Watch your Watch: Inferring Personality Traits from Wearable Activity Trackers

Noé Zufferey[1], Mathias Humbert[1], Romain Tavenard[2], Kévin Huguenin[1]
[1]*University of Lausanne, Switzerland*
[2]*University of Rennes, CNRS, LETG, France*

Abstract

Wearable devices, such as wearable activity trackers (WATs), are increasing in popularity. Although they can help to improve one's quality of life, they also raise serious privacy issues. One particularly sensitive type of information has recently attracted substantial attention, namely personality, as it provides a means to influence individuals (e.g., voters in the Cambridge Analytica scandal). This paper presents the first empirical study to show a significant correlation between WAT data and personality traits (Big Five). We conduct an experiment with 200+ participants. The ground truth was established by using the NEO-PI-3 questionnaire. The participants' step count, heart rate, battery level, activities, sleep time, *etc*. were collected for four months. By following a *principled* machine-learning approach, the participants' personality privacy was quantified. Our results demonstrate that WATs data brings valuable information to infer the openness, extraversion, and neuroticism personality traits. We further study the importance of the different features (i.e., data types) and found that step counts play a key role in the inference of extraversion and neuroticism, while openness is more related to heart rate.

1 Introduction

The number of wearable device[1] owners is increasing every day. The International Data Corporation states that global shipments of wearable devices reached 138.4 million units during the third quarter of 2021 [6], which means that there are more than one billion wearable devices worldwide [5]. These devices collect large amounts of physiological and contextual data, such as step counts and continuous heart rate (for those equipped with the appropriate sensors). Such data can help wearable device users to better monitor their physical activities and health, following a *quantified-self* [25] approach. However, wearable devices raise new privacy and security issues. For instance, Eberz et al. [40] showed that data collected from wearable devices can be used to bypass biometric authentication systems by using accelerometer data to impersonate users. Furthermore, accelerometer data can be used to infer keystrokes (e.g., on pin-pads) [72, 73, 74]. Moreover, WAT data can be used to infer daily activities and habits [7, 43, 64, 85] (e.g., eating) and drug usage [86] (e.g., cocaine), and even to identify SARS-CoV-2 infections [56]; such inferences are highly sensitive from a privacy perspective. Finally, WAT data, such as running routes, can be used to infer sensitive locations (e.g., user's home), even when using protection mechanisms [51, 81], and aggregated location data have been used to locate military bases and infer their internal structures [54], specifically in remote areas where unusual activity patterns were observed.

In the context of the *quantified-self*, questioning the effect of such data collection (and sharing) on people's privacy is becoming increasingly relevant, especially as many users express concerns about the misuse of their data [11, 45]. Personal information, such as personality, socioeconomic status, sexual orientation, and religion can probably be inferred from data collected by wearable devices, similarly to what was shown to be possible for location and social network data (e.g., [18, 82, 119]). Moreover, third-party entities like advertisers, marketers, health insurers, employers, and governments might have an interest in learning sensitive information derived from the data collected by wearables [2]. Some organizations, encouraged in particular by Fitbit (one of the market leaders for WATs [53, 70]), are now offering their employees tracking devices through health programs [97]. More recently, the former US President Trump suggested using data from wearable devices for national security purposes, essentially to preemptively detect mass shooters [41].

One particularly valuable type of personal information, as illustrated by the Cambridge Analytica scandal [21], is personality. Personality is often characterized by the Big Five OCEAN traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) [96], and it is known to influence behavior. Information about an individual's personality enables others to manipulate this individual more efficiently

[1] In particular, (wrist-worn) wearable activity trackers (WATs).

by sending them appropriate signals (e.g., targeted advertisements), thus raising serious ethical concerns. For instance, Cambridge Analytica used data from social networks to infer the personality traits of US voters and to influence them during the 2016 Presidential Election [46, 95]. Similarly, credit card companies have exploited clients' purchase history to profile debtors and craft the appropriate strategies to recover their debts [38] (e.g., by determining whether a specific client would respond better to a comforting or threatening message). As a result, individuals are increasingly worried about the potential misuses of automatic personality assessment [60]. Besides social networks, prior work has demonstrated that personality could be inferred from the data collected by individuals' (smart)phones [24, 34, 35, 83, 104].

In this work, we focus on the problem of personality inference in the context of WATs. As such devices collect a large amount of behavioral and physiological data, they bring valuable information to infer personality. Indeed, behavioral indicators are one of the three types of indicators that are used to assess an individual's personality [120]. Furthermore, previous research extensively studied the relationship between personality and physical activity [94] and identified multiple correlations between the two. Moreover, recent works show that WAT data can be used to infer characteristics related to personality, such as stress resilience [8] and mood [67]. It has also been shown that some personality traits are correlated to sleep [55]. Finally, WAT data can also be combined with other types of data that are already known to be helpful for personality inference (e.g., social network behavior, smartphone usage). Data brokers can indeed easily link different types of users' data from different databases [9] and build accurate inference models using such a larger and more diverse data set. To the best of our knowledge, this is the first work to address the concrete (privacy) risks of personality inference from data collected by WATs.

Contributions and Results

We present the first study on the inference of personality traits from data collected by WATs. We equipped 200+ volunteers with Fitbit wearable devices (namely, Fitbit's Inspire HR WAT) and captured their step counts, heart rate, battery level, activity, and sleep time over the course of a four-month period, as well as data available on their user profile, such as gender. To determine the personality profile of our participants, we used the Big Five personality scores captured through the standardized NEO-PI-3 questionnaire [77]. Our longitudinal data collection enabled us to precisely evaluate to what extent data collected by wearable devices are correlated to personality traits, and thus may be used together with other types of data, to conduct personality inference attacks.

In particular, we rely on a machine learning model trained on the data collected by the wearable devices to predict the given personality trait tercile. Although our model does not reach high levels of accuracy for any Big Five personality trait, it is evaluated using a rigorous leave-one-out (LOO) cross-validation, and we show that it can classify WAT users according to openness, extraversion, and neuroticism with statistical significance compared to the random-guess baseline. We also report on the most relevant features by analyzing those that are the most used by the inference model.

Furthermore, we collected our participants' concerns and perceptions regarding personality inference from their WAT data in an exit questionnaire. Nearly half of our participants thought that such inference would not be possible at all, while nearly two-third of them reported that they would be worried if it was. This is in line with a recent qualitative study, using interviews, that shows that a substantial fraction of users are worried about personality assessment and about its potential misuse [60]. Finally, we analyzed related prior work based on phone and smartphone data, discuss their methodologies, and compare our results and methodology to theirs. We show that the accuracy level achieved by our model outperforms that of the current state-of-the-art found in literature about (smart)phone-based inference using similar methodologies (ternary classification) [83] for all five personality traits. Furthermore, we are the first to show, with a rigorous evaluation process, correlations between wearable data and neuroticism and openness. Based on our analysis, we also discuss the design of potential privacy-preserving solutions.

2 Background

We provide the necessary background regarding two key aspects of our work: personality assessment and WATs.

2.1 Personality

The assessment of an individual's personality is generally based on the Big Five personality traits, also known as the five-factor model. The Big Five personality traits constitute a psychological model that defines an individual's personality through five main traits (specifically openness, conscientiousness, extraversion, agreeableness, and neuroticism; conveniently abbreviated OCEAN) that are subdivided into six sub-traits each [77]. This model, which has been proven to be robust and stable over time [28], is structured as follows [96]:

- **Openness** to experience — Individuals who score high on this dimension tend to be intellectual, imaginative, sensitive, and open-minded. Those who score low tend to be down-to-earth, insensitive, and conventional.
- **Conscientiousness** — Individuals who are high in conscientiousness tend to be careful, thorough, responsible, organized, and scrupulous. Those low on this dimension tend to be irresponsible, disorganized, and unscrupulous.
- **Extraversion** — Individuals who score high on extraversion tend to be sociable, talkative, assertive, and active.

Whereas, those who score low tend to be retiring, reserved, and cautious.

- **Agreeableness** — Individuals who score high on agreeableness tend to be good-natured, compliant, modest, gentle, and cooperative. Individuals who score low on this dimension tend to be irritable, ruthless, suspicious, and inflexible.
- **Neuroticism** — Individuals high on neuroticism tend to be anxious, depressed, angry, and insecure. Those low on neuroticism tend to be calm, poised, and emotionally stable.

The NEO-PI-3 (third version of the NEO-PI) is a standardized questionnaire for assessing an individual's personality, along the five aforementioned traits. It is considered to be a reference in the personality assessment research field [77]. The NEO-PI-3 is a 240-item questionnaire describing and analyzing the five main aforementioned personality traits.This questionnaire delivers, for each of the five personality traits, a score between 0 and 192. The Big Five personality traits and the NEO-PI questionnaires are deeply related and have been developed mainly by Costa and McCrae [29]. Official translations of this questionnaire exist in many languages. In this work, we used the official translation of the full NEO-PI-3 questionnaire, in [redacted for blind review], the local language at our institution.

2.2 Wearable Activity Trackers

Wearable Activity Trackers (WATs) are wearable devices designed to collect diverse physiological and contextual data about their users and are generally acknowledged to be decently accurate [10]. People usually wear such devices either to increase their physical performance, to improve their quality of life and lifestyle, or simply because they like collecting data about their life and habits. This type of device collects a large amount of diverse data such as step count, heart rate, activities, and sleep time (we expand more on this in Section 6.2).

Fitness-tracking devices generally collect an individual's data by using multiple embedded sensors such as accelerometers, gyroscopes, ambient light, and temperature sensors. These data are then sent via Bluetooth to a synchronized smartphone or tablet where the data are processed and stored by the corresponding application. Generally, this application also transfers the user's data to a server either to facilitate data sharing, prevent possible data loss, or for further processing and analytics in order to provide additional services/insights to the user.

With its 29.6 millions of users in 2019 [32], Fitbit is considered one of the leaders of the WAT market. In the case of Fitbit's services, in addition to the standard functionalities that they provide, the possibility to grant read/write access authorization to third parties is given to users so that third parties can access the user's data through the dedicated Fitbit API [4] to provide supplementary services or data analysis. To this end, Fitbit relies on the OAuth 2.0 protocol. During the authorization process, a user can choose which type of data they agree to share with a third party.

3 Adversarial Model

We focus on an adversary that can access some or all of a users' data processed by Fitbit. There exist multiple adversaries who correspond to this description. One such adversary is typically the service provider itself, like Fitbit or other companies that base their business on WAT data collection such as WeWard, that offers their users to be paid according to the number of steps they take [1]. In this case, the risks we can measure represent a lower bound of the actual risks as the service provider has access to the raw WAT data and the smartphone data collected by the companion mobile app. It could also be any of its business partners, or any third party to whom many users have granted, knowingly or not, access to their data (e.g., have given a token pair through OAuth 2.0). Such an adversary can potentially obtain years of data collected from millions of users (there were 29.6 million Fitbit users in 2019 [32] and, 4 millions for WeWard in 2021 [1]). A recent study shows that 70% of WAT users share their data with at least one third-party app [123], and that users who share their data with third-party apps tend to forget that they do. Furthermore, it also shows the users' lack of knowledge about the data sharing process and demonstrates that they are not aware of the actual amount of data they share. Moreover, 9% of the participants in this study claimed to grant Fitbit access to at least one of their social media accounts, so that Fitbit can automatically make posts on their social media profiles related to their activity (e.g., step counts). An adversary could use such information, alone or combined with other information available on the social profiles [61, 62, 69], to infer users' personality. Also, an employer could offer free WATs to their employees if they accept to share the collected data with their employer. Over the past few years US companies have engaged in such corporate-wellness programs using Fitbit devices [97]. A government could gain access to a WAT service provider's data, for national security reasons, as recently suggested by a former US president [41]. An insurance company (e.g., health) could provide tracking devices to their policyholders to better analyze risks. For instance, Google acquired Fitbit [91] and Alphabet, Google's parent company, is growing rapidly in the health insurance market [20], furthermore, they plan to force Fitbit users to migrate to their Google accounts [111]. Finally, other adversaries could obtain such information by accessing tracking-device companies' leaked databases or by using eavesdropping techniques, as WATs and their related mobile applications are known to use poorly protected wireless communication protocols and data storage [15, 27, 33, 47, 71, 118].

