space was scaled using min-max-normalization since normalizing the feature space can have a positive impact on the model performance in various classification tasks [60], [61].

2) Built-In Feature Selection: The *tsfresh* package also offers built-in feature selection based on significance tests as part of the FRESH algorithm [57]. By default, a Mann-Whitney U-test [62] between the feature and the target variable (in this case the personality traits) is performed on the 95% level, and only statistically relevant features are kept. Since this step depends on the currently considered personality trait and therefore the set of informative features per trait might not be congruent, the built-in feature selection is repeated for all personality traits. Many features in the FRESH algorithm are parameterized and potentially similar, hence the problem of multicollinearity can arise, where many highly correlated features introduce a bias because the model attributes too much weight to this group of features.

To tackle the problem of multicollinearity, we combined the built-in feature selection with a correlation-based filtering approach as follows: first, pairwise Spearman correlation coefficients are calculated for all remaining features. Because of the presence of ordinal features, the Spearman correlation coefficient is preferred over the Pearson correlation coefficient. Next, for each pair where the correlation coefficient exceeds a threshold t, the feature with higher p-value from the previous Mann-Whitney U-test is dropped. Since the notion of high correlation is not general but domain- and problem-specific, we used a conservative threshold of $|t| > 0.8$ for the filtering. On average, the feature space was decreased by 78.6% (SD = 6.8%) using these feature selection steps, resulting in 794 (SD = 17) features for openness, 1239 (SD = 79) for conscientiousness, 1788 (SD = 37) for extraversion, 628 (SD = 25) for agreeableness, and 1238 (SD = 185) for neuroticism on average over 10 random splits, see Table III.

TABLE III
NUMBER OF OCCURRENCES OF EACH INPUT FEATURE AND HIGHER-ORDER FEATURE IN THE FINAL FEATURE SETS AVERAGED OVER 10 RANDOM SPLITS (STANDARD DEVIATION IS GIVEN IN BRACKETS) FOR *OPENNESS TO EXPERIENCE* (O), *CONSCIENTIOUSNESS* (C), *EXTRAVERSION* (E), *AGREEABLENESS* (A), AND *NEUROTICISM* (N)

		O	C	E	A	N
Time-based	DayTimestamp	3.4 (1.7)	4.5 (1.4)	1.2 (0.4)	0.0 (0.0)	1.0 (0.0)
	Typing Speed	3.3 (2.5)	7.2 (1.0)	20.7 (2.1)	3.7 (1.6)	1.0 (0.0)
	Elapsed Time	4.2 (2.1)	7.8 (1.4)	5.8 (0.7)	1.7 (0.5)	2.0 (0.0)
Position-based	Eucl. Distance	1.7 (0.7)	5.6 (2.1)	7.5 (1.9)	1.9 (0.5)	6.9 (1.4)
	x-Coordinate	4.2 (2.0)	11.6 (2.9)	6.5 (0.9)	11.4 (2.5)	8.6 (1.5)
	y-Coordinate	8.8 (2.4)	10.2 (1.1)	1.0 (0.0)	2.0 (1.1)	9.3 (0.9)
	Sum of x and y	4.5 (2.7)	1.2 (0.4)	1.0 (0.0)	1.1 (0.3)	10.8 (1.5)
Total No. of Features		28.3 (11.7)	50.1 (5.7)	43.9 (3.4)	22.4 (5.5)	38.7 (3.8)

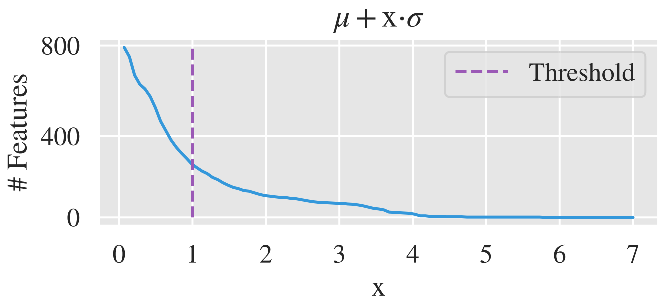


Fig. 4. Example of the empirical threshold for *openness* on the mutual information. The x-axis denotes the number of standard deviations σ above the mean μ. The threshold is chosen conservatively at one standard deviation.