In this article, we consider one such adversary who subsequently uses the collected data to infer the psychological profiles of the associated users. Such personal information

is highly sensitive, from a security and privacy point of view as explained in the introduction. This information is highly valuable for adversaries, thus pushing them to conduct such attacks. In particular, psychological profiling enables discrimination and manipulation. Indeed, assessing an individual's personality can help influence their behavior. For instance, it can be used to influence consumers' choices through targeted advertisements [37, 38] and even voters' choices [116] as in the Cambridge Analytica scandal related to the 2016 US presidential election [21], and thus have an impact beyond manipulating individuals, by influencing global politics.

4 Related Work

Prior research shows that data collected by wearable devices such as WATs can be used to infer information which, although sometimes useful and desired by the user, can be considered as sensitive or can cause security breaches. For example, data from altimeters can be used to reveal the user's location [78] and data from sensors such as accelerometers and gyroscopes can be used to monitor individuals' activities [63, 102]. Such data can also be used to infer more precise information, such as which keys a user pressed on a keyboard (e.g., computer keyboard, smartphone keypad, ATM pin pad) [68, 73, 74, 75, 98], handwritten text [12], food consumption [113], alcohol consumption [48], or smoking [103]. Sensor data can be used to impersonate an individual by allowing an attacker to imitate the user's biometrics in order to bypass identification systems [39, 40] or to study individuals' behavior at work [84]. It can also help to monitor individuals' sleep quality [101], health (by inferring the presence of diseases [63, 88, 115]), mental state (such as their levels of stress [115]), and to identify SARS-CoV-2 infections [56].

In parallel, another line of research shows that an individual's personality can be inferred from various types of data [57, 108]. It can be inferred from location-based social media or location logs (e.g., Foursquare logs) [26, 110], from online social-network profiles, networks, and behavior (e.g., number of "friends", likes, sharing) [61, 62, 69], from pictures (e.g., social-media profile picture) [22, 49], from nonverbal-speech feature analysis (everything except the speech content), from speech features (such as prosody and intonation), body features (such as head or hand movements) [16, 58, 117], from written texts (e.g., Facebook status and posts, Tweets) [52, 79, 80]. Dietary habits were shown to be correlated with personality [114]; therefore, this correlation could be exploited to predict personality from dietary habits reported in the WAT app or detected from the tracker data. Finally, personality can be predicted from call-detail records (CDR) and smartphone data [23, 34, 35, 83, 105]. Below, we review the articles related to data collected by (smart)phones in more detail, as these data are the most similar to WAT data (yet much richer).

Prior studies about mobile-phone-related data highlighted the link between collected personal data and personality traits. Table 1 compares all the related-work experimental layout and results that we discuss in detail next.

de Oliveira et al. studied to which extent it is possible to infer personality traits from call-detail records using regression. Their model obtained mean square errors (MSE) significantly ($p < 0.05$) lower than the baseline (MSE of 1.184) for openness (MSE of 0.670), extraversion (MSE of 0.650), and agreeableness (MSE of 0.615) [35].

Chittaranjan et al. evaluated the accuracy of personality-trait inference from smartphone data by using binary classification methods [23, 24]. They obtained an average accuracy of 72% (+25% of accuracy compared to the baseline on average) for all traits.

de Montjoye et al. evaluated the accuracy of personality-trait inference from phone-based metrics by using ternary classification methods [34]. They obtained an average accuracy of 53% (+42% of accuracy compared to the baseline on average) for all traits.

However, Mønsted et al. show that the inference results were overestimated in the aforementioned articles [23, 34, 35]. More specifically, the authors of these works optimized some parameters (e.g., feature, model, and hyperparameter selection) based on the *entire* dataset instead of doing so based on only the training set considered in each iteration of the cross-validation loop; this corresponds to the common pitfalls P3 and P5 listed in Arp et al.'s recent work on the dos and don'ts of machine learning in computer security [13]. Mønsted et al. further proceed to a ternary classification of the five traits by using the same models, features, and approach as de Montjoye et al.'s article [34]. They show that, based on their correlation with the trait to infer without using cross-validation (i.e., on the entire dataset), previous research about inferring personality from phone data overestimated model performance by selecting certain features. After following the same approach and obtaining similar results to de Montjoye et al., Mønsted et al. show that by using a more rigorous methodology with the same data, only extraversion can be inferred (with an accuracy significantly better than the baseline) from (smart)phone data. They obtained an accuracy improvement of +36% (wrt the baseline) for that specific trait. Therefore, we cannot compare our work with their results, except for those of Mønsted et al. who used a (rigorous) methodology similar to ours. Hence, we can assert that personality inference models using WAT data outperform those using CDR as they achieve a higher accuracy for extraversion as well as accuracies significantly higher than the baseline for neuroticism and openness.

More recently, Stachl et al. inferred personality traits from richer smartphone data [104] using smartphone data of 624 participants collected over 30 days. Their features were more diverse and richer than those used in the other studies. The features were derived from call detail records, music consumption, application usage, mobility, overall phone activities, and

Table 1: Comparative table of the most relevant publications. The 'year' is the year of publication, the 'source' represents the data source used to build the features for the inference process, 'N' is the number of participants, 'var.' means that the data collection duration is not fixed among the different participants, 'CDR' stands for Call Detail Records, the inference type is either regression or classification, k is the number of classes in case of classification, 'SVR' stands for Support Vector Regression, 'SVC' for Support Vector Classification, 'RF' for Random Forest, and 'LOO' stands for Leave-One-Out evaluation. Finally, the 'Results' column shows, in bold, which traits were inferred statistically significantly better than their respective baseline.

Article	Year	Source	N	Dur.	Inference	Model	Eval.	Results
de Oliveira et al. [35]	2011	CDR	39	var.	Regression	SVR	10-fold	**OCEAN***
Chittaranjan et al. [23]	2011	Smartphone	83	8 m	Class. ($k = 2$)	SVC	LOO	**OCEAN***
de Montjoye et al. [34]	2013	CDR	69	16 m	Class. ($k = 3$)	SVC	10-fold	**OCEAN***
Mønsted et al. [83]	2018	CDR	636	24 m	Class. ($k = 3$)	SVC	10-fold	OC**E**AN
Stachl et al. [104]	2020	Smartphone	624	30 d	Regression	RF	10-fold	**OCE**AN
→ This article	2022	Fitness Tracker	204	4 m	Class. ($k = 3$)	SVC	LOO	**OCEAN**

* Mønsted et al. [83] showed that these articles suffer from test-data leakage (i.e., when data from the test data is used for training, for instance, in the feature selection step), which leads to overfitting. Therefore, the performance reported in those articles is largely overestimated. For example, according to Mønsted et al. [83], if de Montjoye et al. had used a rigorous experimental setup, they would have only obtained statistically significant results for extraversion (leading to OC**E**AN instead of **OCEAN** in the table).

daily activities. They show that it is possible to infer openness, extraversion, and conscientiousness from these data.

In summary, we are the first to demonstrate that WAT data brings valuable information to classify users according to their personality traits. Moreover, regarding related work that used similar methodological approaches (ternary classification), we show WAT data is more helpful for such classification than phone data. Also, by using a rigorous evaluation methodology, and thus, in comparison with most of the previous works, fairly evaluating the inference performance, we are the first to show how users can be classified according to their neuroticism level with an accuracy significantly higher than the baseline. Finally, we show that WAT data are correlated to openness, which was not the case with the data considered in prior work (e.g., CDR).

5 Data Collection and Statistics

We describe our data collection campaign and we report on the general statistics regarding our participant pool.

5.1 Data-Collection Campaign

Evaluating the privacy of WAT users, with respect to their personality, requires having access to both WAT and personality data for a number of individuals. In order to collect such data, we organized a large-scale experiment. We recruited the participants through LABEX, a dedicated structure of the University of Lausanne (UNIL); it manages a pool of around 8'000 students from two universities (a technical one, i.e., EPFL, and a general one, i.e., UNIL itself, that covers a broad range of disciplines). Those who were interested in our experiment responded to a screener questionnaire that we used to verify their eligibility for participating. 981 individuals answered the screener questionnaire and 429 were compatible with the experiment criteria: to own a smartphone compatible with the Fitbit application, to speak French(i.e., the local language at the universities; the questionnaires were in the local language), and to not already own a WAT. We finally recruited 230 individuals.[2]

In order to ensure a better diversity of personality profiles, we selected the participants from different academic institutions and various study disciplines. Every selected participant received a Fitbit Inspire HR bracelet. We chose to use a Fitbit device because Fitbit is one of the leaders in the WAT market [70] and because it provides a well-documented and effective API [4] to collect users' data. Moreover, the Fitbit Inspire HR is a high-end general-purpose WAT; as such, it gave us access to a wide range of data types (including step count, activities, sleep time, and heart rate) while still being used by a large user base. Using Fitbit trackers introduced some minor limitations such as the limited accuracy of some of their sensors (compared to higher-end devices) [106] as well as limited access to the data that they collect (only processed data, unlike specialized devices).

We only recruited new users because we wanted to provide them all the same WAT model, for data homogeneity and data collection infrastructure (Fitbit API). Furthermore, recruiting individuals who already owned a WAT could have increased the dropout rate as they would have been tempted to switch back to their own devices during the data collection.

The participants were instructed to wear the bracelet daily and all day long (they were free to remove it for comfort reasons, for example, at night) and to regularly synchronize with the Fitbit app running on their smartphones. They also had to answer a questionnaire that consisted of demographic ques-

[2]Part of the participants agreed to share their WAT data. The dataset is available at https://dx.doi.org/10.5281/zenodo.7621224

tions and the NEO-PI-3 standardized personality assessment items [77] (see Section 2.1), which were used to compute their Big Five scores.[3] We chose that specific questionnaire because it is a reference questionnaire and because it provides results with high confidence and fine granularity. The purpose of this questionnaire was to collect the necessary ground truth.

The WAT data were collected for four months (between May and September 2020)[4] using the Fitbit API (the participants had to grant us an access authorization by using the OAuth2 protocol).[5] We collected the step count for every one-minute interval; the average heart rate for every one-minute interval; the sleep related data such as the bedtime, wake-up time, sleep quality or the number of times that a user was restless during their sleep (for those who wore the device at night); as well as the sports activities (e.g., running, biking) that were automatically detected by Fitbit. Finally, in order to ensure high data-utility of our dataset, we decided to only keep the 204 individuals who wore their devices at least 50% of the time.

Ethical Considerations

During the distribution of the devices, the participants had to sign a consent form that described the conditions of participation, the data being collected (and the associated data management plan), the procedure to withdraw from the study, and information about the financial incentive. The institutional review board at our university validated the consent form and approved the entire experiment. As a reward, participants were paid 60 CHF ($\sim$ 60 USD) at the end of the data collection campaign, and they were allowed keep their device for personal use, which they all did.

5.2 Descriptive Statistics

Among the 204 selected participants, 64.7% were women, 34.8% were men, and 0.5% (1 participant) preferred not to indicate their gender. The women/men ratio is representative of the Fitbit user base. Indeed, we can observe that, in the general population, two thirds of Fitbit users are women [3]. 72% of our participants are students from the general university (where a majority of students are women), and 28% are from the technical university. They are on average 22.6 years old with a standard deviation of 2.7 years. The youngest is 18 years old and the oldest is 33 years old. Note that even if the age range is not representative of the general population, as the Big Five model is known to be stable over time [28], this should not substancially influence our results. Regarding the national statistics in our country, the age distribution corresponds to the student population. However, the proportion of women is slightly higher in our dataset than in the global student population. The scores for all personality traits correspond to a normal distribution. The medians of the scores for the five different personality traits have values between 96.5 and 125 points, depending on the trait. With terciles of 84 and 109 points, neuroticism has the highest score variability, which helps us to better infer that personality trait (this is confirmed by our results; see Section 7), as the difference between individuals appears to be substantial. Table 2 shows the distribution of participants across each tercile of each personality trait. We can observe that the distribution is globally well balanced with no majority class containing more than 35% of the samples. Because participants can have the exact same scores, the terciles are not always of size exactly 33%. The participants wore their devices during 88% of the data collection period on average. They have an average heart rate of 75 bpm (beats per minutes) with a standard deviation of 7 bpm. During the data collection period, the participants took 8,669 steps per day on average with a standard deviation of 2,740.They slept for 8 hours and 17 minutes per day on average with a standard deviation of 2 hours and 4 minutes. Physical activities are automatically detected and recorded by the device, however, it only takes into account activities lasting 15 minutes or more. Walking was, by far, the most practiced activity (63% of the activities). As the participants were free to sometimes remove their bracelets, they probably took steps, slept, or did activities that were not taken into account by the device, therefore, the previously discussed statistics about Fitbit collected data could be slightly underestimated.

5.3 Participants' Privacy Concerns

In the exit questionnaire, we asked the participants to evaluate on a 5-point Likert scale: (1) *To what extent (that is, with what precision) can personality be inferred based on the data collected from your Fitbit tracker? (from "Not at all precise" to "Extremely precise")* and (2) *To what extent would you be worried if the user's personality could be inferred accurately based on the data collected by your Fitbit tracker? (from "Not at all worried" to "Extremely worried")*. For the first question, 47% of the participants answered "Not at all precise" or "Slightly precise", 34% answered "Moderately precise" and 19% answered "Very precise" or "extremely precise". For the second question, 38% of the participants answered "Not at all worried" or "Slightly worried", 26% answered "Moderately worried" and 36% answered "Very worried" or "extremely worried". Our participants also ranked personality as one of the most concerning types of information in a proposed list (age, alcohol, and tobacco consumption, illegal drugs consumption, menstrual cycles, political views, religion, sexual activity, sexual orientation, socio-economic status), and they were far more concerned with personality

[3] The questionnaire is available on https://www.parinc.com/Products/Pkey/275, unfortunately, we cannot directly share it due to copyright issues.

[4] The data collection campaign was conducted during the COVID-19 pandemic. However, there was no lockdown or restriction from May to September in Switzerland; only large events were canceled.

[5] Our access was revoked shortly after the end of the experiment.

Table 2: Distribution of the number of samples for each tercile and each personality trait.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Low	71 (35%)	68 (33%)	72 (35%)	69 (34%)	69 (34%)
Medium	70 (34%)	71 (35%)	64 (31.5%)	69 (34%)	67 (33%)
High	63 (31%)	65 (32%)	68 (33.5%)	66 (32%)	68 (33%)

being inferred than religion or sexual orientation.

6 Inference

Privacy is commonly characterized as the (in)accuracy of an inference process [109], conducted by an adversary, that takes user data as input (data collected from WATs in our case) and outputs (a probability distribution across possible) values for some private attributes of the users (scores for the OCEAN personality traits in our case). In order for the privacy quantification to be fair and unbiased, it is paramount to properly design the inference framework and methodology, as shown by Mønsted et al. [83].

In this section, we describe the machine-learning-based inference methodology, the data extracted from the WATs for the inference (i.e., the features), and we report on our empirical results regarding the quantification of the privacy of WAT users, with respect to their personality.

6.1 Methodology

We define an inference framework which consists in training and testing a machine-learning (ML) model for predicting the scores for each of the OCEAN personality traits, for a given user and the WAT data associated to them. Based on the participants' "actual" scores, computed from their responses to the NEO-PI-3 questionnaire [77] by following a standardized methodology, we establish the ground truth for the personality traits. We use this ground truth to train the ML model, in a supervised manner, and to evaluate its performance in terms of accuracy.

Inference Method

We chose to rely on classification methods because (1) the category within a general population to which an individual belongs to is the most important aspect from a psychological point of view [77] (as explained in Section 2.1) and (2) it is the most common method used in prior work [23, 34, 83]. Classes can be defined based on quantiles in order to get evenly sized groups (in terms of their number of individuals). For example, in the case of two classes (i.e., binary classification), the first class is defined as the individuals whose score is lower than the median and the second class as those whose score is higher than the median. In the case of three classes (i.e., ternary classification), the class boundaries correspond to terciles. A common problem of using the aforementioned technique with an even number of classes is that, for bell-shaped distributions of scores, it splits participants in classes in the middle of the bell, where most of the participants lie. To minimize this issue, we defined the inference attack as a ternary classification process, similarly to previous works [34, 83]. Therefore, the classification problem consists in inferring, for each individual and each personality trait, if they belong to the bottom, middle or top personality score class (regarding the score terciles), with respect to the whole dataset.

Evaluation

We evaluated privacy for each of the five main personality traits (OCEAN) independently. For each trait, we defined three classes from the whole dataset as explained above, and we conducted the inference and the evaluation. In order to train and evaluate the model, we proceeded to a nested Leave-One-Out (LOO) cross-validation. More specifically, for a dataset $S = \{x_i \,|\, i \in [1..N]\}$, where x_i denotes the data of participant i, the model was trained and evaluated N times using $S \setminus \{x_i\}$ as training set and $\{x_i\}$ as testing set for each $i \in [1..N]$. Moreover, for each of the N iterations, the feature selection strategy and its hyper-parameters (i.e., number of selected features) as well as the hyper-parameters of the model were chosen using a grid search with LOO cross-validation on the $N-1$ elements of the training set.

By proceeding this way, we make sure that the results presented are fair in the sense that information leakage (e.g., when the feature selection is done on the entire dataset) is prevented. As pointed out by Mønsted et al. [83], sharing data between model selection and model evaluation steps leads to overestimating performance of the models at stake. In particular, they show that some of the works related to ours [23, 34] are subject to such methodological biases. We use the accuracy (i.e., the proportion of correctly classified instances) as our evaluation metric. This metric is the most suitable for comparing different models, and it provides a clear understanding of their performance. Moreover, it is the only metric that is used in all prior work performing classification [23, 34, 83]. However, we are aware that accuracy is limited since, as it aggregates the confusion matrix into a single value, it does not distinguish between different types of errors and their associated magnitudes (e.g., misclassifying a participant as "bottom" instead of "top" is worse than misclassifying them as "middle"). Finally, we compare our results to the baseline defined by a uniformly-random naive

classifier (the probability of inferring the correct class for each trait and each test individual is therefore 33%). Due to slight differences between the class sizes, we decided not to use majority baseline. When the difference between two class sizes is zero or one, holding-out a single sample from the training set would result in the corresponding class being under-represented in the training set and the majority-class classifier would then underperform the random baseline.

The inner loop of this nested cross-validation performs both feature and model selection. The feature selection strategy is cross-validated among (i) univariate feature selection, (ii) a greedy feature elimination strategy, and (iii) a model-based feature importance approach. The models at stake in this inner cross-validation loop are Support Vector Machines (SVM) and Random Forests (RF). Cross-validated hyper-parameters for SVMs are the kernel (Gaussian and linear kernels are considered), C and γ (for Gaussian kernels), while for RFs, we have cross-validated the number of trees in the forest and the split criterion. For all traits, in all iterations of the inner loop, the selected model is an SVM. Note that, as it can be observed in Table 1, SVM is the most common ML method used in prior work for solving similar problems. For the implementation, we have relied on the `scikit-learn` [100] machine learning library for Python.

6.2 Feature Extraction

We collected different types of data through the Fitbit API: time series (steps, heart rate, battery level), events (sleep, activities) and standalone features (gender, resting heart rate). The extraction of most of our features consisted of aggregating time-series data over time intervals, with some periodicity using the following method: for each day of the week, we aggregated data according to predefined periods of the day. To this end, we partitioned the day into six periods of four hours with boundaries at: 2AM, 6AM, 10AM, 2PM, 6PM and 10PM. Previous studies highlighted that personality is correlated with individuals' circadian rhythm (natural process that regulates a 24-hour biological cycle) [36, 65]. We thus defined $6 \times 7 = 42$ different periods (e.g., "Monday between 10AM and 2 PM") for aggregating the data into features. We then computed features corresponding to their two first statistical moments (i.e., the mean and the variance for the heart rate and step count taken across each of these periods).

Note that, although the extracted features refer to physiological and behavioral information, they are not as rich as those that can be collected from a (smart)phone [24, 34, 35, 83, 105]. They could also contain errors as, for example, the sensor signal analysis might sometimes not detect the right activity or confuse a step with certain arm gestures.

Furthermore, they are particularly centered on the user's activities and, unlike phone data, contain no direct social information, even though multiple personality traits have a strong social component.

Steps and Heart Rate

Steps and heart rate have the same data structure: they are sequences of pairs (t,x), where t a timestamp, and x a scalar value. The sampling period is of one minute. We extracted features from the data of both types by using the periodic aggregation method explained above. As Fitbit "rewards", on a daily basis, its users whose step counts exceed a certain so-called "daily step goal" (set to $10,000$ by default), we added the following three related features: the number of times this goal is achieved, the number of times it is just achieved (up to 5% *more* than the step goal), and the number of times it is almost achieved (up to 5% *less* than the step goal). Furthermore, the Fitbit API directly provides the resting heart rate for each user, which we used as such as a feature. As mentioned in Section 2.1, a relatively high score in extraversion is, for example, linked to sociable and active individuals whose traits could influence the step count. One of the extraversion sub-traits is excitement seeking, which can lead to an augmentation of an individual's heart rate. Neuroticism is linked to impulsivity and stress, which can also cause variations in heart-rate. Moreover, it has been shown that heart-rate variability and an individual's personality are correlated [122].

Sleep and Activities

Sleep data are composed of a start time, a duration, and other information such as the sleep quality, the number of times the user wakes up during their sleep, and the number of times they are agitated. We built features of the same structure as steps and heart rate. We generated, the mean and standard deviation of sleep time, for each four-hour and day-of-the-week periods. We also computed the mean and standard deviation of the awake duration during sleep, the awaking count, the sleep duration, the time (in minutes) it takes to fall asleep, the restless-moment count and duration, and the sleep efficiency (all these details are directly provided by Fitbit). The data structure of the activities is similar to that of sleep data. We therefore built similar features. We computed the number and proportion of each practiced activity, as well as the entropy of the distribution of practiced activities. As mentioned previously, active individuals tend to obtain higher scores in extraversion. As for sleep, previous studies established that an individual's sleep quality and habits are correlated with their personality [55, 93, 99].