3) Linear and Non-Linear Correlation Metrics: To further decrease the number of features and separate informative from uninformative features, we use two powerful approaches that have proven to be helpful in the context of feature selection in the past [63], [64]. First, we compute ANOVA F-statistics over the feature space and the target variable as suggested by Chen et al. [63], whereby a high F-score indicates a linear relationship between the feature and the target variable. Simultaneously, we use an information-theoretic approach introduced by Cover et al. [64] and compute the Mutual Information (MI) between the features and the target variable. Here, a high MI indicates a non-linear relationship between the feature and the personality trait. The cut-offs for the two methods are selected empirically by plotting potential cut-off values against the number of features whose scores (i.e., F-scores and MI) surpass the cut-off, and then setting the cut-off at the beginning of the resulting knee in the plot (see Fig. 4). For all personality traits, the cut-off was one standard deviation above the mean (± 0.25 standard deviations). The two feature sets are then intersected (see green branch in Fig. 3). In the end, 28.3 (SD = 11.7) features were selected for openness, 50.1 (SD = 5.7) for conscientiousness, 43.9 (SD = 3.4) for extraversion, 22.4 (SD = 5.5) for agreeableness, and 38.7 (SD = 3.8) for neuroticism on average over ten randomly generated splits, see Table III.

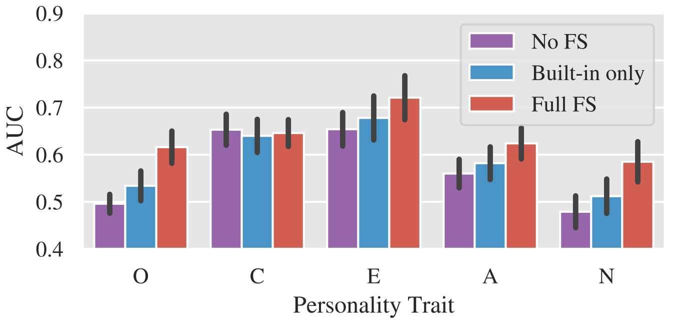


Fig. 5. Performance comparison of different feature selection methods: no feature selection, the built-in feature selection (FS1), and the full feature selection (FS1 and FS2) for all personality traits using one-day windows and grouped by *openness to experience* (O), *conscientiousness* (C), *extraversion* (E), *agreeableness* (A), and *neuroticism* (N). The error bars show the standard deviation from a 10-fold cross-validation.

D. Model Training

We generated ten randomly sampled and disjoint training-validation-test splits (70% training, 20% validation, 10% test). Stratification was used to ensure comparable split sizes. For each split, we extracted the relevant features on the training set using the feature extraction and selection process described before. We then trained an SVM classifier with an RBF kernel on each of the training sets and evaluated the model performance on the corresponding validation sets using the previously extracted set of features. We used bootstrap aggregation (bagging) to increase the model stability. The hyperparameters were tuned using Optuna [65], an optimization framework that tests different hyperparameter sets while optimizing a given objective function. We used the area under the receiver operating characteristic curve (AUC) as the objective function and for assessing the model performance (random level = 0.5). In contrast to accuracy, AUC is not affected by class imbalance. Since the optimization problem might not be convex, Optuna can run into a local optimum. Therefore, the optimization was repeated five times and the model parameters with the highest AUC were selected (see the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2023.3253202, for an extensive list of the hyperparameters). The best hyperparameter set was then used for prediction on the previously unseen test sets.

V. RESULTS

A. Feature Selection

We investigate the impact of our feature selection techniques by comparing the model performance when using no feature selection at all, when using only the built-in feature selection (FS1) provided by the *tsfresh* package, and when using the full feature selection pipeline (FS1 and FS2), which includes feature filtering based on the F-score and the mutual information (see Fig. 3). Fig. 5 shows that for four out of five personality traits the improvements of FS1 and FS2 are substantial compared to when no feature selection is used (+0.064 AUC for agreeableness, +0.12 AUC for openness, +0.106 AUC for neuroticism and

+0.067 AUC for extraversion). However, there is no improvement for conscientiousness. In summary, we conclude that the additional feature selection steps (FS1 and FS2) are necessary for removing uninformative features and thereby increasing the model performance.

B. Feature Importance

We investigate the importance of each input feature and higher-order feature (see Table II for the definitions) by counting their occurrences in the final feature sets after performing the full feature selection. Table III shows the occurrences grouped by personality trait and input features, or respectively, higher-order features. Time-related features were very important for extraversion, especially the typing speed (47% of all features on average). For openness, agreeableness and neuroticism, position-related features were of high relevance (62%, 76%, and 92% of all features on average, respectively). For conscientiousness, a mix of both spatial and time-related features were indicative. We conclude that different input and higher-order features are important for each personality trait.

C. Moving Average

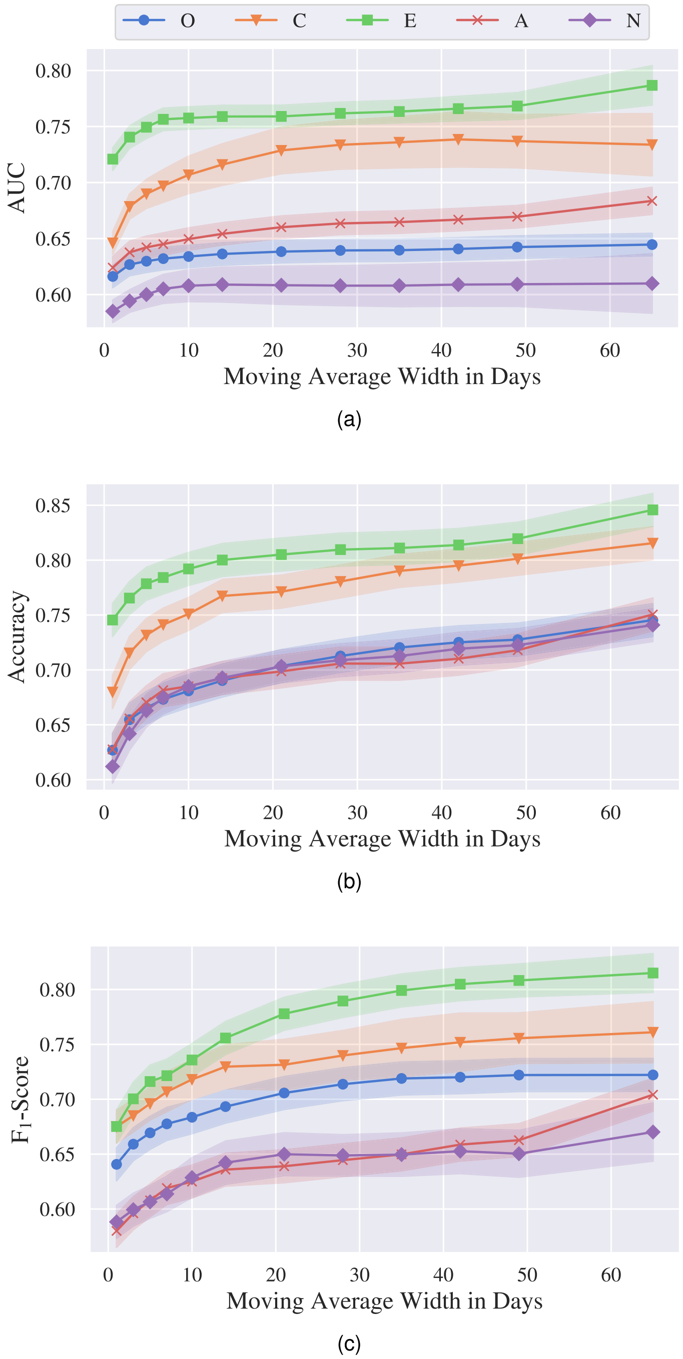
Since personality traits are merely behavioral tendencies rather than strict patterns [1], the typing characteristics of certain days could constitute outlier days. For example, a neurotic person might be unusually calm and relaxed during holidays, or an extroverted and open person might behave unsociably due to fatigue and drowsiness after sleeping poorly. Thus, we propose to smooth the predictions of the SVM over multiple days to account for such outlier days by leveraging previous predictions as prior knowledge and increasingly improving future predictions using a moving average on the previous w predictions:

$$\hat{p}_i = \frac{1}{\min(w,i)+1} \sum_{k=0}^{\min(w,i)} p_{i-k}, \qquad (1)$$

where p_i denotes the prediction at Day i. The predictions of all days are weighted equally since the personality traits are considered to be constant in short time frames and the predictions are independent of each other. Hence, each prediction is *a priori* equally informative. Fig. 6 shows the performance improvement in terms of AUC, accuracy, and F_1-score when averaging the predictions of the previous w days (random chance level at 0.5 for all metrics over all averaging widths due to balanced folds). The performance improves for all traits significantly, especially in the first ten days after which the AUC plateaus. This indicates that ten days are enough to effectively extract stable personality trait patterns. Averaging more than ten days has almost no effect on the AUC. The absolute performance increase after applying the moving average is reported in Table IV).