Battery

The "current" battery level of the device is available at any point in time through the profile endpoint of the Fitbit API. To eventually obtain a battery data time series for each participant, we collected this twice a day, at a fixed time. Note that the API returns the battery level at the time of the last synchronization (together with the time of the last synchronization). Then, we extracted the average battery level right before and

Table 3: List of all features used in the evaluation. "Std." stands for standard deviation. The "+" operation for data aggregation means that both aggregating methods were used to obtain the given feature. The dots in the last 5 columns indicate that the corresponding features of this data type were selected by the model for inferring the corresponding trait in our evaluation.

Data Type	Statistics	Aggregation Method	O	C	E	A	N
Step count	Mean, Std.	Days of the week + 4-hour period	•	•	•	•	•
Step goals	Nb. of occurrences	The whole data collection period	•				•
Heart rate	Mean, Std.	Days of the week + 4-hour period	•	•	•	•	•
Sleep time	Mean, Std.	Days of the week + 4-hour period	•	•		•	•
Other sleep details	Mean, Std.	No aggregation				•	•
Activity time	Mean, Std.	Days of the week + 4-hour period	•	•		•	
Activity types	Entropy, Nb., Proportion	Activity type	•	•	•	•	•
Battery charging	Entropy, Nb. of occurrences	Days of the week, 4-hour period		•			
Gender	Category	N/A	•			•	•

after a charge, as well as its standard deviation. We also computed how many times each participant charged their device and the entropy of the time elapsed between these events, for each day of the week. We also created similar features by using only the six previously defined periods of the day (without again aggregating with the days of the week). However, the Fitbit API provides only the battery level at the time of the last synchronization between the bracelet and the smartphone. Therefore, we might have lost information if users had not synchronized their data regularly (e.g., if Bluetooth was not continuously activated on their phone).

Gender

As gender is known to be correlated with the score of some personality traits [112], and as such information is often available through the profile endpoint of the Fitbit API, we included gender data as a feature. All the participants specified a gender in their profiles. We observed a mismatch between the gender they specified in their Fitbit profiles and that specified in their responses to our questionnaire for only 0.98% ($n = 2$) of the participants. Self-reported gender data can therefore be considered as a readily-available (to an adversary) and trustworthy data in the inference process.

7 Results

Inference Accuracy

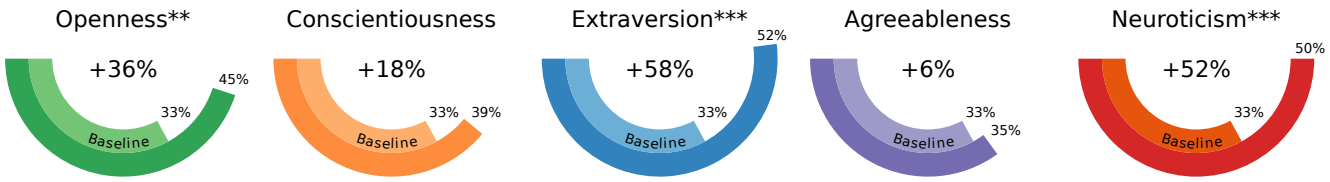
As shown in Figure 1, we obtained results that are statistically significantly better than the baseline[6] for openness ($p < 0.01$), extraversion, and neuroticism ($p < 0.001$). The trained model correctly classified 45% of the participants' scores in openness (+36% with respect to the baseline), 52% of the participants' scores in extraversion (+58% with respect to the baseline), and 50% of the participants' scores in neuroticism (+52% with respect to the baseline). We further observe that Fitbit data brings some valuable information for the inference of other traits, such as agreeableness and conscientiousness, but these results are not statistically significant. Regarding the definition of each personality trait, it is relatively intuitive that WAT data are less informative for a trait such as agreeableness than for neuroticism or extraversion. Table 5 in the appendix provides more performance metrics, namely precision, recall and f1-scores for each tercile. For openness, extraversion, and neuroticism, the weighted mean of the f1-score (respectively 0.45, 0.51, and 0.50) is clearly higher than the baseline (0.33), which confirms the results presented above.

Influential Features

In Table 3, we can see which general-data types were used to extract the relevant features for inferring each personality trait. For each inference, we looked at the three most informative features. We considered the features selected more times during the inner loop of our cross validation as more informative. For features used to infer openness, extraversion, and neuroticism, we conducted statistical tests (Kruskal-Wallis with Bonferroni correction) to reject the natural null hypothesis that the differences between terciles are incidental to the collected data. We show that we can reject the null hypothesis for all of these features with $p < 0.05(*), p < 0.01(**), p < 0.001(***)$ or $p < 0.0001(****)$. More details and figures are available in Appendix A. The three most informative features for each inference process are (when there are more than three features, all the presented features are considered as equally important by the model):

- **Openness****
  - Step-goals ($\geq$ 10k steps) just achieved.**
  - Number yoga activities.*
  - HR std from 2AM to 6AM on Thu.**
  - HR std from 10AM to 2PM on Fri.**
  - HR std from 2PM to 6PM on Thu.*
- **Conscientiousness**

[6] All statistical tests for model comparison were conducted using McNemar's test, with Bonferroni correction.



$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 1: Accuracy of the ternary classification with respect to the baselines for each of the five main traits. For each trait, we display the increase of accuracy (in percentage) compared to the random baseline, the accuracy of the baselines and the accuracy of the prediction. Percentages are rounded to the unit. The accuracy of the prediction outperforms both baselines with statistical significance with Bonferroni correction (i.e., using an α value of $0.05/m$ with m the number of inferences, 5 in our case) for openness ($p < 0.01$), extraversion, and neuroticism ($p < 0.001$).

  - Std of HR btw Wed. and Thu. (10PM-2AM)
  - Sleep-time mean from 10AM to 2PM on Sun.
  - Sleep-time mean from 2AM to 6AM on Sat.
- **Extraversion*****
  - Step mean btw Fri. and Sat. (10PM-2AM).****
  - Step mean on Mon. btw 6PM and 10PM.****
  - Step mean btw Thu. and Fri. (10PM-2AM).****
  - Number of distinct activities.***
  - HR mean btw Sun. and Mon. (10PM-2AM).****
- **Agreeableness**
  - Steps std on Sun. btw 6PM and 10PM.
  - Sleep-time mean (global).
  - Std of HR on Thu. btw 10AM and 2PM.
- **Neuroticism*****
  - Gender.****
  - Steps mean on Mon. btw 6PM and 10PM.**
  - Sleep-time mean from 10AM to 2PM on Sun.*

Interestingly, we can see that the practice of yoga is highly informative for the inference of openness. This is coherent as users with high openness tend to seek new experiences and to engage in self-examination and individuals who practice yoga are known to obtain higher score in openness [19]. However, we cannot make a general conclusion here with that information as only eight participants recorded yoga activities during the data collection. Among those participants, only one was not classified in the high openness tercile. HR-related features are important for the inference of openness. Psychology studies have shown that features related to cardiac activity (including heart rate), are correlated with openness [30, 90]. This is confirmed by Table 4 which shows that without HR-related features, our model is not able to correctly classify individuals according to their openness level significantly higher than the baseline. Note that most of these HR-related features are relative to Thursday and Friday afternoons. One possible reason is that openness is related to art sensitivity and creativity and that these time slots are the most favorable for such activities (museums or art galleries, for example, are often closed at the beginning of the week). Thursday and Friday evenings/nights or Saturday, however, are time periods related to extravert-oriented activities (e.g, clubbing). We can also observe that steps goals are used to infer the score of openness, however, there is no previous research that can help us understand the reason of this correlation.

Looking at Table 3, we can first observe that, information related to steps, heart rate, and activities are used to infer extraversion. This can be explained by the fact that people with higher scores in extraversion tend to be more active, assertive, and sociable (see Section 2.1). The three most informative features relate to the average step count *at night*, thus showing that the level of (social) activity plays a key role in the inference of extraversion. Indeed, the more extraverted a participant is, the more steps they take at night (especially at night between Thursday and Friday, on Monday evenings, and at night between Friday and Saturday). This could be explained by the fact that the more extraverted the individual, the more they go out at night (e.g., to meet friends, to go clubbing, etc.). That may also be supported by the mean heart-rate on Sunday night being higher for the most extraverted individuals. Furthermore, we observe that the most extraverted individuals tend to do more distinct activities, which corresponds to the activity and excitement seeking component of extraversion as described in Section 2.1. Moreover, to assess personality traits, standard tests combine behavioral, cognitive, and affective indicators [120], and behavioral indicators are the most informative to assess extraversion [59]. This explains why WAT data, which are almost exclusively related to behavior, are the most informative for this trait.

Steps, heart rate, and activities are also used to infer neuroticism. However, we observe that HR-related features do not appear to be the most informative features for neuroticism. Instead, these features relate to gender, sleep, and steps. Previous works show that information such as step count, heart

rate, or duration of sleep are indicators of stress resilience, which by definition is highly correlated with neuroticism [8]. We also observe that sleep and gender are used to infer neuroticism. Both are indeed known to be correlated with this personality trait [55, 112]. More specifically, there is a significant difference among the terciles regarding the sleep time (here on Sunday between 10am and 2pm). It also shows that there is a significant difference between genders regarding their neuroticism score. Figures showing how the most informative features are distributed over the different terciles are available in Appendix A. As gender is correlated with neuroticism, we trained and evaluated a simple decision tree to infer the neuroticism class from gender only with the same methodology as described before. Such a model reaches an accuracy score of 48%. Additionally, we also evaluated our model without using gender and showed that it reaches 47% of accuracy. Therefore, a model using WAT data is similar, in terms of accuracy, to a model based on gender for inferring neuroticism. However, considering that WAT users can easily lie about their gender without decreasing their utility, which is not the case with step count or sleep data, a model based on WAT data (possibly helped by gender), is therefore more reliable than a model based on gender only.

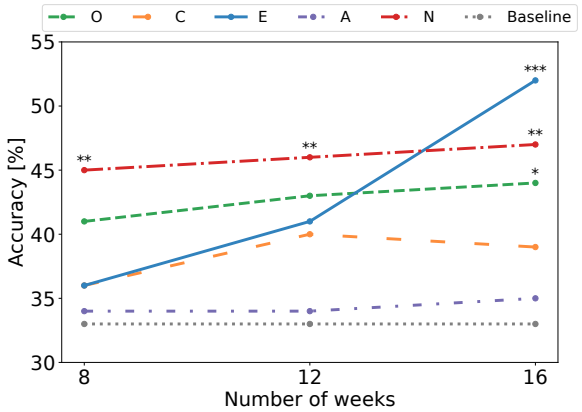
Note that the list of informative features for the conscientiousness and agreeableness traits should be considered with caution, because it corresponds to prediction tasks for which our models do not significantly outperform the baseline.

Sensitivity Analysis

We evaluated the inference performance by using a subset of data sources. Indeed, when giving access to the API, WAT users can choose to restrict access to some information by selecting only some types of data or, simply, by choosing to not report personal information (i.e., gender). Some devices can simply not collect certain data due to the lack of sensors (e.g., unlike the Fitbit Inspire HR, the Fitbit Inspire does not collect heart-rate data). Table 4 summarizes the results obtained by evaluating the inference model which uses different data source combinations. The accuracies of the extraversion and neuroticism inferences are still significantly higher than the baseline when using only step-count related features. This demonstrates that even devices that do not collect the heart rate, such as the Fitbit Inspire bracelet, can be used to accurately infer the personality of their users. However, the results from Table 4 suggest that heart rate data is essential to infer openness as the inference accuracy significantly declines when we remove this data source from the features set.