D. Runtime

For our model to be applicable in a real-world application, certain runtime requirements need to be met. For example, the feature extraction should be computed in a feasible amount of time to allow deployment on mobile devices such as smartphones. We report an extraction time of 4.5 minutes (SD = 3.2 minutes) per participant on an eight-core Intel Xeon E5-2697v4 with 128 GB of memory when extracting all available features over the entire duration of the study. This boils down to approximately 30 seconds per sample for all features on a single core, and a few milliseconds if only the relevant features are extracted.



(a)

(b)

(c)

Fig. 6. The effect of using a moving average of different widths in days on the model performance in terms of (a) AUC, (b) accuracy, and (c) F_1-score grouped by the personality traits *openness to experience* (O), *conscientiousness* (C), *extraversion* (E), *agreeableness* (A), and *neuroticism* (N). The standard deviation from the 10-fold cross-validation is shown in shades. The random chance level is at 0.5 for all averaging widths.

TABLE IV
INCREASE IN AUC, ACCURACY, AND F_1-SCORE FOR ALL TRAITS AFTER APPLYING THE MOVING AVERAGE OVER ALL AVAILABLE DATA POINTS

	AUC	Accuracy	F_1-Score
Openness	+0.03	+11.9%	+0.08
Conscientiousness	+0.09	+13.6%	+0.09
Extraversion	+0.07	+10.0%	+0.14
Agreeableness	+0.06	+12.3%	+0.12
Neuroticism	+0.02	+12.9%	+0.08

Once the SVM is trained on the dataset, the prediction time is in the range of a few milliseconds, which is negligibly small. The training time itself is irrelevant for real-world applications since training is carried out beforehand and depends on the size of the dataset and the number of iterations for optimizing the hyperparameters.

VI. DISCUSSION

In this work, we presented a novel method for inferring the personality traits of the Five Factor Model from in-the-wild smartphone typing characteristics. We presented a system with a fully automated feature extraction, selection and machine learning pipeline that achieves high classification accuracy of up to 74.5% accuracy and 0.72 AUC for a single day of data collection, and up to 84.5% accuracy and 0.79 AUC after subsequent refinement over 10 weeks on binary personality trait labels. In addition to the high prediction accuracy and the short data collection time of a few days, we also discuss in more detail other insights our method offers, such as the type of features that are predictive for each trait, and how previous predictions can be integrated into a long-term aggregation scheme. We conclude our discussion with two potential applications of our method and the resulting implications for the users' privacy.

A. Feature Importance

Counting the occurrences of each higher-order feature in the final feature sets revealed that different higher-order features are relevant for each personality trait (see Section V-B). Especially for openness, agreeableness, and neuroticism, position-related features were most relevant (17.4 out of 28.3 features on average for openness, 17.0 out of 22.4 features on average for agreeableness, and 35.7 out of 38.7 features on average for neuroticism). Position-related features indicate that the typed content was predictive. Since none of the extracted features in the FRESH algorithm explicitly targeted the semantics of the typed text, we believe that special characters (e.g., the space bar, shift key, backspace key, and punctuation characters) were most predictive for neuroticism and openness. For example, neurotic people might tend to use the backspace character much (more) less often than less neurotic people, indicating that they (often) seldom misspelled words and subsequently corrected their input, which would be reflected in a high x- and low y-coordinate due to the backspace character being located in the lower right corner of the keyboard. Using the space bar (more) less often than usual is also reflected by low y-coordinates, indicating that the person used rather (short) long words.

In contrast, for extraversion, we conclude that the typing speed is most relevant since on average 20.7 out of 43.9 features were typing speed-related. In particular, one of the selected features from all feature sets of extraversion was the mean typing speed, which was positively correlated to extraversion (Spearman correlation coefficient $r = 0.31$, $p = 5.2 \cdot 10^{-118}$), indicating that extroverted people tend to type faster than introverted people.

In summary, these results highlight the importance of considering each personality trait individually by training separate classifiers since it allows for a tailored and refined feature selection and feature importance analysis.

B. Moving Average

The moving average improves the prediction accuracy for all personality traits monotonically over time (see Fig. 6). While the accuracy increases monotonically with increasing averaging window sizes, the AUC plateaus after ten days for all personality traits, indicating that averaging more than ten days of data does not yield any additional gain in terms of model performance. We conclude that a time period of at least 10 days is necessary to extract stable trait predictions from everyday typing data. The discrepancy between the AUC and accuracy can be attributed to the fact that accuracy is based on an optimal threshold for separating the output probabilities of the SVM into two classes, whereas the AUC considers many possible thresholds simultaneously, some of which might decrease the overall AUC. While the optimal separation threshold is known at training time, it is not known when deploying the model in a real-world application and has to be fixed beforehand. The same holds for the F_1-score which also depends on a single separation threshold.

C. Implications and Potential Applications

Our method leverages short-term variability of smartphone typing characteristics and integrates it into a long-term aggregation scheme, which enables the deployment in multiple applications where data is collected over long periods, but the personality assessments should be available right away. In the following, we detail two applications that could take advantage of the personality assessments of our system.

1) Mental Health: In psycho-therapeutic smartphone applications, certain facets of pre-selected personality traits are changed over time using different therapeutic schemes. Tracking of the personality over time and a fast and accurate personality assessments are desirable. For example, PEACH [66] is a smartphone application for chatbot-based personality coaching. The application uses in-app personality tests to asses the users' personality. Our model could automate these personality tests based on the text conversations with the chatbot (or conversations from other chat applications if the users provide consent), thus making the interaction with the application less demanding and more natural. Furthermore, our method would allow for an automatic tracking of the traits over time, capturing the natural personality trait drifts occurring in young adulthood [5], [6].

2) Recommender Systems: In personality-aware recommender systems, product suggestions could be made after the first day while refining the suggestions over time using our aggregation scheme without the need to first collect user preferences over longer time spans to output a suitable recommendation, which would benefit both the users and the service providers. For example, YouTube recommends video clips based on the past usage history and user preferences. Assuming consent of the user, our model could be used to improve such video clip recommendations and tailor them to the users' personalities.

3) User Privacy: While our method has numerous potential applications and benefits, it also raises privacy issues. Smartphone users are increasingly concerned about their privacy and want to remain in control over what data they share [67]. Thus, passive recording of touch positions that could potentially leak information about the keystrokes might not be easily adopted and accepted by the users. However, once trained, our model can be deployed fully offline on a user's smartphone due to the light-weight SVM model running on a CPU. Furthermore, the collected data can be instantly deleted after a prediction is obtained, both for privacy and storage reasons. Finally, users can maintain full control over the data collection by pausing and resuming the data recording at any time, and the source code of the application can be made openly available. We believe that these countermeasures can greatly reduce user's privacy concerns. However, personality traits are also considered privacy-sensitive information [28], and we believe that users should be made aware of the usage and functionality of personality trait recognition. More specifically, users should be made aware that the personality traits can be inferred through passively analyzing typing characteristics without semantically analyzing the typed text nor collecting external context data (e.g., location and application usage data), as shown in this work.

VII. Conclusion

In this work, we showed how typing characteristics from in-the-wild smartphone data can be used to classify the personality traits of the Five Factor Model on two levels (low and high) with an accuracy of up to 84.5% and an AUC of up to 0.79. To the best of our knowledge it is the first approach that investigates the relationship between daily typing behavior and the personality traits of the Five Factor Model in the wild.

In addition, our model surpasses related work in two key aspects. First, our method comes with a fully automated feature extraction and selection pipeline that does not require any manual feature engineering or feature selection while at the same time provably succeeding in finding a set of predictive features for the classification task at hand. Second, our approach constitutes an important step towards direct deployment in real-world applications since it decreases the required data collection time of current models from one month of data collection to only a few days of data collection by leveraging short-term typing characteristics, and then subsequently refining and improving the prediction by merging previous predictions using a moving average. In particular, our method shows that personality trait patterns become stable after collecting 10 days of typing data.

Our work highlights the importance of considering separate classification models for each individual personality trait since the traits differ in the types of features that are most predictive. Future work may take these insights into consideration when modeling their classification schemes as it allows for more fine-grained and accurate results in the context of personality trait recognition.

VIII. Limitations & Future Work

In the following, we discuss potential limitations of our work. First, the acquired labels for the dataset do not necessarily reflect the real ground truth because participants may have tried to achieve high scores on socially desirable characteristics, and low scores on undesirable characteristics. Although there was no incentive for the participants to intentionally distort their test scores, we cannot rule out a potential bias. Furthermore, the small sample size of 76 participants, 68 of which were university students, may not be representative of the general population, which can be observed by the skew on openness, agreeableness, and neuroticism.