Performance Evolution over Time and Training Set Size

Additionally, we analyzed how the inference performance evolves with training data collected over an increasing period of time. As it does not evolve over time, we did not use gender as a feature. Figure 2 shows, for each trait, how the inference accuracy evolves using training data collected for 8, 12, and 16 weeks.



$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 2: Evolution of the performance of the inference with training data collected for the first 8, 12, and 16 weeks. As it does not evolve over time, gender is not used as a feature.

We can observe that only 8 weeks are necessary to obtain an accuracy significantly better than random for neuroticism while 16 weeks are required to significantly outperform the baseline for openness and extraversion. We can also postulate that the inference performance would be better with a few more months of data (which would capture additional seasonal phenomena), especially for extraversion, that shows the highest growth with time. We observe that the inference of extraversion is highly dependent on the data collection duration. This is probably due to seasonal behavior change (e.g., people tend to go out more often during the summer), and due to the fact that the most important features are probably related to social events, and thus that more time is necessary to collect enough data related to these specific, and possibly short, events. However, the results tend to show that an augmentation of data collection duration would not highly impact the inference of conscientiousness and agreeableness. Note that we use the same set of participants for all inferences, which may introduce a bias due to the fact that we selected the ones who wore their devices at least 50% of the time during the whole four-month period. Results with fewer months could so be slightly underestimated considering that some participants may have been selected while they were not wearing the device much during that specific period.

Finally, we also evaluated our model using k-fold cross validation with $k \in \{2,3,4,5,10\}$ (details are available in Figure 6 in the appendix) and show that, especially for openness, neuroticism, and extraversion, the inference accuracy tends to increase with the size of the training set. For all traits, the accuracy does not plateau for larger training sets, which indicates that the accuracy would increase if the sample included

Table 4: The obtained inference accuracy using different combinations of data sources. The increase in accuracy is computed using the random baseline. The last line corresponds to aggregations by day (i.e., not 4-hours time slots) for heart rate and steps.

Data source	O	C	E	A	N
All data sources	45% (+36%)**	39% (+18%)	52% (+58%)***	35% (+6%)	50% (+52%)***
All data but gender	44% (+33%)*	39% (+18%)	52% (+58%)***	35% (+6%)	47% (+42%)**
All data but heart rate	35% (+6%)	32% (-3%)	50% (+52%)***	34% (+3%)	50% (+52%)***
All data but heart rate and sleep	34% (+3%)	33% (+0%)	50% (+52%)***	33% (+0%)	48% (+45%)**
Only step count	34% (+3%)	32% (-3%)	47% (+42%)**	34% (+3%)	44% (+33%)*
All data (aggregated) but gender	38% (+15%)	35% (+6%)	33% (+0%)	34% (+3%)	34% (+3%)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

data of more individuals.

Obfuscation

Finally, we evaluated the inference performance using heart rate and step count data *aggregated by day* (instead of 4-hour intervals), mimicking the case where the adversary would only have access to the average daily heart rates and total daily step counts (other features such as sleep and activities are used in the same way as described previously). Indeed, previous research suggests that such aggregation may be used as an obfuscation technique to reduce privacy risks and shows high acceptance among WAT users [107]. Table 4 shows that aggregating heart rate and step count results in an important drop in accuracy and that none of the inferences are significantly better than the baseline in this case. Note that we also removed gender from the features to properly evaluate the impact of such an obfuscation technique on neuroticism.

8 Discussion

Our experimental results demonstrate that processed data from WATs bring valuable information about at least three of the Big Five personality traits. Indeed, WAT data correlates with at least three of the five personality traits, which is consistent with multiple previous findings showing that behavior indicators are particularly informative for some traits (especially for extraversion) [59, 120], that WAT data can help assess stress resilience [8], or that it can be used to infer someone's mood [67]. As the results of this work are based on a limited period of time and on *processed* data, they constitute a lower bound of what an adversary, such as the service provider, could achieve in terms of inference. As we used only WAT data collected from a limited number of individuals during a limited amount of time, our results constitute a lower bound of what data brokers can do. On the one hand, they can access training data from many more individuals, and thus can build stronger models. On the other hand, they can easily link WAT data with other types of data to improve the inference models. In their research, Aimeur et al. [9] showed how easy it is to link data of the same individual through different data broker databases. They voiced concerns about how easy it is to collect personal data about given individuals in general. Furthermore, it is known that few individuals read privacy policies, and that among those who do, one-third claim to have no (or very little) understanding of what they read [14]. Considering this, and that most WAT users tend to forget about the (not always honest [42, 76, 89]) third-party apps they share their data with and highly underestimate their number [123], it is likely that many data brokers have access to individuals' WAT data along with other types of personal data that can be used together to accurately infer personality profiles. Moreover, as Google recently acquired Fitbit [91] and plans to force Fitbit users to migrate their Fitbit account into their Google accounts [111], they will be in position to build the strongest possible inference models. Furthermore, the magnitude of this threat can only increase as the technology improves with the addition of new sensors (e.g., ECG), better sensor accuracy, and more efficient machine-learning algorithms. This raises obvious privacy and societal issues, especially in the light of the recent scandals related to personality-based influence campaigns.

To address this threat, a first step is to raise awareness of it. This article makes a contribution by providing concrete evidence of this threat based on a rigorous risk assessment. Based on this assessment, privacy protection techniques should be designed. A first protection technique would be to limit the amount of data shared with the service provider, keeping as much data as possible on the users' devices. As all Fitbit users collected data are stored on Fitbit's servers, a simple solution would be to let the user choose whether to store each type of data on Fitbit's servers or to only store them on a personal synchronized smartphone/tablet. Except for some specific data, the raw sensor signal-processing is directly computed either on the WAT or on the smartphone. This means that as long as the user does not need to share personal data and the smartphone's storage capacity is sufficient, they could increase their privacy while keeping the same level of utility. Furthermore, if a given piece of information needs more computing power than provided by the user's smartphone, so it has to be processed on Fitbit's servers, it can simply be deleted from the servers once transferred back to the user.

This will leave the data inaccessible to most of the potential adversaries and reduce the data-leakage risks. Additionally, the data shared could be obfuscated to further enhance users' privacy. A commonly used solution is to add noise to the data, which should be done in a controlled way in order to provide formal guarantees, such as differential privacy. However, we decided to evaluate a different, simpler (and so more understandable by users), technique which consists in aggregating data over some period of time. For instance, only the daily step count or the daily average heart rate could be shared with the service provider. We showed the efficacy of such an obfuscation technique in Section 7. By doing so, an adversary loses substantial information about when the data has been collected, which is particularly useful as seen in Section 7 (e.g., steps at night). Indeed, our results suggest that only intra-day data brings information about personality traits. Therefore, an adversary whose goal is to infer individuals' personality would probably not obtain significant results using aggregated WAT data. Furthermore, in the case of the adversary being the service provider, it would still be able to store their users' (aggregated) data, and to provide them with attractive services. Indeed, recent works show that most users view this obfuscation technique as having little impact on their utility [107], and are inclined to use it when sharing their data [123]. Another possible solution would be to empower users by letting them choose which sensors to enable or disable and which data to keep on the device or share with the servers of the service provider.[7]

An important lead for future work is to evaluate the acceptability of such protection techniques by end users. Would users be interested in disabling some of their WAT sensors (and which ones)? Do users need to synchronize their data with the service provider (which data)? Do users need to synchronize their step counts for every minute and with a one-step precision? Indeed, research has shown that users usually do risk-benefit analysis or so-called privacy calculus when using wearable devices [50]. For example, when purchasing healthcare wearable devices, users trade off receiving relevant and personalized health information, the sensitivity of this information, and the existence of legislative data-protection mechanisms [66]. Some individuals are willing to decrease their privacy for an increase in utility, especially when they consider that the device provides them considerable benefits [121], whereas other individuals are willing to accept lower benefits to gain more privacy [17]. The latter users probably prefer to use WATs that implement protection mechanisms, even if the activation of such mechanism decreases their utility. They could then trade off utility and privacy directly when using the device and fine-tune the parameters with respect to their concerns. This could be studied through the lens of privacy calculus [31, 50, 66].

[7]Note that Fitbit already enables their users to deactivate some sensors directly on some of its devices [44]. However, this option is not particularly highlighted on the user interface and is limited to a binary choice.

Our work has some limitations, beyond those related to the use of Fitbit, as mentioned above. In particular, we only show that, for three of the five traits, WAT data can be used to reach significant higher inference accuracies compared to the random baseline. Thus, future studies are needed to optimize the model and show that WAT data can be used to develop highly effective models for personality inference. Also, while we can assume that our ground truth is particularly accurate given the detailed questionnaire we relied upon, we want to highlight that the participants' answer quality could be degraded due to the well-documented respondent fatigue [92], as well as the social desirability bias [87]. There is clearly a trade-off between the details of the psychological profiles and the quality of the collected survey data. Furthermore, the participants' responses about their privacy concerns may have been biased as they were aware of the study's purpose. Additionally, while the study participants are somewhat representative of the local student population, they are not representative of the general adult population. Finally, a larger duration and a larger number of participants would have increased the significance of our results.

9 Conclusion and Future Work

In this article, we showed that WAT data can help classify users according to their personality traits, especially for openness, extraversion, and neuroticism. We demonstrated that the use of WATs can create privacy risks that an adversary can potentially exploit. Our study is based on the WAT data of 204 individuals collected over a period of four months. We conducted ternary classification and used accuracy as the evaluation metric and obtained results significantly higher than the baseline for openness, extraversion, and neuroticism. Also, we showed that, regarding prior work, using WAT data outperforms the use of call detail records (CDR) for inferring individuals regarding their personality traits. Moreover, we analyzed the selected features and highlighted the most informative ones for each personality trait. We also showed that aggregating step count and heart rate by day is an effective obfuscation technique. Finally, we drew links with related studies and compared our results with theirs.

For future work, as noted in Section 7, we consider that it would be interesting to optimize inference models by exploring more feature combinations and by training and evaluating such models on larger datasets. To this end, additional data collection may be useful. For example, knowing that some WATs provide logging functionalities (e.g., meals and food intake), those data may be used to build features to improve the inference model (prior studies state that personality and dietary habits are correlated [114]). Also, profile information or device-usage data, as the number of "Fitbit friends" or the number of times where a user taps on the device's screen, could be helpful to increase the inference accuracy. It would also be interesting to design and evaluate other obfuscation

techniques. Indeed, it might be relevant to develop obfuscation techniques that result in less data loss, and thus, would have an even better acceptability than the one that we evaluated.

In this study, we focus on a particular adversary who has full access to user data. However, it could be interesting to consider adversaries who would have only partial access to the data and study what methods they might use to obtain these data. Furthermore, we focus on only one given type of device. It would be interesting to extend our study to multiple kinds of devices and evaluate, for instance, how the quality/quantity of sensors affects the inference accuracy. Finally, in our study, we used data collected on a very specific population. Conducting a similar experiment on a more diverse population would be useful for studying whether our results can be extended to all categories of the population.

Acknowledgments

This work was partially funded by the Swiss National Science Foundation with Grant #200021_178978 (PrivateLife), and by armasuisse S+T with Grant #CYD-C-2020007. We thank Didier Dupertuis, Dimitri Percia David, Kavous Salehzadeh Niksirat, Rita Abi Akl, and Yamane El Zein for their help in distributing the bracelets. We thank Samuel Lew for his help with data collection. We thank Gaël Bernard and Rémi Coudert for their help in feature engineering as well as Marie Reignier and Patrick Rousseau for their help with machine learning and data processing. We also thank Robin Zufferey for his help in psychology regarding the metrics for personality assessment. Finally, we thank Holly Cogliati and Vincent Vandersluis for proofreading this article.