Second, the feature selection pipeline filters the features based on their direct relationship with the target variable although multiple features combined might contain more discriminative power than each feature separately. Although we tackled this problem by combining the input time series into higher-order features before the feature extraction, the high number of extracted features from the FRESH algorithm impedes a detailed analysis of all feature combinations. Furthermore, our method does not account for potential dependencies among subsequent days. Future work might investigate a way to efficiently equip our pipeline with the ability to combine features during the selection process and consider dependencies of data points of users when applying the moving average.

Finally, our work is based on the premise that users regularly type on their smartphones in order to produce a sufficient amount of data for the personality trait inference. However, a major paradigm shift in the way people use their phones might be under way. For example, people might increasingly start to replace typed messages by voice messages, or use different emerging interaction technologies such as smartwatches or smartglasses. Future work might investigate other input modalities apart from typing patterns (e.g., voice or gesture input) and adapt our pipeline accordingly to enable multimodal personality trait recognition.

References

[1] G. W. Allport, "Concepts of trait and personality," *Psychol. Bull.*, vol. 24, no. 5, pp. 284–293, 1927.

[2] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychol. Assessment*, vol. 4, no. 1, pp. 26–42, 1992.

[3] L. R. Goldberg, "The structure of phenotypic personality traits," *Amer. Psychol.*, vol. 48, no. 1, 1993, Art. no. 26.

[4] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality Traits*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[5] B. W. Roberts and D. Mroczek, "Personality trait change in adulthood," *Curr. Directions Psychol. Sci.*, vol. 17, no. 1, pp. 31–35, 2008.

[6] C. J. Soto, O. P. John, S. D. Gosling, and J. Potter, "Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample," *J. Pers. Social Psychol.*, vol. 100, no. 2, pp. 330–348, 2011.

[7] M. A. Harris, C. E. Brett, W. Johnson, and I. J. Deary, "Personality stability from age 14 to age 77 years," *Psychol. Aging*, vol. 31, no. 8, pp. 862–874, 2016.

[8] M. Braunhofer, M. Elahi, and F. Ricci, "User personality and the new user problem in a context-aware point of interest recommender system," in *Information and Communication Technologies in Tourism 2015*. Berlin, Germany: Springer, 2015, pp. 537–549.

[9] J. A. Recio-Garcia, G. Jimenez-Diaz, A. A. Sanchez-Ruiz, and B. Diaz-Agudo, "Personality aware recommendations to groups," in *Proc. 3rd ACM Conf. Recommender Syst.*, New York, NY, USA: Association for Computing Machinery, 2009, pp. 325–328.

[10] B. Ferwerda, M. Tkalcic, and M. Schedl, "Personality traits and music genres: What do people prefer to listen to?," in *Proc. 25th Conf. User Model. Adapt. Personalization*, 2017, pp. 285–288.

[11] M. A. S. Nunes and R. Hu, "Personality-based recommender systems: An overview," in *Proc. 6th ACM Conf. Recommender Syst.*, 2012, pp. 5–6.

[12] R. R. McCrae and P. T. Costa, *Personality in Adulthood: A Five-Factor Theory Perspective*. New York, NY, USA: Guilford Press, 2003.

[13] L. S. Martin, L. G. Oades, and P. Caputi, "Intentional personality change coaching: A randomised controlled trial of participant selected personality facet change using the five-factor model of personality," *Int. Coaching Psychol. Rev.*, vol. 9, no. 2, pp. 196–209, 2014.

[14] R. Hogan, J. Hogan, and B. W. Roberts, "Personality measurement and employment decisions: Questions and answers," *Amer. Psychol.*, vol. 51, no. 5, pp. 469–477, 1996.

[15] J. F. Youngman, "The use and abuse of pre-employment personality tests," *Bus. Horiz.*, vol. 60, no. 3, pp. 261–269, 2017.

[16] M. X. Zhou, G. Mark, J. Li, and H. Yang, "Trusting virtual agents: The effect of personality," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 2-3, pp. 1–36, Mar. 2019, doi: 10.1145/3232077.

[17] K. R. Black, "Personality screening in employment," *Amer. Bus. Law J.*, vol. 32, pp. 69–124, 1995.

[18] N. B. Afini Normadhi, L. Shuib, H. N. Md Nasir, A. Bimba, N. Idris, and V. Balakrishnan, "Identification of personal traits in adaptive learning environment: Systematic literature review," *Comput. Educ.*, vol. 130, pp. 168–190, 2019.

[19] R. McCrae and O. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, Jun. 1992.