References

[1] WeWard - The mobile app that motivates you to walk. https://en.weward.fr/.

[2] Fitness trackers chase after the corporate market. http://www.washingtonpost.com/blogs/on-leadership/wp/2014/12/18/fitness-trackers-chase-after-the-corporate-market/, 2014.

[3] Fitbit Counts on Women as Device Buyers, Just Not Board Members. https://www.businessoffashion.com/articles/technology/fitbit-counts-on-women-as-device-buyers-just-not-board-members, 2015.

[4] Fitbit Development: Fitbit Web API Basics. https://dev.fitbit.com/build/reference/web-api/basics/, 2021.

[5] Number of connected wearable devices worldwide from 2016 to 2022. https://www.statista.com/statistics/487291/global-connected-wearable-devices/, 2021.

[6] Wearables Shipments Grew 9.9% in the Third Quarter of 2021 as Watches Start to Displace Wristbands in the Wrist-worn Device Category, Says IDC. https://www.idc.com/getdoc.jsp?containerId=prUS48460121, 2021.

[7] R. Abdel-Salam, R. Mostafa, and M. Hadhood. *Human Activity Recognition Using Wearable Sensors: Review, Challenges, Evaluation Benchmark*. Communications in Computer and Information Science. 2021. doi: 10.1007/978-981-16-0575-8_1.

[8] D. A. Adler, V. W.-S. Tseng, G. Qi, J. Scarpa, S. Sen, and T. Choudhury. Identifying Mobile Sensing Indicators of Stress-Resilience. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021. doi: 10.1145/3463528.

[9] E. Aimeur, G. Brassard, and M. Guo. How data brokers endanger privacy. *Transactions on Data Privacy*, 2022.

[10] M. Alharbi, A. Bauman, L. Neubeck, and R. Gallagher. Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population. *European Journal of Preventive Cardiology*, 2016. doi: 10.1177/2047487316634883.

[11] A. Alqhatani and H. R. Lipford. "There is nothing that I need to keep secret": Sharing Practices and Concerns of Wearable Fitness Data. In *Proc. of the USENIX Symp. on Usable Privacy and Security (SOUPS)*, 2019.

[12] L. Ardüser, P. Bissig, P. Brandes, and R. Wattenhofer. Recognizing text using motion data from a smartwatch. In *IEEE Int. Conf. on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2016. doi: 10.1109/PERCOMW.2016.7457172.

[13] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck. Dos and Don'ts of Machine Learning in Computer Security. In *Proc. of the USENIX on Security Symp.* 2020. doi: https://doi.org/10.48550/arXiv.2010.09470.

[14] B. Auxier, L. Rainie, M. Anderson, A. Perrin, M. Kumar, and E. Turner. Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information, 2019.

[15] L. Barman, A. Dumur, A. Pyrgelis, and J.-P. Hubaux. Every Byte Matters: Traffic Analysis of Bluetooth Wearable Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021. doi: 10.1145/3463512.

[16] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *Proc. of the Int. Conf. on Multimodal Interfaces (ICMI)*, 2011. doi: 10.1145/2070481.2070528.

[17] I. Bilogrevic, K. Huguenin, S. Mihaila, R. Shokri, and J.-P. Hubaux. Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms. In *Symp. of the Network and Distributed System Security (NDSS)*, 2015. doi: 10.14722/ndss.2015.23032.

[18] A. Boutet and S. Gambs. Inspect What Your Location History Reveals About You: Raising user awareness on privacy threats associated with disclosing his location data. In *Proc. of the ACM Int'l Conf. on Information and Knowledge Management (CIKM)*. 2019. doi: 10.1145/3357384.3357837.

[19] S. Bright, E. Gringart, E. Blatchford, and S. Bettinson. A quantitative exploration of the relationships between regular yoga practice, microdosing psychedelics, wellbeing and personality variables. *Australian Journal of Psychology*, 2021. doi: 10.1080/00049530.2021.1882266.

[20] G. Bruce. Google parent Alphabet's health insurance company grew nearly sixfold in '22. https://www.beckershospitalreview.com/disruptors/google-parent-alphabets-health-insurance-company-grew-nearly-sixfold-in-22.html, 2023.

[21] C. Cadwalladr and E. Graham-Harrison. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, 2018.

[22] F. Celli, E. Bruni, and B. Lepri. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. In *Proc. of the ACM Int. Conf. on Multimedia (MM)*, 2014. doi: 10.1145/2647868.2654977.

[23] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones. In *ISWC*, 2011. doi: 10.1109/ISWC.2011.29.

[24] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 2013. doi: 10.1007/s00779-011-0490-1.

[25] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz. Understand-

ing quantified-selfers' practices in collecting and exploring personal data. In *Proc. of the Conf. on Human Factors in Computing Systems (CHI)*, 2014. doi: 10.1145/2556288.2557372.

[26] M. J. Chorley, R. M. Whitaker, and S. M. Allen. Personality and location-based social networks. *Computers in Human Behavior*, 2015. doi: 10.1016/j.chb.2014.12.038.

[27] J. Classen, D. Wegemer, P. Patras, T. Spink, and M. Hollick. Anatomy of a Vulnerable Fitness Tracking System: Dissecting the Fitbit Cloud, App, and Firmware. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018. doi: 10.1145/3191737.

[28] D. A. Cobb-Clark and S. Schurer. The stability of big-five personality traits. *Economics Letters*, 2012. doi: 10.1016/j.econlet.2011.11.015.

[29] P. T. Costa and R. R. McCrae. Four ways five factors are basic. *Personality and Individual Differences*, 1992. doi: 10.1016/0191-8869(92)90236-I.

[30] I. Čukić and T. C. Bates. Openness to experience and aesthetic chills: Links to heart rate sympathetic activity. *Personality and Individual Differences*, 2014. doi: 10.1016/j.paid.2014.02.012.

[31] M. J. Culnan and P. K. Armstrong. Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation. *Organization Science*, 1999. doi: 10.1287/orsc.10.1.104.

[32] D. Curry. Fitbit Revenue and Usage Statistics (2020). https://www.businessofapps.com/data/fitbit-statistics/, 2020.

[33] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra. Uncovering Privacy Leakage in BLE Network Traffic of Wearable Fitness Trackers. In *Proc. of the ACM Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2016. doi: 10.1145/2873587.2873594.

[34] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. Pentland. Predicting Personality Using Novel Mobile Phone-Based Metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction (SBP)*, 2013. doi: 10.1007/978-3-642-37210-0_6.

[35] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. Armenta Lopez de Vicuña, and N. Oliver. Towards a psychographic user model from mobile phone usage. In *CHI · Work-in-Progess*, 2011. doi: 10.1145/1979742.1979920.

[36] C. G. DeYoung, L. Hasher, M. Djikic, B. Criger, and J. B. Peterson. Morning people are stable people: Circadian rhythm and the higher-order factors of the Big Five. *Personality and Individual Differences*, 2007. doi: 10.1016/j.paid.2006.11.030.

[37] N. A. Dodoo and C. M. Padovano. Personality-Based Engagement: An Examination of Personality and Message Factors on Consumer Responses to Social Media Advertisements. *Journal of Promotion Management*, 2020. doi: 10.1080/10496491.2020.1719954.

[38] C. Duhigg. What Does Your Credit-Card Company Know About You? *The New York Times*, 2009.

[39] S. Eberz, N. Paoletti, M. Roeschlin, A. Patani, M. Kwiatkowska, and I. Martinovic. Broken Hearted: How To Attack ECG Biometrics. In *Proc. of the Network and Distributed System Security Symp. (NDSS)*, 2017. doi: 10.14722/ndss.2017.23408.

[40] S. Eberz, G. Lovisotto, A. Patane, M. Kwiatkowska, V. Lenders, and I. Martinovic. When Your Fitness Tracker Betrays You: Quantifying the Predictability of Biometric Features Across Contexts. In *S&P*, 2018. doi: 10.1109/SP.2018.00053.

[41] M. Ehrenkranz. The Plan to Use Fitbit Data to Stop Mass Shootings Is One of the Scariest Proposals Yet. https://gizmodo.com/the-plan-to-use-fitbit-data-to-stop-mass-shootings-is-o-1837710691, 2019.

[42] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. *ACM Transactions on Computer Systems*, 2014.

[43] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, and S. Shah. AutoSense: Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proc. of the 9th ACM Conf. on Embedded Networked Sensor Systems (SenSys)*, 2011. doi: 10.1145/2070942.2070970.

[44] Fitbit. Fitbit Inspire HR User Manual, 2019.

[45] S. Gabriele and S. Chiasson. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Proc. of the Conf. on Human Factors in Computing Systems (CHI)*, 2020. doi: 10.1145/3313831.3376651.

[46] E. Gibney. The scant science behind Cambridge Analytica's controversial marketing techniques. *Nature*, 2018.

[47] O. M. Gouda, D. J. Hejji, and M. S. Obaidat. Privacy Assessment of Fitness Tracker Devices. In *Int'l Conf. on Computer, Information and Telecommunication Systems (CITS)*, 2020. doi: 10.1109/CITS49457.2020.9232503.

[48] M. A. Gutierrez, M. L. Fast, A. H. Ngu, and B. J. Gao. Real-Time Prediction of Blood Alcohol Content Using Smartwatch Sensor Data. In X. Zheng, D. D. Zeng, H. Chen, and S. J. Leischow, *Smart Health*. 2016.

[49] J. A. Hall, N. Pennington, and A. Lueders. Impression management and formation on Facebook: A lens model approach. *New Media & Society*, 2014. doi: 10.1177/1461444813495166.

[50] C. Hallam and G. Zanella. Wearable Device Data and Privacy: A study of Perception and Behavior. *World Journal of Management*, 2016. doi: 10.21102/wjm.2016.03.71.06.

[51] W. U. Hassan, S. Hussain, and A. Bates. Analysis of Privacy Protections in Fitness Tracking Social Networks -or- You can run, but can you hide? In *Proc. of the USENIX on Security Symp.*, 2018.

[52] Q. He, C. A. Glas, M. Kosinski, D. J. Stillwell, and B. P. Veldkamp. Predicting self-monitoring skills using textual posts on Facebook. *Computers in Human Behavior*, 2014. doi: 10.1016/j.chb.2013.12.026.

[53] A. Henriksen, M. Haugen Mikalsen, A. Z. Woldaregay, M. Muzny, G. Hartvigsen, L. A. Hopstock, and S. Grimsgaard. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. *Journal of Medical Internet Research*, 2018. doi: 10.2196/jmir.9157.

[54] A. Hern. Fitness tracking app Strava gives away location of secret US army bases. *The Guardian*, 2018.

[55] M. Hintsanen, S. Puttonen, K. Smith, M. Törnroos, M. Jokela, L. Pulkki-Råback, T. Hintsa, P. Merjonen, T. Dwyer, O. T. Raitakari, A. Venn, and L. Keltikangas-Järvinen. Five-factor personality traits and sleep: Evidence from two population-based cohort studies. *Health Psychology*, 2014. doi: 10.1037/hea0000105.

[56] R. P. Hirten, M. Danieletto, L. Tomalin, K. H. Choi, E. Golden, S. Kaur, D. Helmus, A. Biello, A. Charney, R. Miotto, B. S. Glicksberg, I. Nabeel, J. Aberg, D. Reich, D. Charney, L. Keefer, M. Suarez-Farinas, G. N. Nadkarni, and Z. A. Fayad. Physiological Data from a Wearable Device Identifies SARS-CoV-2 Infection and Symptoms and Predicts COVID-19 Diagnosis: Observational Study. *Journal of Medical Internet Research*, 2021. doi: 10.2196/26107.