[20] O. P. John et al., *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*. Univ. California Berkeley, Berkeley, California, 1999.

[21] P. T. Costa and R. R. McCrae, "Four ways five factors are basic," *Pers. Individual Differences*, vol. 13, no. 6, pp. 653–665, 1992.

[22] R. R. McCrae, "The five-factor model of personality traits: Consensus and controversy," in *The Cambridge Handbook of Personality Psychology*. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 148–161.

[23] C. J. Soto and O. P. John, "The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power," *J. Pers. Social Psychol.*, vol. 113, no. 1, pp. 117–143, 2017.

[24] P. T. Costa and R. R. McCrae, "NEO five-factor inventory (NEO-FFI)," *Psychol. Assessment Resour.*, vol. 3, 1989.

[25] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the big-five personality domains," *J. Res. Pers.*, vol. 37, no. 6, pp. 504–528, 2003.

[26] R. S. Woodworth, "Autobiography of Robert S. Woodworth," *Int. Univ. Ser. Psychol. A Hist. Psychol. Autobiography*, vol. 2, pp. 359–380, 1932.

[27] G. Fahey, "Faking good and personality assessments of job applicants: A review of the literature," *DBS Bus. Rev.*, vol. 2, pp. 45–68, 2018.

[28] S. T. Völkel, R. Haeuslschmid, A. Werner, H. Hussmann, and A. Butz, "How to trick AI: Users' strategies for protecting themselves from automatic personality assessment," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.

[29] F. P. Morgeson, M. A. Campion, R. L. Dipboye, J. R. Hollenbeck, K. Murphy, and N. Schmitt, "Reconsidering the use of personality tests in personnel selection contexts," *Personnel Psychol.*, vol. 60, no. 3, pp. 683–729, 2007.

[30] M. Ziegler, C. MacCann, and R. D. Roberts, *New Perspectives on Faking in Personality Assessment*. Oxford, England: Oxford Univ. Press, 2011.

[31] L. Küster, C. Trahms, and J.-N. Voigt-Antons, "Predicting personality traits from touchscreen based interactions," in *Proc. 10th Int. Conf. Qual. Multimedia Experience*, 2018, pp. 1–6.

[32] S. Berkovsky et al., "Detecting personality traits using eye-tracking data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.

[33] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Who's who with big-five: Analyzing and classifying personality traits with smartphones," in *Proc. IEEE 15th Annu. Int. Symp. Wearable Comput.*, 2011, pp. 29–36.

[34] I. A. Khan, W.-P. Brinkman, N. Fine, and R. M. Hierons, "Measuring personality from keyboard and mouse use," in *Proc. 15th Eur. Conf. Cogn. Ergonom.: Ergonom. Cool Interact.*, 2008, pp. 1–8.

[35] C. Montag et al., "Smartphone usage in the 21st century: Who is active on whatsapp?," *BMC Res. Notes*, vol. 8, no. 1, pp. 1–6, 2015.

[36] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Pers. Ubiquitous Comput.*, vol. 17, no. 3, pp. 433–450, 2013.

[37] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, M. Gross, and C. Holz, "Affective state prediction from smartphone touch and sensor data in the wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–14.

[38] R. Kaplan and D. Saccuzzo, *Psychological Testing: Principles, Applications, and Issues*. Boston, MA, USA: Cengage Learning, 2008.

[39] D. Schmitt et al., "The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations," *J. Cross-Cultural Psychol.*, vol. 38, pp. 173–212, 2007.

[40] P. T. Costa and R. R. McCrae, *The Revised NEO Personality Inventory (NEO-PI-R)*. Thousand Oaks, CA, USA: SAGE, 2008, pp. 179–198.

[41] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, Jul./Sep. 2014.

[42] L. V. Phan and J. F. Rauthmann, "Personality computing: New frontiers in personality assessment," *Social Pers. Psychol. Compass*, vol. 15, no. 7, Jun. 2021.

[43] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "Tensorflow-based automatic personality recognition used in asynchronous video interviews," *IEEE Access*, vol. 7, pp. 61 018–61 023, 2019.

[44] M.-A. Carbonneau, E. Granger, Y. Attabi, and G. Gagnon, "Feature learning from spectrograms for assessment of personality traits," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 25–31, First Quarter 2020.

[45] Y. Zhao, D. Miao, and Z. Cai, "Reading personality preferences from motion patterns in computer mouse operations," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1619–1636, Second Quarter 2022.

[46] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Second Quarter 2018.