[57] Z. Ihsan and A. Furnham. The new technologies in personality assessment: A review. *Consulting Psychology Journal: Practice and Research*, 2018. doi: 10.1037/cpb0000106.

[58] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc. Recognition of Personality Traits from Human Spoken Conversations. In *Proc. of the Annual Conf. of the Int. Speech Communication Association (ISCA)*, 2011. doi: 10.21437/Interspeech.2011-467.

[59] J. A. Johnson. Units of Analysis for the Description and Explanation of Personality. In *Handbook of Personality Psychology*. Elsevier, 1997. doi: 10.1016/B978-012134645-4/50004-4.

[60] S. Kim, A. Thakur, and J. Kim. Understanding Users' Perception Towards Automated Personality Detection with Group-specific Behavioral Data. In *Proc. of the CHI Conf. on Human Factors in Computing Systems (CHI)*, 2020. doi: 10.1145/3313831.3376250.

[61] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc. of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1218772110.

[62] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 2014. doi: 10.1007/s10994-013-5415-y.

[63] P. Kumari, L. Mathew, and P. Syal. Increasing trend of wearables and multimodal interface for human activity monitoring: A review. *Biosensors and Bioelectronics*, 2017. doi: 10.1016/j.bios.2016.12.001.

[64] H. Kwon, G. D. Abowd, and T. Plötz. Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In *Ubicomp*, 2018. doi: 10.1145/3267242.3267258.

[65] R. J. Larsen. Individual differences in circadian activity rhythm and personality. *Personality and Individual Differences*, 1985. doi: 10.1016/0191-8869(85)90054-6.

[66] H. Li, J. Wu, Y. Gao, and Y. Shi. Examining individuals' adoption of healthcare wearable devices: An empirical study from privacy calculus perspective. *International Journal of Medical Informatics*, 2016. doi: 10.1016/j.ijmedinf.2015.12.010.

[67] J. Li, Z. He, Y. Cui, C. Wang, C. Chen, C. Yu, M. Zhang, Y. Liu, and S. Ma. Towards Ubiquitous Personalized Music Recommendation with Smart Bracelets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2022. doi: 10.1145/3550333.

[68] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li. Deep Learning Based Inference of Private Information Using Embedded Sensors in Smart Devices. *IEEE Network*, 2018. doi: 10.1109/MNET.2018.1700349.

[69] A. C. E. Lima and L. N. de Castro. A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, 2014. doi: 10.1016/j.neunet.2014.05.020.

[70] R. Llamas, J. Ubrani, and M. Shirer. Xiaomi and Apple Tie for the Top Position as the Wearables Market Swells 17.9% During the First Quarter, According to IDC. https://www.businesswire.com/news/home/20170605005391/en/Xiaomi-and-Apple-Tie-for-the-Top-Position-as-the-Wearables-Market-Swells-17.9-During-the-First-Quarter-According-to-IDC, 2017.

[71] K. Lotfy and M. L. Hale. Assessing Pairing and Data Exchange Mechanism Security in the Wearable Internet of Things. In *IEEE Int. Conf. on Mobile Services (MS)*, 2016. doi: 10.1109/MobServ.2016.15.

[72] A. Maiti, M. Jadliwala, J. He, and I. Bilogrevic. (Smart)Watch Your Taps: Side-channel Keystroke Inference Attacks Using Smartwatches. In *Proc. of the ACM Int. Symp. on Wearable Computers (ISWC)*, 2015. doi: 10.1145/2802083.2808397.

[73] A. Maiti, O. Armbruster, M. Jadliwala, and J. He. Smartwatch-Based Keystroke Inference Attacks and Context-Aware Protection Mechanisms. In *Proc. of the ACM on Asia Conf. on Computer and Communications Security (AsiaCCS)*, 2016. doi: 10.1145/2897845.2897905.

[74] A. Maiti, R. Heard, M. Sabra, and M. Jadliwala. Towards Inferring Mechanical Lock Combinations Using Wrist-Wearables As a Side-Channel. In *Proc. of the ACM Conf. on Security & Privacy in Wireless and Mobile Networks (WiSec)*, 2018. doi: 10.1145/3212480.3212498.

[75] A. Maiti, M. Jadliwala, J. He, and I. Bilogrevic. Side-Channel Inference Attacks on Mobile Keypads Using Smartwatches. *IEEE Transactions on Mobile Computing*, 2018. doi: 10.1109/TMC.2018.2794984.

[76] A. M. Mandalari, D. J. Dubois, R. Kolcun, M. T. Paracha, H. Haddadi, and D. Choffnes. Blocking Without Breaking: Identification and Mitigation of Non-Essential IoT Traffic. *Proceedings on Privacy Enhancing Technologies*, 2021. doi: 10.2478/popets-2021-0075.

[77] R. R. McCrae, P. T. Costa, Jr., and T. A. Martin. The NEO–PI–3: A More Readable Revised NEO Personality Inventory. *Journal of Personality Assessment*, 2005. doi: 10.1207/s15327752jpa8403_05.

[78] Ü. Meteriz, N. Fazıl Yıldıran, J. Kim, and D. Mohaisen. Understanding the Potential Risks of Sharing Elevation Information on Fitness Applications. In *ICDCS*, 2020. doi: 10.1109/ICDCS47774.2020.00063.

[79] A. Minamikawa and H. Yokoyama. Blog tells what kind of personality you have: Egogram estimation from Japanese weblog. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work (CSCW)*, 2011. doi: 10.1145/1958824.1958856.

[80] A. Minamikawa and H. Yokoyama. Personality Estimation Based on Weblog Text Classification. In *Modern Approaches in Applied Intelligence (IEA/AIE)*. 2011. doi: 10.1007/978-3-642-21827-9_10.

[81] J. Mink, A. R. Yuile, U. Pal, A. J. Aviv, and A. Bates. Users Can Deduce Sensitive Locations Protected by Privacy Zones on Fitness Tracking Apps. In *CHI Conference on Human Factors in Computing Systems*, 2022. doi: 10.1145/3491102.3502136.

[82] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *Proc. of the ACM Int'l Conf. on Web Search and Data Mining (WSDM)*. 2010. doi: 10.1145/1718487.1718519.

[83] B. Mønsted, A. Mollgaard, and J. Mathiesen. Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, 2018. doi: 10.1016/j.jrp.2017.12.004.

[84] A. Montanari, C. Mascolo, K. Sailer, and S. Nawaz. Detecting Emerging Activity-Based Working Traits through Wearable Technology. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2017. doi: 10.1145/3130951.

[85] V. S. Murahari and T. Plötz. On attention models for human activity recognition. In *Proc. of the ACM Int. Symp. on Wearable Computers (ISWC)*, 2018. doi: 10.1145/3267242.3267287.

[86] A. Natarajan, A. Parate, E. Gaiser, G. Angarita, R. Malison, B. Marlin, and D. Ganesan. Detecting cocaine use with wearable electrocardiogram sensors. In *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2013. doi: 10.1145/2493432.2493496.

[87] A. J. Nederhof. Methods of coping with social desirability bias: A review. doi: 10.1002/ejsp.2420150303.

[88] K. Niazmand, K. Tonn, Y. Zhao, U. M. Fietzek, F. Schroeteler, K. Ziegler, A. O. Ceballos-Baumann, and T. C. Lueth. Freezing of Gait detection in Parkinson's disease using accelerometer based smart clothes. In *IEEE Biomedical Circuits and Systems Conf. (BioCAS)*, 2011. doi: 10.1109/BioCAS.2011.6107762.

[89] M. Nobakht, Y. Sui, A. Seneviratne, and W. Hu. PGFit: Static permission analysis of health and fitness apps in IoT programming frameworks. *Journal of Network and Computer Applications*, 2020. doi: 10.1016/j.jnca.2019.102509.

[90] P. S. Ó Súilleabháin, S. Howard, and B. M. Hughes. Openness to experience and adapting to change: Cardiovascular stress habituation to change in acute stress exposure. *Psychophysiology*, 2018. doi: 10.1111/psyp.13023.

[91] J. Porter. Google completes purchase of Fitbit. https://www.theverge.com/2021/1/14/22188428/google-fitbit-acquisition-completed-approved, 2021.

[92] S. R. Porter, M. E. Whitcomb, and W. H. Weitzer. Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004. doi: 10.1002/ir.101.

[93] C. Randler. Morningness–eveningness, sleep–wake variables and big five personality factors. *Personality and Individual Differences*, 2008. doi: 10.1016/j.paid.2008.03.007.

[94] R. E. Rhodes and N. E. I. Smith. Personality correlates of physical activity: A review and meta-analysis. *British Journal of Sports Medicine*, 2006. doi: 10.1136/bjsm.2006.028860.

[95] C. A. Rivera. The Big-Five Personality Test and Cambridge Analytica. https://galindes.wordpress.com/2019/05/03/the-big-five-personality-test-and-cambridge-analytica/, 2019.

[96] S. Roccas, L. Sagiv, S. H. Schwartz, and A. Knafo. The Big Five Personality Factors and Personal Values. *Personality and Social Psychology Bulletin*, 2002. doi: 10.1177/0146167202289008.

[97] C. Rowl. With fitness trackers in the workplace, bosses can monitor your every step - and possibly more. https://www.washingtonpost.com/business/economy/with-fitness-trackers-in-the-workplace-bosses-can-monitor-your-every-step–and-possibly-more/2019/02/15/75ee0848-2a45-11e9-b011-d8500644dc98_story.html, 2019.

[98] M. Sabra, A. Maiti, and M. Jadliwala. Keystroke inference using ambient light sensor on wrist-wearables: A feasibility study. In *Proc. of the ACM Workshop on Wearable Systems and Applications (WearSys)*, 2018. doi: 10.1145/3211960.3211973.

[99] A. Sano, A. J. Phillips, A. Z. Yu, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman, and R. W. Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *IEEE Int. Conf. on Wearable and Implantable Body Sensor Networks (BSN)*, 2015. doi: 10.1109/BSN.2015.7299420.

[100] scikit-learn. Scikit-learn: Machine learning in Python — scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/.

[101] A. V. Shelgikar, P. F. Anderson, and M. R. Stephens. Sleep Tracking, Wearable Technology, and Opportunities for Research and Clinical Care. *Chest*, 2016. doi: 10.1016/j.chest.2016.04.016.

[102] S. Shen, H. Wang, and R. Roy Choudhury. I Am a Smartwatch and I Can Track My User's Arm. In *Proc. of the Annual Int. Conf. on Mobile Systems, Applications, and Services (MobiSys)*, 2016. doi: 10.1145/2906388.2906407.

[103] M. Shoaib, O. D. Incel, H. Scholten, and P. Havinga. SmokeSense: Online Activity Recognition Framework on Smartwatches. In K. Murao, R. Ohmura, S. Inoue, and Y. Gotoh, *Mobile Computing, Applications, and Services*. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-319-90740-6_7.

[104] C. Stachl, Q. Au, R. Schoedel, S. D. Gosling, G. M. Harari, D. Buschek, S. T. Völkel, T. Schuwerk, M. Oldemeier, T. Ullmann, H. Hussmann, B. Bischl, and M. Bühner. Predicting personality from patterns of behavior collected with smartphones. *Proc. of the National Academy of Sciences*, 2020. doi: 10.1073/pnas.1920484117.