[47] W. Li et al., "Quantitative personality predictions from a brief EEG recording," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1514–1527, Second Quarter 2022.

[48] G. Zhao, Y. Ge, B. Shen, X. Wei, and H. Wang, "Emotion analysis for personality inference from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 362–371, Second Quarter 2018.

[49] T. Tandera, D. HendroR. SuhartonoWongso, and Y. L. Prasetio, "Personality prediction system from facebook users," *Procedia Comput. Sci.*, vol. 116, pp. 604–611, 2017.

[50] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, "Fusing social media cues: Personality prediction from twitter and instagram," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 107–108.

[51] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, KNN and SVM," in *Proc. Int. Conf. Data Softw. Eng.*, 2015, pp. 170–174.

[52] G. Farnadi et al., "Computational personality recognition in social media," *User Model. User-Adapted Interact.*, vol. 26, no. 2, pp. 109–142, 2016.

[53] D. Danner et al., "Die deutsche version des big five inventory 2 (BFI-2)," Mannheim, Germany: GESIS – Leibniz–Institut für Sozialwissenschaften 2016, p. 20, doi: 10.6102/zis247.

[54] T. Chamorro-Premuzic and A. Furnham, "Personality predicts academic performance: Evidence from two longitudinal university samples," *J. Res. Pers.*, vol. 37, no. 4, pp. 319–338, 2003.

[55] R. Wever, "Characteristics of circadian rhythms in human functions," *J. Neural Transmiss. Supplementum*, vol. 21, pp. 323–373, 1986.

[56] G. M. Harari et al., "Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life," *J. Pers. social Psychol.*, vol. 119, no. 1, pp. 204–228, 2020.

[57] M. Christ, A. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial Big Data applications," 2016, *arXiv:1610.07717*.

[58] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[59] "Tsfresh python package," Accessed: Dec. 01, 2021. [Online]. Available: https://tsfresh.readthedocs.io/en/v0.18.0/index.html

[60] A. B. A. Graf, A. J. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 597–605, May 2003.

[61] S. Ali and K. A. Smith-Miles, "Improved support vector machine generalization using normalized input space," in *Proc. Australas. Joint Conf. Artif. Intell.*, 2006, pp. 362–371.

[62] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.

[63] Y.-W. Chen and C.-J. Lin, *Combining SVMs With Various Feature Selection Strategies*. Berlin, Germany: Springer, 2006, pp. 315–324.

[64] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1999.

[65] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," 2019, *arXiv: 1907.10902*.

[66] M. Stieger, M. Nißen, D. Rüegger, T. Kowatsch, C. Flückiger, and M. Allemand, "PEACH, a smartphone-and conversational agent-based coaching intervention for intentional personality change: Study protocol of a randomized, wait-list controlled trial," *BMC Psychol.*, vol. 6, no. 1, pp. 1–15, 2018.

[67] P. E. Ketelaar and M. van Balen, "The smartphone as your follower: The role of smartphone literacy in the relation between privacy concerns, attitude and behaviour towards phone-embedded tracking," *Comput. Hum. Behav.*, vol. 78, pp. 174–182, 2018.

Nikola Kovačević received the BSc and MSc degrees in computer science from ETH Zurich, in 2019 and 2021, respectively. He is currently working toward the doctoral degree with the Computer Science Department, ETH Zurich, Switzerland. His research interests include affective computing, human-computer interaction, and conversational AI.

Christian Holz received the PhD degree in computer science from the University of Potsdam, Germany, in 2013. He is an assistant professor in computer science with ETH Zurich, where he leads the Sensing, Interaction & Perception Lab. His research investigates the interface between the virtual world and the physical world of people and objects.

Tobias Günther received the PhD degree in scientific visualization from the University of Magdeburg, in 2016. He is a professor of computer science with the Friedrich-Alexander-University Erlangen-Nuremberg, Germany. His research interests include visualization, light transport, and real-time rendering.

Markus Gross received the PhD degree in computer graphics and image analysis from Saarland University, in 1989. He is a professor of computer science with ETH Zurich, head of the Computer Graphics Laboratory, and the director of Disney Research Studios, Zurich. His research interests include physically based modeling, computer animation, immersive displays, and video technology.

Rafael Wampfler received the PhD degree in computer science from ETH Zurich, in 2021. He is a postdoctoral researcher with the Computer Science Department, ETH Zurich, Switzerland. His research interests include affective computing, human-computer interaction, applied machine learning, and digital avatars.