[105] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, and M. Bühner. Personality Research and Assessment in the Era of Machine Learning. *European Journal of Personality*, 2020. doi: 10.1002/per.2257.

[106] S. Tedesco, M. Sica, A. Ancillao, S. Timmons, J. Barton, and B. O'Flynn. Accuracy of consumer-level and research-grade activity trackers in ambulatory settings in older adults. *PLOS ONE*, 2019. doi: 10.1371/journal.pone.0216891.

[107] L. Velykoivanenko, K. S. Niksirat, N. Zufferey, M. Humbert, K. Huguenin, and M. Cherubini. Are Those Steps Worth Your Privacy?: Fitness-Tracker Users' Perceptions of Privacy and Utility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021. doi: 10.1145/3494960.

[108] A. Vinciarelli and G. Mohammadi. A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 2014. doi: 10.1109/TAFFC.2014.2330816.

[109] I. Wagner and D. Eckhoff. Technical Privacy Metrics: A Systematic Survey. *ACM Computing Surveys*, 2019. doi: 10.1145/3168389.

[110] S. S. Wang and M. A. Stefanone. Showing Off? Human Mobility and the Interplay of Traits, Self-Disclosure, and Facebook Check-Ins. *Social Science Computer Review*, 2013. doi: 10.1177/0894439313481424.

[111] J. Weatherbed. All Fitbit users will require a Google account by 2025. https://www.theverge.com/2022/9/26/23372438/fitbit-changes-update-google-account-new-2025, 2022.

[112] Y. J. Weisberg, C. G. DeYoung, and J. B. Hirsh. Gender Differences in Personality across the Ten Aspects of the Big Five. *Frontiers in Psychology*, 2011. doi: 10.3389/fpsyg.2011.00178.

[113] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber. Smartwatch-based activity recognition: A machine learning approach. In *IEEE-EMBS Int. Conf. on Biomedical and Health Informatics (BHI)*, 2016. doi: 10.1109/BHI.2016.7455925.

[114] S. J. Weston, G. W. Edmonds, and P. L. Hill. Personality traits predict dietary habits in middle-to-older adults. *Psychology, Health & Medicine*, 2020. doi: 10.1080/13548506.2019.1687918.

[115] D. R. Witt, R. A. Kellogg, M. P. Snyder, and J. Dunn. Windows into human health through wearables data analytics. *Current Opinion in Biomedical Engineering*, 2019. doi: 10.1016/j.cobme.2019.01.001.

[116] B. Zarouali, T. Dobber, G. De Pauw, and C. de Vreese. Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research*, 2020. doi: 10.1177/0093650220961965.

[117] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: Towards socially and personality aware visual surveillance. In *Proc. of the ACM Int. Workshop on Multimodal Pervasive Video Analysis (MPVA)*, 2010. doi: 10.1145/1878039.1878048.

[118] Q. Zhang and Z. Liang. Security analysis of bluetooth low energy based smart wristbands. In *Int'l Conf. on Frontiers of Sensors Technologies (ICFST)*, 2017. doi: 10.1109/ICFST.2017.8210548.

[119] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. In *Proc. of the ACM Int'l Conf. on Web Search and Data Mining (WSDM)*. 2015. doi: 10.1145/2684822.2685287.

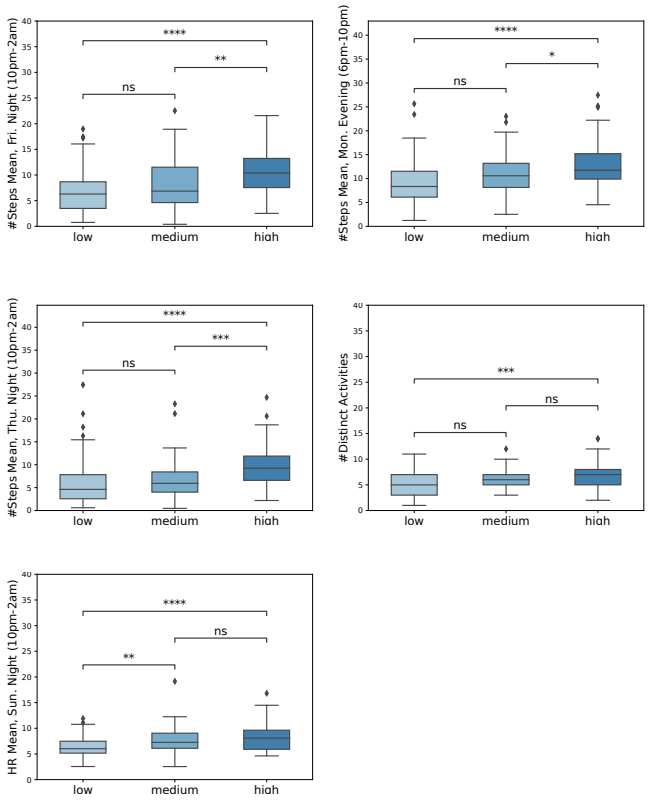
[120] L. M. P. Zillig, S. H. Hemenover, and R. A. Dienstbier. What Do We Assess when We Assess a Big 5 Trait? A Content Analysis of the Affective, Behavioral, and Cognitive Processes Represented in Big 5 Personality Inventories. 2002. doi: 10.1177/0146167202289013.

[121] M. Zimmer, P. Kumar, J. Vitak, Y. Liao, and K. Chamberlain Kritikos. 'There's nothing really they can do with this information': Unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 2020. doi: 10.1080/1369118X.2018.1543442.

[122] A. H. Zohar, C. R. Cloninger, and R. McCraty. Personality and Heart Rate Variability: Exploring Pathways from Personality to Cardiac Coherence and Health. *Open Journal of Social Sciences*, 2013. doi: 10.4236/jss.2013.16007.

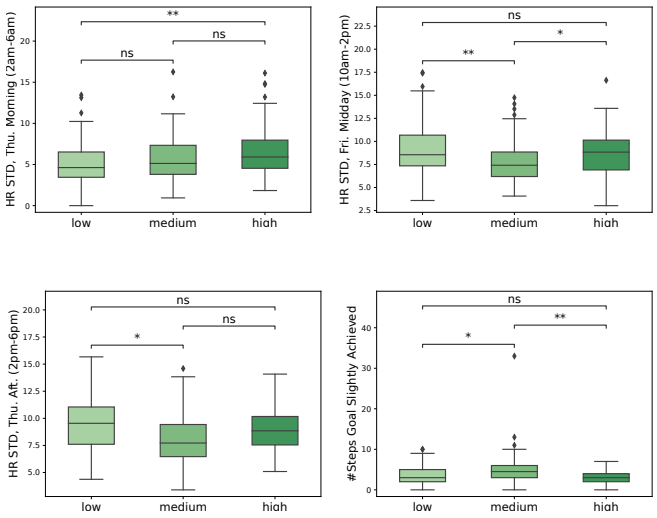
[123] N. Zufferey, K. Salehzadeh Niksirat, M. Humbert, and K. Huguenin. "Revoked just now!" Users' Behaviors Toward Fitness-Data Sharing with Third-Party Applications. *Proceedings on Privacy Enhancing Technologies*, 2023. doi: 10.56553/popets-2023-0004.

A Features Importance



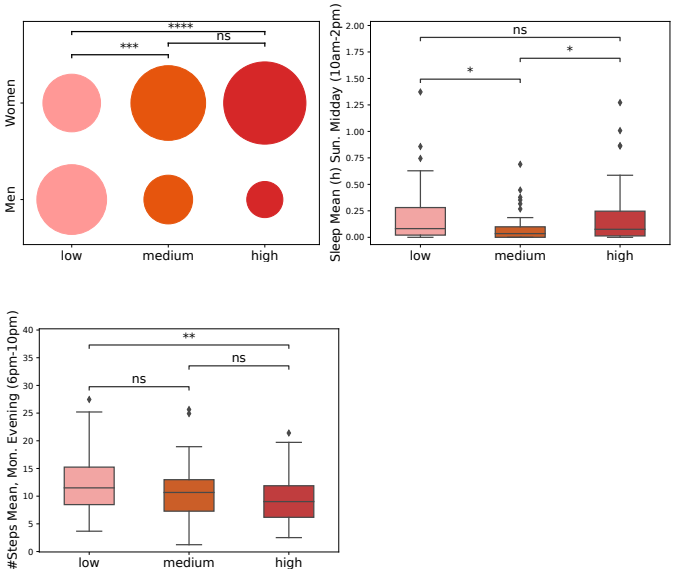
ns: $p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$

Figure 3: Distribution of the five main features used for extraversion inference for each tercile. Step count means are weighted regarding the bracelet wearing time, HR mean is weighted regarding the individual's resting HR.



ns: $p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$

Figure 4: Distribution of four of the main features used for openness inference for each tercile. HR mean is weighted regarding the individual's resting HR.



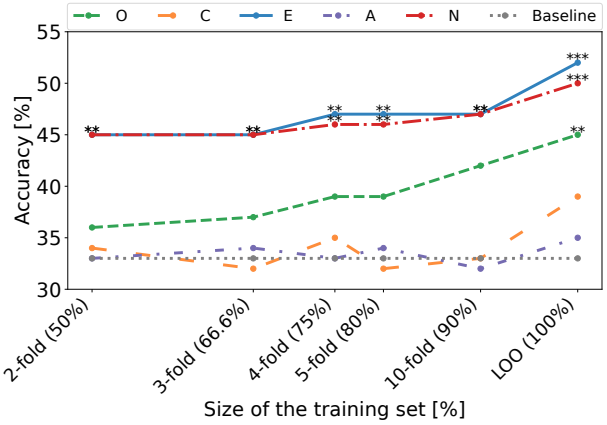
ns: $p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$

Figure 5: Distribution of the three main features used for neuroticism inference for each tercile. Step count means are weighted regarding the bracelet wearing time. The sleep time is in hours. The area of each circle in the gender plot is proportional to the number of participants who corresponds to the given gender.

Table 5: Precision, recall and f1-score for each class.

B Results Details

Openness	**Prec.**	**Rec.**	**f1-score**	**B. f1-score**
Low	0.47	0.39	0.43	0.34
Medium	0.48	0.60	0.53	0.34
High	0.39	0.35	0.37	0.32
Weighted Mean	0.45	0.45	0.45	0.33
Conscien.	**Prec.**	**Rec.**	**f1-score**	**B. f1-score**
Low	0.39	0.43	0.41	0.33
Medium	0.33	0.31	0.32	0.34
High	0.44	0.26	0.43	0.33
Weighted Mean	0.39	0.39	0.39	0.33
Extraversion	**Prec.**	**Rec.**	**f1-score**	**B. f1-score**
Low	0.54	0.61	0.57	0.34
Medium	0.44	0.31	0.37	0.32
High	0.56	0.63	0.59	0.33
Weighted Mean	0.51	0.52	0.51	0.33
Agreeab.	**Prec.**	**Rec.**	**f1-score**	**B. f1-score**
Low	0.35	0.36	0.36	0.34
Medium	0.39	0.41	0.40	0.34
High	0.31	0.29	0.30	0.33
Weighted Mean	0.35	0.35	0.35	0.33
Neuroticism	**Prec.**	**Rec.**	**f1-score**	**B. f1-score**
Low	0.55	0.59	0.57	0.34
Medium	0.41	0.42	0.41	0.33
High	0.53	0.49	0.51	0.33
Weighted Mean	0.50	0.50	0.50	0.33



$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 6: Evolution of the performance of the inference with training dataset size by evaluating the model with k-fold cross validation with $k \in \{2, 3, 4, 5, 10\}$.